2
24.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
"CAMPUS ARAGÓN"

"ELEMENTOS PROBABILISTICOS EN EL RECONOCIMIENTO DE PATRONES"

T E S I S

QUE PARA OBTENER EL TITULO DE

INGENIERO EN COMPUTACION

P R E S E N T A N:

AIDA ALEJANDRA ALBARRAN SANCHEZ
CLAUDIA CARDENAS HERNANDEZ

ASESOR: ING. AMILCAR A. MONTERROSA ESCOBAR

MÉXICO 1997

La vida es acción. Quien no actúa, es como si hubiera nacido muerto. Transcurre por su época, como la sombra de una sombra. No es ascender cuando no se ama la luz. El reptil repugna a la cima que es meta reservada al impulso divino de las alas.

Carlos A. Madrazo

ING. AMILCAR A. MONTERROSA E.

A quien agradecemos el tiempo que dedicó para el desarrollo de este trabajo, así como el haber compartido sus conocimientos desinteresadamente para concluirlo.

Gracias a todos los profesores de Ingeniería y de otras áreas que participaron en nuestro desarrollo profesional y nos brindaron sus conocimientos; también a cada uno de nuestros compañeros que de una u otra forma nos apoyaron.

Agradecemos al Instituto Mexicano del Petróleo, al Departamento de Costos de Inversión y a todas las personas que nos apoyaron y nos dieron las facilidades para hacer posible la terminación de este trabajo.

Agradecemos a los sinodales por haber colaborado en la revisión de este trabajo.

Y a todas aquellas personas que merezcan nuestro agradecimiento por su apoyo.

A MIS PADRES

Con todo el cariño y respeto que se merecen, por el buen ejemplo que me inculcaron, por su dedicación, apoyo incondicional, constancia, comprensión y sabiduría y sobre todo por el gran esfuerzo que pusieron en mi realización personal y profesional.

A MIS HERMANOS

Con cariño por su apoyo incondicional y comprensión que tuvieron hacia mí.

A MIS SOBRINOS

Por su alegría, su inocencia y el cariño que me han brindado.

CLAUDIA CARDENAS H.

A DIOS

Por permitirme llegar a este momento y compartirlo con las personas que quiero.

A MIS PADRES

Con toda mi admiración, cariño y respeto por la confianza incondicional y comprensión que me han brindado, y por el esfuerzo que han hecho para lograr que se cumpla una de mis metas, que es también la suya. Pero sobre todo por ser mis mejores amigos.

A MI HERMANA

Con cariño por el apoyo que me ha dado en todo momento.

A MIS ABUELITAS

Eustoquia Albarrán (†) y Consuelo Soriano
Por su cariño y el buen ejemplo que me han dado.

A AGUSTIN

Por su comprensión y apoyo en todo este tiempo.

AIDA A. ALBARRAN S.

INDICE

INTRODUCCION

Debido a la necesidad por optimizar sus tareas el ser humano desarrolla continuamente técnicas y tecnologías en las diversas ciencias, en nuestro caso en las diferentes ramas de la informática siendo una de ellas la inteligencia artificial, donde se han tenido importantes avances en todas sus áreas particularmente en la del Reconocimiento de Patrones, el cual trabaja con métodos de comparación de patrones para tratar de clasificar los objetos, acontecimientos ó procesos en función de sus características.

En este escrito se mencionarán algunos elementos probabilísticos que se emplean en el Reconocimiento de Patrones para la clasificación de entradas. Este puede ser útil a las personas que se desenvuelvan en las áreas que se relacionan con el tema. Consta de seis capítulos que a continuación se describen brevemente:

En el capítulo 1: "Introducción al Reconocimiento de Patrones", se mencionan los conceptos básicos que se involucran en el Reconocimiento de Patrones, tales como: clase, patrón, funciones de decisión, etc. Y se presenta un panorama general de las áreas donde se aplica.

En el capítulo 2: "Teoría de la probabilidad", se citan conceptos de probabilidad y estadística tales como: probabilidad de un evento, probabilidad condicional, teorema de Bayes, distribuciones de probabilidad, etc. que sirven de base para el desarrollo de capítulos posteriores.

El capítulo 3: "Clasificación de patrones mediante funciones likelihood (de verosimilitud)", retoma del capítulo anterior el teorema de Bayes, de donde se deriva un tipo de clasificador de patrones que se utiliza comúnmente en la práctica y es conocido como clasificador bayesiano, del cual se determinan las funciones discriminantes que darán lugar a la clasificación.

En el capítulo 4: "Clasificadores entrenables. Una aproximación determinística", se describen algoritmos de aprendizaje mediante procesos iterativos, esto es, mediante aproximaciones deterministicas, siendo la más representativa el *perceptrón*.

En el capítulo 5: "Clasificadores entrenables. Una aproximación estadística", se analizan los clasificadores que son capaces de estimar información desconocida mediante aproximaciones estadísticas, resultando un proceso de aprendizaje iterativo que genera la solución correcta al problema de clasificación.

En el capítulo 6: "Aplicación", se utiliza la fórmula del clasificador bayesiano mencionada en el capítulo 3 aplicándola a un ejemplo.

CAPITULO 1

INTRODUCCION AL RECONOCIMIENTO DE PATRONES

CAPITULO 1
INTRODUCCION AL RECONOCIMIENTO DE PATRONES

1.1 INTRODUCCION

El ser humano desde que nace posee conocimientos innatos y aprende a reconocer "patrones" poco a poco, inicialmente aprende a reconocer a sus padres y familiares más cercanos detectando si alguien es extraño a él. Conforme va creciendo perfeccionará su Sistema de Reconocimiento de Patrones convirtiéndolo en un sistema muy sofisticado, adquiriendo la habilidad de distinguir diferentes tipos de patrones tales como: sonidos de voz e imágenes; leer manuscritos sin importar omisiones, variaciones o distorsiones; distinguir una sinfonía de Beethoven de una de Bach; el sabor de un helado y los patrones táctiles de frío y calor. Existen patrones que presentan manifestaciones físicas, como caracteres escritos sobre una hoja de papel o señales electromagnéticas que retoma un radar.

La corriente del Reconocimiento de Patrones se dió con el resurgimiento de los Sistemas Neuronales Artificiales, llevándola a un nivel alto de excitación al principio de la década de los sesentas.

A semejanza del pensamiento humano se han diseñado máquinas y lenguajes de programación que tratan de imitar su comportamiento ayudándose de métodos estadísticos que hacen mucho más fácil la exploración del Reconocimiento de Patrones.

Algunos programas del Reconocimiento de Patrones han reconocido y diferenciado obras de arte como un Picasso de un Matisse, ya que sus pinturas muestran una variedad de características simples con las que podemos diferenciar a los dos artistas. En la actualidad tratamos el Reconocimiento de Patrones simples a los cuales las personas han asignado nombres. Ejemplos de éstos son las letras de un alfabeto tales como A, B, C, ..., Z; los números 0, 1, 2, ..., 9; los objetos simples de nuestro mundo (mesas, sillas, perros, hojas, árboles, etc.); decir palabras y fonemas; y los patrones individuales de cada individuo.

Las técnicas que han sido desarrolladas y los modelos que se han construido, hoy en día son capaces de reconocer una amplia gama de patrones derivados de diferentes áreas de aplicación.

1.2 CONCEPTOS BASICOS

Los patrones son los medios por los cuales interpretamos al mundo y de acuerdo a la manera en que los distinguimos, se pueden clasificar en concretos y abstractos. El reconocimiento de **patrones concretos** incluye la identificación y clasificación de patrones espaciales (visual y aural) y temporales; para los cuales se necesita una ayuda sensorial. Ejemplo de patrones espaciales son caracteres, huellas digitales, pinturas, una pieza de música de cámara, objetos físicos, etc. y dentro de los temporales se incluyen formas de onda, lenguajes, electrocardiogramas (ECG), esto es, situaciones en base al tiempo. El reconocimiento de **patrones abstractos** se refiere al proceso de identificación de conceptos e ideas a través del pensamiento el cual puede hacerse sin la ayuda sensorial (vista y oído) por ejemplo, patrones sociales o datos económicos.

> *El Sistema de Reconocimiento de Patrones casi siempre forma parte de sistemas de información complejos, es una ciencia inexacta y por consiguiente admite muchas aproximaciones, algunas veces complementarse o algunas veces compitiendo con la solución aproximada de un problema dado. En un estilo diferente, podemos ver al reconocimiento de patrones como una mayor área de búsqueda común e impulsando al desarrollo por la necesidad para procesar datos e información obtenidos de la interacción entre científicos, tecnólogos y sociedad en general. Otra motivación para el estímulo de actividad en este campo es la necesidad de la gente para comunicarse con máquinas programadas en lenguajes naturales. La tercera y más importante motivación para el estudio en esta área es que científicos e ingenieros están interesados en la idea de diseñar y fabricar autómatas (máquinas inteligentes) que pueden llevar a cabo ciertas tareas con las habilidades comparables a la actuación humana* [1].

> *"Una **clase de patrón** es una categoría determinada por algún atributo común dado. Un **patrón** es la descripción de cualquier miembro de una categoría que representa a una clase de patrón"* [2]

El Reconocimiento de Patrones da por hecho la existencia de un "reconocedor", también un organismo viviente o un dispositivo físico programado. El reconocedor pretende discriminar entre una combinación de dos ó más patrones sobre fundamentos de observaciones realizadas; así, existirán por lo menos dos patrones distinguibles. Los patrones pueden definirse "a priori" [3] de acuerdo a ciertas relaciones señaladas para estar sujetos por las observaciones, dichas relaciones son consideradas como mediciones de las propiedades o los atributos de las muestras que se quieren clasificar. En la figura 1 se observa una representación de las clases de patrón.

[1] C. Bezdek y K. Pal Sankar, "Fuzzy Models for Pattern Recognition", p 6
[2] Tou Julius T. y Gonzalez Rafael C., "Pattern Recognition Principles", p.7
[3] "A priori" se refiere a una probabilidad antes de que se realice un experimento.

Figura 1. Representación de las clases de patrón.

Existe una variedad de definiciones de Reconocimiento de Patrones de las que mencionaremos sólo algunas:

"El Reconocimiento de Patrones puede ser definido como la categorización de un dato de entrada en clases identificables mediante la extracción de rasgos ó atributos significativos". [4]

"El Reconocimiento de Patrones puede ser considerado como un mapeo de muchos a uno del conjunto de todos los ejemplos variantes de los diferentes patrones que pueden ser identificados por el conjunto de estos patrones con nombre". [5]

"El Reconocimiento Automático de Patrones intenta automatizar una clase de procesos perceptual y cognoscitivo. Estos procesos incluyen extracción, identificación, clasificación y descripción de datos de patrones reunidos de ambientes reales y simulados". [6]

"El Reconocimiento de Patrones se refiere a la identificación y al nombramiento de objetos presentados a un dispositivo de sensación-percepción". [7]

Desde nuestro punto de vista podemos dar la siguiente definición:
"El Reconocimiento de Patrones consiste en clasificar apropiadamente como miembro de una clase, un patrón de entrada a partir de la extracción de rasgos o atributos comunes".

Los sistemas no necesariamente han sido diseñados para reconocer un conjunto de patrón específico, algunos han aprendido a reconocer una gran variedad de patrones, incluyendo rostros, mesas, sillas, figuras, etc. Un sistema de reconocimiento de caracter es un Sistema de Reconocimiento de Patrones el cual recibe señales ópticas como la entrada de dato e identifica el nombre del caracter.

[4] Tou Julius T. y González Rafael C., op. cit., p.6.
[5] Uhr Leonard Merrick, "Pattern Recognition, learning and thought", p.18
[6] C. Shapiro, "Encyclopedia of Artificial Intelligence", p.1116.
[7] Uhr Leonard Merrick, op. cit., p.29.

El propósito de un Sistema de Reconocimiento de Patrones es el de tomar la mejor decisión para clasificar un patrón de entrada X como miembro de alguna de las clases en consideración. Se realizan muchos procesos perceptuales cognoscitivos, desde que el patrón de entrada llega al Sistema de Reconocimiento y hasta el momento en que tome una decisión. Estos procesos obtienen de los patrones de entrada las características principales necesarias para llevar a cabo la clasificación.

Un patrón se describe por un número finito de variables escalares llamadas rasgos o características: $X_1, X_2, \ldots, X_n$. Un patrón particular es un punto $X=(X_1, X_2, \ldots, X_n)$ en un espacio de patrón X n-dimensional cercando la región en la cual los patrones pueden encontrarse. Existe un número finito de clases de patrón $W_1, W_2, \ldots, W_n$ dentro de los cuales deseamos clasificar puntos del espacio de patrón. La clase W, es referida muchas veces como clase i.

Para dar paso al reconocimiento automático de los objetos individualizados "segmentados" se lleva a cabo una transformación de éstos, de manera que se transforma en un vector X en donde sus componentes se definen características ó rasgos. Cada vector de características es comparado con un conjunto de vectores ya establecidos, formado por los vectores de rasgos de todos los objetos del universo de trabajo. En la figura 2 se representa el proceso de Reconocimiento de Patrónes.

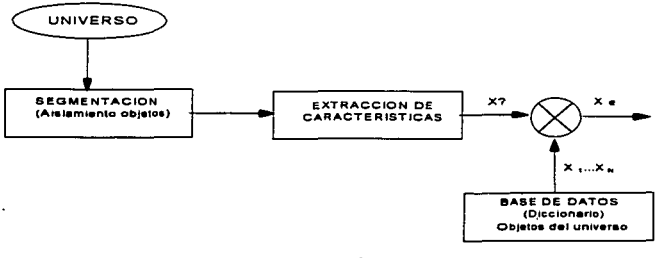


Figura 2

En la figura 2 se observa que los objetos individuales son convertidos en vectores numéricos X antes de que se realice el reconocimiento. Esta operación de cálculo ó extracción de las características ó rasgos diferenciadores de un objeto, es necesario en el rendimiento general de un

Sistema de Reconocimiento de Patrones específico. La notación que se emplea para los vectores de características forman un vector columna

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_n \end{pmatrix}$$

en donde X_1, X_2, ..., X_n son números reales que cuantifican, para un objeto en particular, las correspondientes características.

"Una vez calculado el vector X asociado a un objeto individual, su reconocimiento automático se basa en determinar su grado de semejanza con los vectores de características prototipos de cada posible clase de objetos previamente definidos".[6]

1.3 FUNCIONES DE DECISION

Hasta este momento, un problema de Reconocimiento de Patrones empieza con la definición de las clases en consideración y muestras etiquetadas de ésas clases en alguna representación practicable. Dicho problema no se puede resolver totalmente tomando en cuenta únicamente la medición de propiedades, es necesario el empleo de funciones lineales ó no lineales de tales propiedades.

Estas funciones se derivan de la medición de las propiedades y son menos en cantidad que el número de propiedades medidas al inicio. El problema se resuelve cuando el Sistema de Reconocimiento de Patrones genera una función de decisión la cual asigna rápidamente una única clasificación como miembro de una clase a un patrón de entrada nuevo.

Una vez derivada la función de decisión ésta puede ser muy simple y eficiente, mientras que la derivación requiere un mayor esfuerzo.

Considerando la figura 3 en donde suponemos que dos clases de patrones son conocidos, se ha trazado una recta que representa la función discriminante fd, la cual divide el plano de características en dos semiplanos que corresponden cada uno de ellos a una clase. En éste caso, hay una separación lineal de las clases debido a que se eligieron correctamente las características. Cuando las características no se eligieron correctamente existe un hiperplano que separa las clases.

[6] Maravall Darío, "Reconocimiento de formas y visión artificial", p 3.

es decir, existe separabilidad no lineal, y ésta se puede obtener con la ayuda de Redes Neuronales Artificiales.

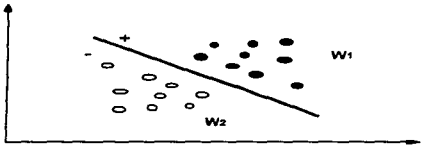


Figura 3

Dado que la ecuación de la función discriminante es una recta, da como ecuación el producto escalar de dos vectores:

$$fd(X) = W^T \cdot X = [W_1\ W_2\ W_3] \begin{pmatrix} X_1 \\ X_2 \\ 1 \end{pmatrix} \qquad (1\text{-}1)$$

donde T es el superíndice que indica que W es un vector traspuesto.

El vector de características se ha ampliado en una componente ficticia para que se pueda emplear una notación algebraica reducida. Hacemos a fd(X) como la ecuación de la línea de separación:

$$fd(X)=W_1X_1 + W_2X_2 + W_3 = 0 \qquad (1\text{-}2)$$

donde las W's son parámetros y X_1, X_2 son variables de coordenadas generales.

Nos podemos dar cuenta que los puntos situados en un semiplano dan lugar a un signo uniforme de fd(X) (positivo ó negativo), esto es, fd(X) será positiva cuando un patrón X es miembro de la clase W_1 y fd(X) resultará negativa cuando pertenezca a W_2.

De lo anterior podemos decir que un patrón X de clasificación desconocida es de la clase W_1 si fd(X)>0 ó a W_2 si fd(X)<0. Todos los patrones que están sobre la recta discriminante causan que fd=0.

El éxito del esquema de clasificación de patrones depende de dos factores:
1) La forma de fd(X) y
2) Habilidades para determinar sus coeficientes.

El primer problema está relacionado con propiedades geométricas de la clase patrón. En cuanto al segundo punto es considerablemente más difícil cuando la dimensionalidad de los patrones es más grande. Bajo éstas condiciones el único recurso racional es una aproximación analítica estricta. En el problema de los coeficientes podrían ayudar algoritmos adaptables como perceptrones y redes neuronales.[9]

En la solución general para n características, las funciones discriminantes pasan a ser hiperplanos pero la notación algebraica de la expresión (1-1) no cambia. Es decir, hay que *"encontrar un hiperplano $W^T \cdot X$ que divide el espacio de las clases en dos regiones, asociadas cada una de ellas a diferentes clases"*. [10] Se necesita obtener más de un hiperplano cuando el número de clases sea mayor a dos.

En el caso de N clases W_1, W_2,..., W_N no es común establecer un número mínimo de funciones discriminantes que serían necesarias para dividir el espacio de las clases en N regiones separadas, cada una relacionada con una sola clase. *"El procedimiento común es obtener una fd de dos clases"*. [11] Es posible discriminar N clases al mismo tiempo a través de una solución multiclase en condiciones de aprendizaje.

El planteamiento anterior del problema de Reconocimiento de Patrones ha dado una solución lineal que tiene sus principios en una "regionalización" del espacio de las clases. Pero existe también otra solución lineal que se basa en la clasificación de patrones mediante funciones de distancia (medidas de distancia), éste método es apropiado cuando la clase tiende a tener propiedades de agrupamiento. Esto es, cuando tenemos un patrón X de entrada para ser reconocido, el principio de ésta clasificación es relacionar el patrón X a la clase cuyo prototipo esté más cercano de X.

1.4 DISEÑO Y METODOLOGIA

A través del tiempo se han desarrollado una gran variedad de técnicas de diseño y metodologías, y continúan surgiendo más aproximaciones las cuales se establecen en base a distintos conceptos que son los siguientes: patrón ó clase, rasgos ó características y funciones discriminantes ó de decisión; de acuerdo a lo anterior se tienen las siguientes etapas de diseño.

PRIMERA ETAPA: RESOLUCION DEL UNIVERSO DE TRABAJO

"La primera etapa en el diseño consiste en el establecimiento de las clases: en lo que se podría denominar como definición del universo del trabajo del sistema".[12] En muchos casos esta etapa es

[9] Tou Julius T. y González Rafael C., op. cit., p.39.
[10] Maravall Darío, op. cit., p.5.
[11] Ibidem, p.5.
[12] Ibidem, p.7

considerada como directa y trivial, ya que la persona encargada del diseño del sistema conoce a la perfección los patrones ó clases de objetos que serán reconocidos. Sin embargo, podría suceder que las clases sean desconocidas a priori. En tales condiciones es necesario apoyarse en técnicas de agrupación (clustering).

SEGUNDA ETAPA: SELECCION Y DEMOSTRACION DE LAS CARACTERISTICAS

Después de haber definido las clases en la primera etapa, se prosigue con esta etapa, la cual consiste en llevar a cabo la elección del vector de características.

Esta etapa es un tanto compleja y la calidad del sistema final está regida por los rasgos escogidos en la etapa anterior.

El vector de características es el elemento principal en un sistema de reconocimiento, y no existe una regla general para determinarlo, esto es, para la elección de los rasgos ó características es necesario recurrir a la intuición y experiencia dependiendo de la aplicación que se desea desarrollar.

Aún así, existen técnicas de apoyo para crear el vector de características, pero se aplican una vez que se eligió un conjunto específico de características. Dichas técnicas se derivan de la estadística como el análisis de componentes principales y el análisis discriminante, que determina la calidad de las características elegidas.

TERCERA ETAPA: CALCULO DE LAS FUNCIONES DISCRIMINANTES

Una vez elegido el vector de características del sistema, esta etapa se encargará de realizar los cálculos necesarios de las funciones discriminantes ó de decisión.

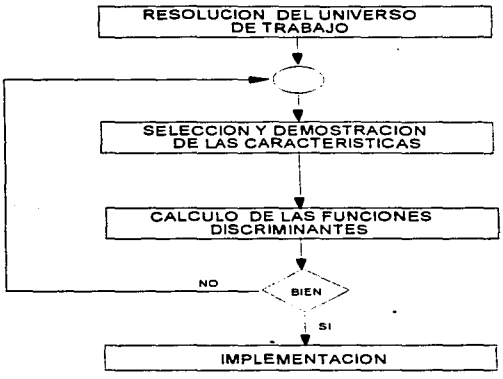


Figura 4. Diagrama de flujo de las etapas en el diseño de un sistema de reconocimiento

Las técnicas de diseño pueden ser implementadas por tres métodos principales

a) Métodos heuristicos.
b) Métodos matemáticos.
c) Métodos lingüísticos ó sintácticos.

a) METODOS HEURISTICOS

Esta aproximación se basa en los conocimientos empíricos del ser humano. Los sistemas que emplean este método se fundan en el desarrollo de procedimientos exactos para tareas de reconocimiento especializadas, ya que no cuenta con principios generalizados necesita de la aplicación especifica de reglas de diseño. Debido a éso, la funcionalidad del sistema heuristico dependerá de la experiencia e intuición de la persona que lo desarrolle.

b) METODOS MATEMATICOS

Se basan en aproximaciones deterministicas y estadisticas. La aproximación determinística está basada en una estructura matemática que no emplea propiedades estadísticas de las clases de

patrón bajo consideración. Un ejemplo de ésta, es el algoritmo de aprendizaje iterativo. La aproximación estadística está basada en reglas de clasificación matemática, dichas reglas se derivan en una estructura estadística. Un ejemplo de esta aproximación es la clasificación de Bayes y sus variaciones, esta regla produce un clasificador óptimo cuando son conocidas la función de densidad de probabilidad de cada patrón y la probabilidad de ocurrencia de cada clase de patrón.

c) METODOS LINGÜISTICOS O SINTACTICOS

La categorización de patrones a través de elementos primitivos (subpatrones) y sus relaciones proponen un Reconocimiento de Patrón automático mediante la aproximación lingüística o sintáctica. Un patrón puede representarse por medio de una estructura jerárquica de subpatrones análogos a la estructura sintáctica del lenguaje. En este método se puede aplicar la teoría de lenguajes formales. Un patrón gramático consta de conjuntos finitos de elementos llamados variables, primitivos y producciones. El tipo de gramática es determinado por las reglas de producción. Esta aproximación es muy útil en los casos cuando los patrones no se pueden describir por medio de mediciones numéricas o porque son muy complejos.

1.5 APLICACIONES

La década anterior ha presenciado los rápidos y considerables avances en la investigación y desarrollo del Reconocimiento de Patrones y Máquinas de Aprendizaje, ejemplos de éstos existen en abundancia y se aplican en muchos problemas de la vida real y en una gran variedad de áreas, a continuación se mencionan algunos de estos ejemplos:

1) PROCESAMIENTO DE INFORMACION MEDICA:
- Análisis de Electrocardiogramas (ECG's),
- Electroencefalogramas (EEG's),
- Análisis de rayos X y diagnósticos,
- Análisis de tejido celular.

2) COMUNICACION HOMBRE-MAQUINA:
- Reconocimiento de lenguajes autómatas,
- Identificación de voces,
- Comprensión de lenguajes e imágenes.

3) PROCESAMIENTO DE LENGUAJES.
4) APLICACIONES EN FISICA Y QUIMICA.

5) CRIMEN Y DETECCION CRIMINAL:
- Identificación de huellas digitales y fotografías,
- Escritura hecha a mano,
- Sonidos de lenguajes.

6) TIERRA Y CIENCIAS DEL ESPACIO:
- Clasificación de movimientos sísmicos.

7) ESTUDIO DE RECURSOS NATURALES Y ESTIMACION:
- Agricultura, silvicultura, medio ambiente,
- Hidrología, geología,
- Patrón de nubes.

8) APLICACIONES ESTEOROLOGICAS:
- Procesamiento de metal y mineral,
- Biología.

9) APLICACIONES MILITARES:
- Detección de explosiones nucleares,
- Radar,
- Detección de señales sonar,
- Detección de submarinos nucleares.

10) APLICACIONES INDUSTRIALES:
- Diseño asistido por computadora y manufactura,
- Simulación gráfica por computadora de pruebas y montaje de productos,
- Inspección automática y control de calidad en fábricas,
- Fallas y defectos en mecánica de aparatos.

11) RECONOCIMIENTO DE PATRONES AUDITIVOS:
- Identificación de palabras y de personas que están hablando.

12) ROBOTICA E INTELIGENCIA ARTIFICIAL:
- Tecnología sensor inteligente,
- Procesamiento del lenguaje natural.

Estas son algunas de las muchas áreas en las que se puede aplicar el Reconocimiento de Patrones, a continuación se listan algunos ejemplos en los que ha sido aplicado exitosamente.

RECONOCIMIENTO DE CARACTERES.

Cuyo objetivo es identificar y clasificar a los miembros de una muestra de un alfabeto dado de una matriz binaria, obtenida de caracteres digitados, en donde el patrón localizará un dispositivo de reconocimiento de caracter óptico, tal como las máquinas que leen el código de caracteres de cheques de bancos comunes. El grupo de caracteres se localiza en la actualidad en los cheques de los Estados Unidos, en la ya conocida American Bankers Association E-13B. El grupo de caracteres fuente consiste de 14 caracteres los cuales están especialmente diseñados sobre una área cuadriculada de 9 X 7 en orden para facilitar su lectura. Los caracteres están generalmente impresos en tinta, la cual contiene un material magnético muy finamente molido. Si el caracter ya fué leído por un dispositivo magnético, la tinta se magnetiza antes de que se realice la lectura en orden, para acentuar la presencia de los caracteres y de esta forma hacer más sencilla la tarea de lectura.

Los caracteres son rastreados regularmente en dirección horizontal con un sólo corte de lectura, el cual es estrecho pero más alto que el caracter, como los movimientos extremos atraviesan el caracter producen una señal eléctrica la cual es condicionada para hacer proporcional a la tasa del incremento del área del caracter por debajo del extremo.

El reconocimiento de caracteres de escritos a mano son aplicados en oficinas postales. Ambas técnicas, la estadística y sintáctica son usadas para este propósito, la última es especialmente adaptada para caracteres chinos, reconoce los rasgos cómo están pintados en segmentos. En general el sistema consiste en un programa de trazo de contornos que determina los límites de cada caracter componente. Una búsqueda es entonces conducida a encontrar un rasgo de segmento para ser usado como un punto de inicio. El algoritmo entonces arrastra a lo largo los rasgos hasta que el fin o unión de rasgos es encontrada. La gráfica de los componentes es entonces construida en términos de los rasgos de segmentos en un orden específico y una secuencia de primitivas es generada. Finalmente el reconocimiento es logrado partiendo de patrones sintácticos con representación de caracteres.

RADAR Y SONAR.

Aplicaciones en ingeniería incluyen análisis de señales en radar o sonar, el problema de reconocimiento de patrones que más se da en radar y sonar es el de dos clases, determinando la presencia ó ausencia de un blanco basado en señales de ruido recibidos desde áreas blanco. Los radares son usados para detectar aviones y blancos de buques por medio de transmisores y de sensores estacionados en tierra. El sonar utiliza ondas acústicas y son útiles únicamente debajo del agua, el ruido acústico es generado por maquinaria de las naves para la detección de los blancos; las

características utilizadas para clasificación por un radar se conoce como diagrama característico de un radar, un ejemplo de un radar multiclase es la clasificación de diferentes tipos de naves.

HUELLAS DIGITALES.

Las huellas digitales son poderosas en la identificación de seres humanos. La necesidad de la aplicación del reconocimiento automático de las huellas digitales en seguridad, surge debido al gran número de plantillas con las cuales tendría que ser comparada una muestra.

El FBI ha estado interesado muchos años en el desarrollo automático de sistemas que sirvan para identificar huellas digitales. Un ejemplo del esfuerzo realizado en esta área es el sistema prototipo llamado FINDER desarrollado por el Calspan Corporation para el FBI. Este sistema automáticamente detecta y localiza características únicas en una impresión.

Las técnicas sintácticas y los árboles gramaticales pueden ser aplicados a este problema.

PROCESAMIENTO DE INFORMACION MEDICA.

Identificación de enfermedades y áreas afectadas desde la caja de Rayos X y Tomografías auxiliadas por computadora (CAT), pueden ser examinadas y probadas por técnicas de Reconocimiento de Patrones. El Reconocimiento de Patrones sintáctico ha sido aplicado con éxito en la clasificación de cromosomas en el cual se realiza un análisis de tejido celular, en donde las clases pueden ser "sanas"; "dañadas" o "indeterminadas", para identificar los tipos de cromosomas desde sus imágenes usando rasgos cursivos primitivos. Las formas de onda médicas son procesadas por el Reconocimiento de Patrones como electrocardiogramas, electroencefalogramas, ondas de pulso, dolores de cabeza, etc. Técnicas sintácticas pueden ser usadas para clasificar los ECG's dentro de diferentes clases normales y anormales.

Otros ejemplos de aplicaciones de Reconocimiento de Patrones son: análisis de placas fotográficas expuestas en partículas de cámara de nubes, detección de materiales por medio de resonancia nuclear magnética, identificación de componentes ó recursos químicos, clasificación de rocas, predicción del comportamiento futuro de un sistema, entre otros.

En general, cualquier sistema puede ser considerado una "caja negra" representado por datos de entrada-salida, donde la salida es una clase clasificada y están disponibles modelos de datos para algoritmos de Reconocimiento de Patrones.

CAPITULO 2

TEORIA DE LA PROBABILIDAD

CAPITULO 2
TEORIA DE LA PROBABILIDAD

2.1 INTRODUCCION

En nuestra vida, estamos acostumbrados a escuchar y a decir sin planearlo deducciones que llevan implícito el concepto de probabilidad, tales como: la probabilidad de que llueva, la probabilidad de que un candidato gane en unas elecciones, la probabilidad de hacer un examen y quedarse en una cierta escuela, la probabilidad de que un grupo termine una carrera universitaria, la probabilidad de vivir muchos años, etc. Pero la teoría de la probabilidad tiene sus raíces en una simple teoría matemática de los juegos de azar por la época del siglo XVII, cuando se tuvo la necesidad de un método racional para calcular la probabilidad de ganar ó perder en un juego ya que, conforme pasaba el tiempo se incluían juegos más complicados con la ruleta, dados, cartas, etc. y se apostaban grandes cantidades de dinero.

En 1654, Blas Pascal (1623-1662) y Pierre de Fermat (1601-1665) 2 matemáticos franceses, formularon la teoría de la probabilidad mediante el intercambio de cartas. Desde entonces y a la fecha han contribuido a su perfeccionamiento varios matemáticos y científicos importantes que han determinado los conceptos de la probabilidad y los han colocado sobre una firme base matemática.

Podemos decir que la teoría de la probabilidad es la rama de las matemáticas que se encarga de los fenómenos que se producen al azar o fenómenos aleatorios (que dependen de un suceso imprevisto); y su finalidad consiste en proveer un modelo matemático adecuado a la descripción de ciertos fenómenos observados.

En casi todos los campos como la física, la química, la biología, la psicología, la sociología, la política, la economía, los negocios y todos las ramas de la ingeniería, se aplica la teoría de la probabilidad, y aparece hoy en día con otra disciplina importante que es la estadística, la cual se basa en los fundamentos de la teoría de la probabilidad.

La estadística implica procesos repetitivos como lanzar un dado 100 veces, el resultado de un examen de 50 estudiantes, etc. Por lo tanto, en la teoría de la probabilidad comenzamos con leyes de azar supuestas que empleamos como modelo para predecir los resultados de ciertos experimentos observados. En la estadística, examinamos los resultados de operaciones repetitivas y después tratamos de interpretarlos con la ayuda de las probabilidades calculadas.

2.2 PROBABILIDAD

2.2.1 EVENTOS Y ESPACIO MUESTRAL

Con el fin de hacer más sencilla la comprensión de términos que se emplearán en la teoría de la probabilidad, se empezará por definir el más sencillo:

- EXPERIMENTO. Proceso ó actividad que genera un conjunto de datos y nos lleva a un resultado u observación. Los experimentos son algo vital dentro de los procesos empleados en probabilidad. A cada uno de los posibles resultados de un experimento, ya sea uno prescrito o de cualquier otro tipo se le llama **evento**.

Y a la agrupación de todos los resultados posibles de un experimento se le llama **espacio muestral**. Este espacio se representa como un conjunto S, donde cada uno de sus elementos se llama un punto muestral ó muestra y está asociado a uno y sólo un resultado posible del experimento. Cada elemento del conjunto S se conoce como un evento simple (o indivisible); de tal manera que un evento simple es un subconjunto único de S. Y un evento compuesto (o divisible) es la unión de dos ó más eventos simples. De esta manera, cualquier evento simple ó compuesto es un subconjunto de S.

Por ejemplo: al lanzar un dado se observará el número que aparece en la cara superior, por lo que el espacio muestral (S) en este caso lo forman los números del 1 al 6.

$S = \{1,2,3,4,5,6\}$

Supongamos que A es el evento de que ocurra un número par, B un número impar y C un número primo:

$A\{2,4,6\}$, $B\{1,3,5\}$, $C\{2,3,5\}$

Por lo tanto:

$A \cup C$[1] $= \{2,3,4,5,6\}$ es el evento del número, sea par ó primo.

$B \cap C = \{3,5\}$ es el evento de que el número sea impar primo.

$C^C = \{1,4,6\}$ es el evento de que el evento no sea primo.

---

[1] Un conjunto es una lista o colección de objetos definidos; los cuales son llamados elementos o miembros.
OPERACION CON CONJUNTOS
Sean los conjuntos A y B.
La unión de A y B se representa por AUB, es el conjunto de elementos que pertenecen a A ó a B.
La intersección de A y B se representa por $A \cap B$, es el conjunto de elementos que pertenecen a A y B.
El complemento de A representado por A^c, es el conjunto de elementos que no pertenecen a A, es decir, A^c, es la diferencia entre el conjunto universal U y el conjunto A. Lipschutz Seymour, "Teoría y problemas de probabilidad", pp. 1,2.

A y B son eventos mutuamente excluyentes: $A \cap B = \varnothing$; es decir, un número par y un número impar no pueden ocurrir simultáneamente.

Los espacios muestrales se clasifican en **continuos** y **discretos** según el tipo de elementos que lo forman. Si los elementos de un espacio muestral pueden enumerarse, es decir, contarse, entonces es **discreto**; y es **continuo** cuando sus elementos no pueden enumerarse.

2.2.2 PROBABILIDAD DE UN EVENTO

"La teoría de la probabilidad es un sistema matemático compuesto de términos definidos e indefinidos y de un conjunto de suposiciones relativos a ellos; de todo esto obtenemos conclusiones lógicas, esto es, demostremos teoremas".[2] Es una disciplina abstracta que se utiliza como modelo para realizar conclusiones relativas a eventos que probablemente pueden ocurrir en una operación física real ó imaginaria.

Se puede definir a la probabilidad de una manera muy simple como un número de 0 a 1, que se le asigna a un fenómeno para indicar su posibilidad de ocurrencia. Se considera pertinente cuantificar las probabilidades conforme a una escala numérica de 0 a 1 (o de 0 a 100%). Entonces; *"La probabilidad de un evento es igual a la suma de las probabilidades de los eventos simples que lo componen; es decir, se cumple con el postulado de que P(S)=1".*[3] La probabilidad de ocurrencia del espacio de eventos es uno.

La definición de la probabilidad $P(A_i)$ de un evento simple A_i está dada por las siguientes reglas:

I. La probabilidad de un evento simple A se define como un número no negativo; es decir, una probabilidad no puede ser menor de 0 ni mayor de 1:

$$0 \leq P(A_i) \leq 1$$

II. La probabilidad de un espacio de eventos (S) que ocurre es igual a uno, para calcular ésta, se utiliza la regla de la suma de probabilidades, su función es la de ir sumando todas las probabilidades de todos los eventos simples de un espacio de eventos que lo componen

$$P(S)=1 \; ; \; P(A_1) + P(A_2) + \ldots + P(A_n) = 1$$

en donde $S = A_1 U A_2 U \ldots U A_n$

[2]Mode Elmer B. "Elementos de probabilidad y estadística", p. 2
[3]Rascón Octavio A., "Introducción a la teoría de probabilidades", p.41

III. Si los eventos A y B son mutuamente excluyentes, ambos eventos tienen igual probabilidad de ocurrir.

$$P(A \text{ ó } B) = P(A) + P(B)$$

La probabilidad de un evento vacío $\varnothing$ está dada por $P(\varnothing) = 0$.

Se puede usar la siguiente regla para calcular la probabilidad de un evento, que resulta de la regla de suma de probabilidades:

Si n(A) es el número de formas igualmente probables en que puede suceder el evento A, entonces

$$P(A)=n(A)/N$$

En donde N es el total de formas en que puede suceder el espacio de eventos.

Por ejemplo: si se lanza un dado al aire, y si el dado no está cargado, todos los eventos simples tienen igual oportunidad de ocurrir.

Sea el espacio de eventos S={1,2,3,4,5,6} y N=6.

Si todos los eventos simples tienen igual oportunidad de ocurrir entonces:

$P(1)=P(2)=P(3)=P(4)=P(5)=P(6)= 1/6$

$P(S)=P(1)+P(2)+P(3)+P(4)+P(5)+P(6)$

$P(S)=1/6+1/6+1/6+1/6+1/6+1/6= 6/6= 1$

2.2.3 PROBABILIDAD CONDICIONAL

En todas y cada una de las probabilidades calculadas hasta ahora, el denominador que representa la parte esperada, ha sido el conjunto universal. Esto es debido a que el conjunto universal es el grupo más grande de elementos por los cuales existe un interés.

"La probabilidad que se calcula utilizando información adicional a la que nos proporciona la sola descripción de un experimento, se llama PROBABILIDAD CONDICIONAL"[4] . En otras palabras, la probabilidad de un evento A es aquélla que se calcula con el anticipado conocimiento de llevar a cabo el experimento correspondiente. Al referirse a una probabilidad condicional se dice sencillamente "la probabilidad de A dado que ocurrió B" ó "probabilidad de A con la condición B".

Así, la forma de simbolizar la probabilidad condicional de ocurrencia del evento A, dado que ocurrió el B, o dicho de otra manera, la probabilidad para el evento A bajo la condición de que ha tenido éxito el evento B; es la siguiente $P(A|B)$ y se calcula por medio de la fórmula:

[4] Reséndiz Octavio A., op. cit., p.30

$$P(A|B) = n(A \cap B)/n(B)$$

Si dividimos entre N el numerador y el denominador de la fórmula anterior se obtiene

$$P(A|B) = [n(A \cap B)/N] / [n(B)/N]$$

en donde N es el total de maneras igualmente probables en que puede ocurrir el espacio de eventos al cual corresponden A y B.

Dado que $n(A \cap B)/N = P(A \cap B)$ y $n(B)/N = P(B)$, de acuerdo a la definición de probabilidad la fórmula queda en la siguiente forma:

$$P(A|B) = P(A \cap B)/P(B) \qquad \text{si } P(B) > 0$$

en donde $P(A \cap B)$ y $P(B)$ se encuentran a partir del espacio muestral original.

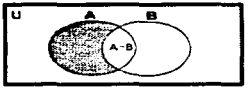


Figura 1. Representación de la probabilidad condicional de A dado B, mediante diagramas de Venn $P(A|B)$.

LA REGLA DE MULTIPLICACION DE PROBABILIDADES se determina a partir de la ecuación

$$P(A|B) = P(A \cap B)/P(B)$$

despejando $P(A \cap B) \Rightarrow P(A \cap B) = P(A|B) \cdot P(B)$

La fórmula $P(A \cap B) = P(B) \cdot P(A|B)$ denota que la probabilidad de que ocurra la intersección de A y B, es igual a la probabilidad de que ocurra B por la probabilidad condicional de A dado B.

Si A y B son independientes entonces $P(A|B) = P(A)$ por lo que $P(A \cap B) = P(B) \times P(A|B) = P(B) \times P(A)$.

De la ecuación $P(B|A) = P(A \cap B)/P(A)$

despejando $P(A \cap B) \Rightarrow P(A \cap B) = P(A) \times P(B|A)$

por lo tanto, la probabilidad de la intersección de A y B se puede calcular con alguna de las 2 fórmulas siguientes:

$$P(A \cap B) = P(A) \times P(B|A)$$
$$P(A \cap B) = P(B) \times P(A|B)$$

Dos eventos A y B son **independientes**, si se cumple que

$$P(A \cap B) = P(A) \times P(B),$$

es decir, si la probabilidad de la intersección de dos eventos es igual al producto de las probabilidades de cada uno de ellos, entonces los dos eventos son independientes, el evento B puede ocurrir sin modificar la probabilidad de ocurrencia del evento A, por lo tanto, la probabilidad de que ocurra el evento A no depende de la ocurrencia del evento B.

Por ejemplo: la probabilidad de que un tren salga a tiempo es P(D)= 0.83; la probabilidad de que llegue a tiempo a su destino es P(A)= 0.92; y la probabilidad de que salga y llegue a tiempo a su destino es $P(D \cap A)$= 0.78.
encontrar:

a) Probabilidad de que el tren llegue a tiempo, dado que salió a tiempo.
$P(A|D) = P(D \cap A)/P(D) = 0.78/0.83 = 0.94$

b) Haya salido a tiempo, dado que llegó a tiempo.
$P(D|A) = P(D \cap A)/P(A) = 0.78/0.92 = 0.85$

2.2.4 REGLA DE BAYES

Al matemático inglés Thomas Bayes (1702-1761) se le atribuye una fórmula para el cálculo de probabilidades. Esta se conoce como el **teorema de Bayes**, que se publicó en 1763 después de que murió, y permite calcular la probabilidad de que cualquier evento (o efecto) que haya ocurrido sea el resultado de una causa. Se ha empleado en diagnósticos médicos por computador y es uno de los fundamentos de la teoría de la decisión (problemas de tomar decisiones bajo la incertidumbre).

Se dice que dos o más eventos son mutuamente excluyentes si en caso de ocurrir uno, los otros no pueden ocurrir. Por ejemplo, los eventos de aprobar una asignatura son mutuamente excluyentes ya que alguien no puede aprobar y reprobar al mismo tiempo una asignatura.

PROBABILIDAD A PRIORI O CLASICA. Es aquella en la que no es necesario realizar el experimento para calcular las probabilidades deseadas, lo único es utilizar previamente un razonamiento lógico. Teniendo la siguiente definición: *"Si en un experimento pueden producirse N resultados igualmente probables y mutuamente excluyentes y si dentro de estos N resultados el evento E puede ocurrir N_E veces, la probabilidad del evento E, que se escribe P(E), está dada por*

$$P(E) = N_E / N \text{ [5]}$$

Por ejemplo: si tiramos un dado, la probabilidad de observar que salga el 1 es 1/6. Y si tiramos una moneda al aire, la probabilidad de que caiga águila es de 1/2.

PROBABILIDAD A POSTERIORI. Consiste en información proveniente de la probabilidad a priori después de que se realizó el experimento.

TEOREMA DE BAYES

"Si $A_1, A_2, \ldots, A_n$ son n eventos mutuamente excluyentes, cuya unión es el conjunto universal. B es un evento arbitrario tal que $P(B)>0$ y $P(B/A_k)$ y $P(A_k)$ son conocidos para $1 \leq k \leq n$, entonces:

$$P(A_i/B) = [P(A_i)\, P(B/A_i)]/[\; P(A_n)P(B/A_n)] \text{ [6]}$$

2.3 VARIABLES ALEATORIAS

2.3.1 CONCEPTO DE VARIABLE ALEATORIA

Se le llama **variable** al conjunto de caracteristicas, de personas ó cosas que presentan diferentes valores cuando se observan, como el número de zurdos en una escuela, el sexo de los niños nacidos en un sanatorio, etc. Cuando las variables toman únicamente valores numéricos se les llama "variables escalares", y cuando sólo toman valores indicados por nombres ó atributos se les llama "variables nominales", es decir, aquéllas que no pueden tomar valores numéricos como: éxito o fracaso; cara ó cruz; masculino ó femenino, etc.

Si un espacio muestral contiene un número finito de posibilidades o una secuencia sin final con igual número de elementos que números enteros, se le denomina ***espacio muestral discreto.***
Si un espacio muestral contiene un número infinito de posibilidades iguales al número de puntos que se encuentran en un segmento de línea, se le denomina ***espacio muestral continuo.*** [7]

Una **variable continua** es aquella que puede escoger algún valor dentro de un intervalo de valores y se mide uniformemente, como las que se miden con una escala de peso, altura, distancia, temperatura ó tiempo. Un ejemplo de variable continua es la estatura de una persona a través del tiempo. Una **variable discreta** es aquella en donde los valores que puede escoger están separados uno de otro por una determinada cantidad. Una característica de ésta es que presenta "vacíos" ó

[5] Wayne Daniel W., "Estadística en aplicaciones a las ciencias sociales y a la educación", p. 34.
[6] Ibidem p. 59.
[7] Walpole Ronald E.y Myers Raymond H., "Probabilidad y estadística para Ingenieros", p. 48.

"interrupciones" entre los valores que puede tomar. Y su conjunto de posibles resultados es contable. Una variable nominal es un ejemplo de variable discreta, mientras que una variable escalar puede ser continua ó discreta. Por lo tanto, se dice que las variables aleatorias continuas representan datos medidos y las variables aleatorias discretas representan datos que se cuentan.

Una **variable es aleatoria** cuando los valores que toma no se pueden predecir exactamente con anticipación al realizar un experimento. De esto, se puede dar la siguiente definición:

"Una variable aleatoria es una función que asocia un número real a cada elemento del espacio muestral."[8] Esto quiere decir que a cada elemento de S (espacio muestral) le corresponde un número real único, esto es, el valor de x.

Las variables aleatorias se representan con letras mayúsculas X, Y ó Z. Y su correspondiente letra minúscula x, y o z para representar algunos de sus valores.

En el caso de que la variable aleatoria Y tenga 6 valores, éstos se representan como: y_1, y_2, y_3, y_4, y_5 y y_6.

Por ejemplo; una variable aleatoria podría ser el número de caras que aparecen cuando se lanzan dos monedas.

De donde tenemos: S= {CC, CT, TC, TT} C= cara T= cruz

S= {(1,1),(1,0),(0,1),(0,0)}

Entonces:

X(1,1)=2, X(1,0)=1, X(0,1)=1 y X(0,0)=0

2.3.2 DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

DISTRIBUCION DE PROBABILIDAD

Una distribución de probabilidad se puede presentar mediante una tabla, fórmula ó gráfica. Se le llama **distribución de probabilidad** de una variable aleatoria al procedimiento que sirva para encontrar $P(X=x_i)$ que es la probabilidad del evento correspondiente a que la variable aleatoria X tome cada uno de los posibles valores x. A la fórmula que se utiliza para determinar $P(X=x_i)$ se le conoce como la **función de probabilidad** y se representa por $f(x_i)$, esto es, $f(x_i)= P(X=x_i)$.

[8] Ibidem p. 46

DISTRIBUCION DISCRETA

Una variable aleatoria discreta escoge cada uno de sus valores con determinada probabilidad, por lo que **la distribución de probabilidad** de la variable aleatoria discreta X es el conjunto de pares ordenados $(x_i, f(x_i))$ si, para cada resultado posible x:

$$1.- f(x_i) \geq 0$$

$$2.- \sum_{i=1}^{i=n} (x_i) = 1$$

$$3.- P(X=x) = f(x)$$

De lo anterior se puede concluir lo siguiente:

** Los elementos de la distribución de probabilidades tienen que ser mayores ó iguales que cero, esto es, $f(x_i) = P(x_i) \geq 0$ para todos los valores x_i que puede tomar la variable aleatoria X.

** Ningún valor de $P(x_i)$ debe ser mayor a 1 ni menor de 0:

$$0 \leq P(x_i) \leq 1$$

** *"Por otra parte, puesto que una distribución de probabilidades es el conjunto de las probabilidades correspondientes a todos los elementos del espacio de eventos de la variable aleatoria, se debe cumplir que la suma de todas esas probabilidades, sea 1."*[9] Por lo que si una variable aleatoria X puede tomar n valores se tiene que

$$\sum_{i=1}^{i=n} P(x_i) = 1$$

En un espacio muestral discreto se puede tabular una variable aleatoria X al enumerar en algún orden todos los puntos del espacio muestral y relacionar con cada uno el valor correspondiente de X.

Ejemplo: En la tabla siguiente tenemos la densidad de probabilidades con los siguientes valores: En la primer columna (x_i) están todos los valores que puede tomar x_i; en la segunda $P(x_i)$ están las probabilidades correspondientes a cada x_i.

[9] Rascón Octavio A., op. cit., p. 212.

x_i	$P(x_i)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

$$\sum_{i=2}^{i=12} P(x_i)=1$$

En la figura 2 se muestra la gráfica de la distribución de probabilidad de la tabla anterior, se nota que cada probabilidad se representa con una línea gruesa que parte del valor correspondiente de la variable aleatoria.

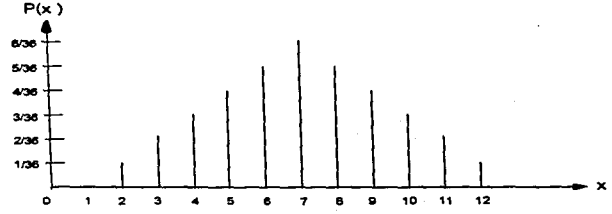


Figura 2.

2.3.3 DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

Se tienen variables aleatorias continuas siempre que se trata con experimentos que consisten en medir cantidades, las cuales pueden tomar cualquier valor dentro de ciertos límites. Por ejemplo, el nivel de azúcar en la sangre, la duración de la conversación telefónica, la distancia que caminó una persona, etc.

La variable aleatoria continua presenta una probabilidad de cero de escoger exactamente cualquiera de sus valores. Por lo tanto, representar su distribución de probabilidad de manera tabular es imposible. La distribución de una probabilidad continua se representa gráficamente a través de una curva suave. A pesar de que la distribución de probabilidad continua no se puede representar en forma tabular, se puede obtener una fórmula. Esta, será función de los valores numéricos de la variable continua X y se representará a través de f(x), que en este caso se conoce como "densidad de probabilidad o función de frecuencia de la distribución". Puede ser que f(x) tenga un número finito de discontinuidades ya que X se define para un espacio muestral continuo. Las gráficas de las funciones de densidad pueden ser de diferentes formas como se observa en la figura 3. Podemos ver que la función de densidad debe estar totalmente por encima del eje de las x's porque se emplearán áreas para representar probabilidades y dichas áreas son valores numéricos positivos.

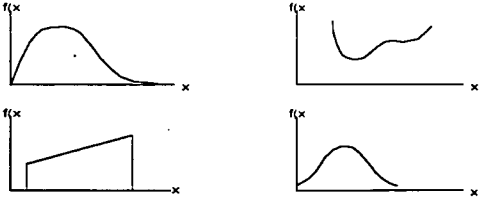


Figura 3.

En la figura 4 tenemos que *"la probabilidad de que X tome un valor entre a y b es igual al área sombreada bajo la función de densidad bajo las abscisas x=a y x=b; del cálculo integral se tiene que:*

$$P(a<X<b) = \int_a^b f(x)dx.$$ [10]

La diferencial f(x)dx se denomina elemento de probabilidad de la distribución.

[10] Walpole Ronald E., op. cit., p.55.

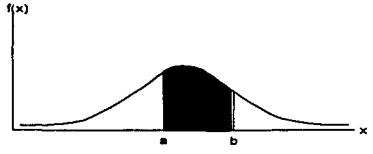


Figura 4. P (a<X<b).

De lo anterior decimos que f(x) es la función de densidad de probabilidad de la variable aleatoria continua X, la cual está definida sobre el conjunto de los números reales R, si:

1.- $f(x) \geq 0$ para toda $x \in R$

2.- $\int_{-x}^{x} f(x)dx = 1$

3.- $P(a<X<b) = \int_{a}^{b} f(x)dx$

2.3.4 DISTRIBUCIONES DE PROBABILIDAD ACUMULADAS

La función de distribución de probabilidad acumulada de X, denotada por F(x) es la probabilidad de que la variable aleatoria X tome datos menores o iguales a x_i. Así, tenemos que

$$F(x) = P(X \leq x_i)$$

Por lo tanto, si los elementos del espacio muestral de la variable aleatoria X están **ordenados en forma creciente de valores**, la probabilidad de que X tome un valor menor ó igual que un número dado x_m es la siguiente:

$$P(X \leq x_m) = \sum_{i=1}^{i=m} P(x_i)$$

donde: la i indica la posición en orden creciente, que tiene cada valor que puede tomar X.

Al conjunto de todas las $P(X \leq x_i)$ se le llama **distribución de probabilidades acumuladas.**

Para calcular la probabilidad acumulada hasta un elemento dado, tenemos que sumarle a la probabilidad de ocurrencia de ese elemento la probabilidad acumulada hasta el elemento inmediato anterior, esto es:

$$P(X \leq x_m) = P(X \leq x_{m-1}) + P(x_m)$$

se debe de satisfacer siempre la siguiente condición:

$$0 \leq P(X \leq x_i) \leq 1$$

es decir, que ningún elemento es menor que cero y la probabilidad acumulada del mayor de los valores que puede asumir X es 1 ya que se suman todos los elementos de la distribución de probabilidades correspondiente.

"La distribución acumulada F(x) de una variable aleatoria discreta X con distribución de probabilidad f(x) está dada por

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) \quad para \quad -\infty < x < \infty \text{ ."}^{[11]}$$

En la figura 5 se representa gráficamente una distribución acumulada discreta, la cual se obtiene dibujando los puntos (x,F(x)).

Figura 5. Distribución acumulada discreta.

"La distribución acumulada F(x) de una variable aleatoria continua X, con función de densidad f(x), está dada por

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad para \quad -\infty < x < \infty \text{ "}^{[12]}$$

En la figura 6 se representa gráficamente una distribución acumulada continua.

[11] Ibidem p. 51
[12] Ibidem p. 55

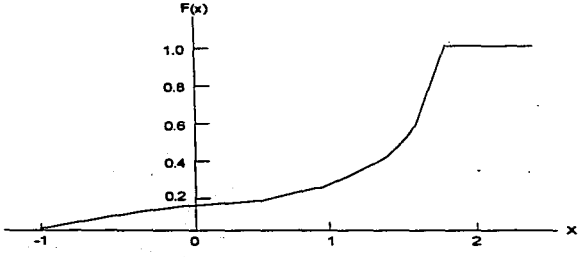


Figura 6. Distribución acumulada continua.

2.4 DISTRIBUCIONES DE PROBABILIDAD DISCRETA

A la agrupación de las probabilidades que corresponden a todos los elementos del espacio muestral de una variable aleatoria discreta se le llama distribución de probabilidades, la cual también describe el comportamiento de la variable por medio de gráficas, histogramas, en forma tabular ó por medio de una fórmula. Frecuentemente las observaciones originadas por experimentos estadísticos poseen el mismo tipo de comportamiento. En consecuencia las variables aleatorias discretas relacionadas con estos experimentos pueden ser definidas por la misma distribución de probabilidad y, por lo tanto se pueden representar por medio de una sola fórmula. Por lo que sólo es necesario utilizar algunas distribuciones de probabilidad para describir la gran mayoría de variables discretas aleatorias encontradas en la práctica.

2.4.1 MEDIDAS ESTADISTICAS IMPORTANTES

Un aspecto importante de la estadística es el análisis de resultados de operaciones repetitivas con el fin de interpretarlos a través de las leyes de probabilidad.

MEDIDAS DE TENDENCIA CENTRAL

Una medida de tendencia central es un número que señala el centro de una serie de números e implica el concepto de promedio. Las medidas de tendencia central más usuales son: la media

aritmética y la mediana, entre otras, en donde interviene el término de muestra la cual es un subconjunto de una población.

MEDIA ARITMETICA (MEDIA)

La media se refiere a un promedio. Se calcula sumando todos los datos para los que se quiere la media y dividiendo el resultado entre el número total de datos que se sumaron.

Si tenemos una variable aleatoria X y se le han tomado n medidas $x_1, x_2, ..., x_n$. La media de las n medidas es:

$$\text{Media} = \mu = [x_1+x_2+...+x_n]/n$$

quedando:

$$\overline{X} = \mu = [\sum_{i=1}^{n} x_i]/n$$

Es común entre los estadísticos llamar a este valor esperanza o expectativa matemática así como valor esperado de la variable X, representándola también como E(X).

Por ejemplo: Un profesor quiere obtener el promedio de aprovechamiento general de un grupo de alumnos y toma una muestra aleatoria de 15 calificaciones, obteniéndose los siguientes datos: 8.5, 8.0, 9.5, 7.0, 6.8, 8.8, 9.1, 10, 7.7, 6.9, 5.9, 8.4, 9.3, 9.7, 8.7. Calcular la media.

$$\overline{x} = (8.5+8.0+9.5+7.0+6.8+8.8+9.1+10+7.7+6.9+5.9+8.4+9.3+9.7+8.7)/15$$
$$\overline{x} = 8.28$$

MEDIANA

La mediana es el valor central de un grupo de datos ordenados en orden de magnitud. Si el número de datos es impar, la mediana es igual al valor de la mitad. Si el número de datos es par, la mediana es igual a la media de los dos valores que quedan a la mitad. Quedando representada de la siguiente forma: Si $X_1, X_2, ..., X_n$ conforman una muestra de tamaño n, establecida en orden creciente de magnitud, la mediana es:

$$X = \begin{cases} X_{(n+1)/2} & \text{si n es impar.} \\ [X_{n/2}+X_{(n/2)+1}]/2 & \text{si n es par .} \end{cases}$$

Por ejemplo: Los contenidos de nicotina en una muestra aleatoria de 8 cigarrillos de cierta marca, fueron de 3.5, 2.9, 3.4, 2.2, 2.6, 1.8, 3.1, 2.5 miligramos. Calcular la mediana.

Ordenando los valores se tiene 1.8 2.2 2.5 2.6 2.9 3.1 3.4 3.5
por lo tanto, la mediana es la media de 2.6 y 2.9. Quedando,

$x = (2.6+2.9)/2$

$x = 2.75$ miligramos

Por ejemplo: El número de computadoras que vende una tienda en 9 días seleccionados al azar fueron de: 2, 1, 5, 5, 1, 3, 4, 1, 2. Calcular la mediana.

Ordenando los valores, se tiene 1 1 1 2 2 3 4 5 5
por lo tanto, $x = 2$

MEDIDAS DE VARIABILIDAD

Una medida de tendencia central no da una descripción completa de un conjunto de datos; las medidas de variabilidad son medidas de la forma en que los valores individuales se desvían del promedio.

RANGO

El rango de una muestra aleatoria $X_1, X_2, \ldots, X_n$ es la diferencia entre el valor máximo y el mínimo de un conjunto de datos, y se representa con la fórmula siguiente:

$$X_{(n)} - X_{(1)}$$

donde:

$X_{(n)}$ es el valor máximo observado de la muestra y $X_{(1)}$ es el valor mínimo.

Por ejemplo: Los precios de una muestra aleatoria de 5 refrigeradores son: $1580.00, $2400.00, $5320.00, $3200.00, $3426.00. Calcular el rango.

El rango para los cinco valores es 5320.00 - 1580.00 = $3740.00

VARIANZA

"La varianza de un conjunto de datos se obtiene restando a cada uno de los valores el valor de la media de todos los valores, elevando al cuadrado cada una de las diferencias resultantes, sumando las diferencias al cuadrado y dividiendo este total por el número de valores menos 1." [13] Lo anterior se calcula con la siguiente fórmula:

$$\text{Varianza} = \sigma^2 = \left[\sum_{i=1}^{n} (X_i - X)^2\right]/(n-1)$$

[13] Wayne W. Daniel, op. cit., p.22.

Si tenemos una muestra aleatoria de tamaño n la varianza se calcula con la siguiente fórmula más útil:

$$\text{Varianza}=\sigma^2 = n\sum_{i=1}^{n} (X_i)^2 \left(\sum_{i=1}^{n} X_i\right)^2 / n(n-1)$$

DESVIACION ESTANDAR

La desviación estándar es la raíz cuadrada positiva de la varianza. Su fórmula es la siguiente:

$$\sigma = \sqrt{\sum_{i=1}^{n} (X_i^2-X)^2 / n-1} \qquad \text{ó}$$

$$\sigma = \sqrt{[n \sum_{i=1}^{n} X_i^2-(\sum_{i=1}^{n} X_i)^2] / n(n-1)}.$$

COVARIANZA

Cuando en una población se tienen los valores X y Y en la que son variables aleatorias, se le llama distribución bivariante. La distribución bivariante es un caso especial de una distribución conjunta en la que hay sólo dos variables (X y Y). Una distribución conjunta es aquélla en la que dos o más variables varían al mismo tiempo.

Si X y Y son variables aleatorias que tienen distribución de probabilidad conjunta, la covarianza de X y Y denotada por σ_{xy} ó cov(X,Y) que es la covariabilidad entre las dos variables, se calcula con la siguiente fórmula:

$$\sigma_{XY} = E(XY)-\mu_x \mu_y \qquad \text{donde } \mu=\text{media.}$$

"La covarianza será positiva cuando valores elevados de X se asocian con valores elevados de Y,y valores bajos de X se asocian a valores bajos de Y. Si los valores relativamente menores de X se relacionan con valores elevados de Y y viceversa, entonces la covarianza será negativa." [14] Cuando X y Y son estadísticamente independientes, la covarianza es igual a cero.

2.4.2 DISTRIBUCION DISCRETA UNIFORME

La más simple de todas las distribuciones de probabilidad discretas es aquélla en donde la variable aleatoria toma cada uno de los valores con igual probabilidad, distribuyéndolos uniformemente y siendo constante en un intervalo dado. Esta distribución de probabilidad se llama **distribución discreta uniforme**.

[14] Walpole Ronald E., op. cit. p.99.

"Si la variable aleatoria X toma los valores $x_1, x_2, \ldots, x_k$, con igual probabilidad, entonces la distribución discreta uniforme está dada por

$$f(x;k) = 1/k, \qquad x = x_1, x_2, \ldots, x_k"$$ [15]

La notación f(x;k) indica que la distribución discreta uniforme depende del parámetro k.

Ejemplo: Si se lanza un dado al aire, cada elemento del espacio muestral S={1,2,3,4,5,6} ocurre con probabilidad de 1/6. Como resultado se tiene la distribución uniforme

$$f(x;6) = 1/6, \qquad x = 1, 2, 3, 4, 5, 6$$

la distribución gráfica del lanzamiento del dado se hace por medio de un histograma, formado por un conjunto de rectángulos de igual altura como se ilustra en la figura 7.

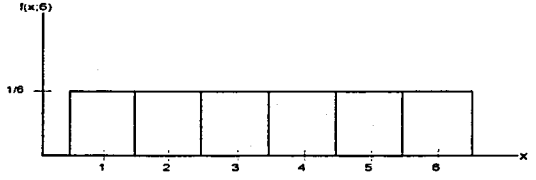


Figura 7. Histograma del lanzamiento de un dado.

Las propiedades de esta distribución son:

Media ------> $\mu = \sum_{i=1}^{k} x_i / k$

Varianza ------> $\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 / k$

[15] ibidem p.118.

2.4.3 DISTRIBUCIONES BINOMIAL Y MULTINOMIAL

DISTRIBUCION BINOMIAL

Considerando un experimento que consiste de pruebas repetidas e independientes con dos posibles resultados, de los cuales se puede producir únicamente uno de los dos; llamaremos a uno de los resultados éxito (ó favorable) s y al otro fracaso (ó desfavorable) f. Esto se comprueba al checar productos que salen de la línea de ensamble, en donde cada prueba o ensayo puede indicar un artículo defectuoso o uno sin defecto. Es indistinto definir uno u otro resultado como éxito.

Un experimento binomial es el que tiene las siguientes características:

1.- El experimento está compuesto de n pruebas ó ensayos repetidos.

2.- Cada ensayo proporciona un resultado que se puede clasificar como éxito ó fracaso.

3.- La probabilidad de éxito, definida por p, se mantiene constante de una prueba a otra.

4.- Los ensayos son independientes.

Las probabilidades de éxito (s) y fracaso (f) son:

$$P(s)= p, \qquad P(f)= q,$$

en donde:

$$p+q=1$$

entonces, sea p la probabilidad favorable, y así que q=1-p es la probabilidad desfavorable (n-x).

La distribución de probabilidad de la variable aleatoria binomial se llama distribución binomial y sus valores serán designados por b(x;n,p). Para determinar la probabilidad de x éxitos exactamente en n pruebas repetidas de un experimento binomial; se debe considerar en primer lugar la probabilidad de x éxitos y de n-x fracasos en un orden determinado. Dado que las pruebas son independientes, se pueden multiplicar todas las probabilidades que corresponden a los distintos resultados. Cada éxito ocurre con una probabilidad p y cada fracaso con una probabilidad q=1-p. En consecuencia, la probabilidad para un determinado pedido es $p^x q^{n-x}$. Se debe especificar ahora, el número total de puntos muestrales en el experimento que tiene x éxitos y n-x fracasos. Este valor es igual al número de particiones de n resultados en dos grupos, con x en un grupo y n-x en el otro, y está dado por $\binom{n}{x}$, como dichas particiones son mutuamente excluyentes, se realiza la suma de las probabilidades de todas las diferentes particiones para poder obtener la fórmula general o simplemente se multiplica $p^x q^{n-x}$ por $\binom{n}{x}$.

Si un ensayo binomial puede resultar en un éxito con probabilidad p y en un fracaso con probabilidad q=1-p, entonces la distribución de probabilidad de la variable aleatoria binomial X, el número de éxitos en n ensayos independientes, es

$$b(x;n,p)= \binom{n}{x} p^x q^{n-x}, \quad x= 0, 1, 2, \ldots, n$$ [16]

[16] Ibidem p.121.

Se llama distribución binomial puesto que para x=0, 1, 2,..., n corresponde a los términos sucesivos del desarrollo binomial. Esta, también se conoce como distribución de Bernoulli, y las pruebas independientes con dos resultados se llaman pruebas de Bernoulli.

Las propiedades de esta distribución son:

Media -----> $\mu = np$

Varianza -----> $\sigma^2 = npq$

Desviación estándar -----> $\sigma = \sqrt{npq}$

DISTRIBUCION MULTINOMIAL

Un experimento binomial se transforma en un **experimento multinomial** si se estima que cada ensayo podrá tener más de dos posibles resultados.

"La distribución binomial puede generalizarse fácilmente al caso de n ensayos repetidos e independientes, donde cada ensayo puede tener uno de varios resultados." [17] Generalmente si un ensayo dado lleva a cualquiera de los k resultados posibles: E_1, E_2, ..., E_k, en donde p_1, p_2, ..., p_k son las probabilidades respectivas de obtener los valores mutuamente excluyentes, entonces la distribución multinomial proporcionará la probabilidad de que E_1 ocurra x_1 veces, que E_2 suceda x_2 veces, ..., y de que E_k ocurra x_k veces, en n ensayos independientes. Así,

$$x_1 + x_2 + \ldots + x_k = n$$

Esta distribución de probabilidad se determinará mediante $f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n)$.

En general $p_1 + p_2 + \ldots + p_k = 1$, por lo que el resultado de cada ensayo debe ser uno de los k resultados posibles.

Dado que los ensayos son independientes, cualquier orden que se dé y que produzca x_1 resultados para E_1, x_2 para E_2, ..., x_k para E_k, ocurrirá con probabilidades $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$.

El número total de órdenes que produce resultados semejantes para los n ensayos es igual al número de repartir n artículos en k grupos, con x_1 en el primero; x_2 en el segundo; ...; y x_k en el k-ésimo. Esto se puede realizar en

$$\binom{n}{x_1, x_2, \ldots, x_k} = n!/x_1!\, x_2! \ldots x_k!$$

formas [18]

La distribución multinomial se obtiene al multiplicar la probabilidad para un orden especificado por el número total de particiones. Entonces la fórmula de la distribución multinomial se representa de la siguiente forma:

[17] Feller William, "Introducción a la teoría de la probabilidad y sus aplicaciones. Vol.I", p.171.
[18] Walpole Ronald E., op. cit., p.125.

$$f(x_1,x_2,\ldots,x_k;\ p_1,p_2,\ldots,p_k,\ n)=\binom{n}{x_1,x_2,\ldots,x_k} p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}$$

donde

$$\sum_{i=1}^{k} x_i = n \quad y \quad \sum_{i=1}^{k} p_i = 1.$$

Se le llama distribución binomial por el hecho de que los términos del desarrollo multinomial de $(p_1+p_2+\ldots+p_k)^n$ corresponden a todos los posibles valores de $f(x_1,x_2,\ldots,x_k;\ p_1,p_2,\ldots,p_k,\ n)$.

Ejemplo: Un dado se lanza 8 veces. La probabilidad de obtener los lados 5 y 6 dos veces y cada uno de los otros una vez es

$f(2,2,1,1,1,1;\ 1/6,1/6,1/6,1/6,1/6,1/6,1/6,1/6,\ 8)=$

$= (8!/2!\ 2!\ 1!\ 1!\ 1!\ 1!)\ (1/6)^2\ (1/6)^2\ (1/6)\ (1/6)\ (1/6)\ (1/6) = 35/5832 \approx 0.006$

2.4.4 DISTRIBUCION DE POISSON

Los experimentos que dan como resultado valores numéricos de una variable aleatoria X, el número de resultados que ocurren durante un periodo de tiempo dado o en una región especificada, frecuentemente es llamado experimento de Poisson. Puede ser cualquier magnitud: un minuto, un día, una semana, un mes, etc.

El experimento de Poisson posee las siguientes propiedades.

1.- El número de resultados que ocurren en un cierto intervalo de tiempo ó en una región especificada, es independiente del número que se tiene en cualquier otro intervalo.

2.- La probabilidad de que un solo resultado ocurra durante un lapso muy corto ó en una pequeña región, es proporcional a la magnitud del intervalo de tiempo o al tamaño de la región, y no depende del número de resultados que se produzcan fuera del intervalo considerado.

3.- La probabilidad de que ocurra más de un resultado en ese breve lapso ó de que caiga en una pequeña región es despreciable. [19]

El proceso de Poisson consta de eventos independientes que ocurren aleatoriamente en el tiempo. Así, la distribución de probabilidad de la variable aleatoria de Poisson X, que representa la cantidad de resultados que se generan en un rango de tiempo dado ó en una región específica, es

$$p(x;\mu)= e^{-\mu}\ \mu^x / x!, \qquad x= 0, 1, 2, \ldots,$$

μ es el número promedio de resultados que ocurren en el intervalo de tiempo dado ó en la región específica y e=2.71828....

[19] Ibidem p.139.

Esta distribución infinita se presenta en muchos fenómenos naturales, tales como el número de llamadas telefónicas por un minuto en un tablero de distribución, el número de errores por página en un texto muy grande, las emisiones de electrones de un cátodo caliente, y el número de partículas α emitidas por una sustancia radiactiva.

En la figura 8 se muestra el diagrama de la secuencia aleatoria típica de eventos con la distribución de Poisson.

Figura 8. Distribución de Poisson.

Las propiedades de la distribución de Poisson son:

Media -----> $\mu = \bar{x}$

Varianza -----> $\sigma^2 = x$

Desviación estándar -----> $\sigma = \sqrt{\bar{x}}$

2.5 DISTRIBUCIONES DE PROBABILIDAD CONTINUA

Anteriormente se ha estudiado la forma de actuar de la distribución binomial normalizada cuando se hace que aumente indefinidamente el número n, mientras que la probabilidad se mantiene constante, sin sufrir variaciones, y se ha visto que la variación binomial discreta se convierte entonces en una distribución límite continua: la distribución normal. También otras distribuciones límite resultan del tipo discreto.

2.5.1 DISTRIBUCION NORMAL

De todas las distribuciones continuas que se conocen, la distribución normal se considera una de las más importantes y también se conoce como distribución de Gauss. Su gráfica es llamada

curva normal, tiene la forma de campana como se muestra en la figura 9 la cual describe la distribución de muchos conjuntos de datos que ocurren, con media μ y varianza σ^2.

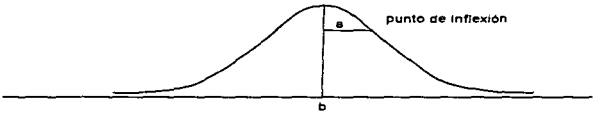


Figura 9. Distribución normal con media μ y varianza σ^2 donde $a=\sigma^2$ y $b=\mu$.

Una variable aleatoria continua X que tiene su distribución en forma de campana, se le llama variable aleatoria normal y depende de los parámetros media μ y desviación estándar σ.

La fórmula de la función de densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es

$$f(x)= [\ 1/\sigma\sqrt{2\pi}\ ]\ e^{-(1/2)[(x-\mu)/\sigma]^2}, \quad -\infty < x < \infty$$

donde μ= la media de la distribución

σ= la desviación típica de la distribución

π= la constante 3.14159...

e= la constante 2.71828...

Algunas características de importancia se mencionan a continuación:

1.- El área total comprendida bajo la curva y por encima del eje horizontal es igual a 1 (unidad cuadrada).

2.- *"La moda, es el punto en el eje horizontal donde la curva tiene su máximo, ocurre en $x=\mu$"* [20] así, la media, la mediana y la moda son todas iguales.

3.- La curva es simétrica respecto a un eje vertical que pasa por la media μ, es decir, el 50% del área está a la derecha de la mediana y el otro 50% se encuentra a la izquierda.

4.- La curva tiene sus puntos de inflexión en $x=\mu\pm\sigma$, es cóncava hacia abajo si $\mu-\sigma< X <\mu+\sigma$ y es cóncava hacia arriba en caso contrario.

[20] Ibidem p 147.

5.- Existe una distribución normal diferente para cada valor de μ y σ , el valor de μ coloca la distribución en el eje horizontal. Las distribuciones que tienen diferentes valores en las medias se sitúan en diferentes posiciones sobre el eje horizontal, mientras más grande sea la desviación típica más plana y más extendida es la gráfica de distribución.

6.- La curva de una distribución normal va desde $-\infty$ hasta $+\infty$.

2.5.2 APROXIMACION DE LA DISTRIBUCION NORMAL A LA BINOMIAL

La aproximación normal a la distribución binomial tiene una gran importancia teórica-práctica y juega un papel importante en el desarrollo del teorema del límite central.

La utilidad que tiene la distribución normal en estadística se entiende mejor si se toma en cuenta el hecho de que esta distribución proporciona una aproximación de la distribución binomial, cuando **n** es grande y **p** y **q** no son muy pequeñas (no está demasiado cercano a 0 ó 1), las distribuciones asociadas con una distribución binomial pueden aproximarse con una exactitud estimable por medio de las áreas correspondientes bajo la curva normal.

"Mediante los métodos del cálculo, se puede demostrar formalmente que el área bajo la curva normal es igual a la unidad. Debido a este hecho, cualquier área parcial bajo la curva se interpreta como una probabilidad " [21]

La curva de una distribución de probabilidad continua ó llamada también función de densidad, se traza de forma que el área bajo la curva delimitada por dos ordenadas $x=x_1$ y $x=x_2$, es igual a la probabilidad de que la variable aleatoria X tome un valor entre $x=x_1$ y $x=x_2$. En la figura 10 en la curva normal, el área bajo la curva se muestra en la región sombreada.

$$P(x_1<X<x_2) = \int_{x_1}^{x_2} f(x)dx$$

$$= [\, 1/\sigma\sqrt{2\pi}\,] \int_{x_1}^{x_2} e^{-(1/2)[(x-\mu)/\sigma]^2}\, dx$$

[21] Mode Elmer B., op. cit., p.150.

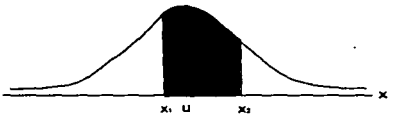


Figura 10. $P(x_1<X<x_2)$ = área de la región sombreada.

Es importante tomar en cuenta que las frecuencias binomiales son funciones de variable discreta, mientras que las frecuencias normales son funciones de variable continua, por esta razón es conveniente usar una corrección por continuidad, una regla práctica establece que la aproximación normal de la binomial es adecuada cuando np y n(1-p) son mayores que cinco.

Para hacer uso de la aproximación normal, hacemos que $\mu=np$, $\sigma = \sqrt{np(1-p)}$ y convertimos los valores de la variable original en valores de z para encontrar las probabilidades que presenten interés.

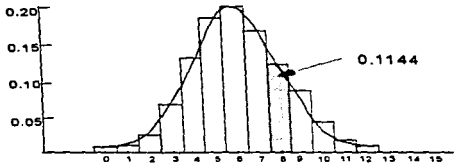
Ejemplo: La corrección de continuidad se muestra en la figura 11, donde n=20 y p=0.3. La probabilidad de X=x es igual al área del rectángulo centrado en x. Por ejemplo, la probabilidad de x=8 es igual al área del rectángulo centrado en 8. Este rectángulo se extiende de 7.5 a 8.5, de acuerdo con la tabla, el área es igual a 0.1144 que se muestra en la figura 11a con el área sombreada.

Si usamos la aproximación normal de la binomial, tenemos en cuenta que para la binomial, P(X=x) es el área del rectángulo centrado en x, cuando realizamos la conversión de los valores de x en valores de z, la corrección de continuidad consiste en sumar 0.5 a, y/o restar 0.5 de, x, según sea más conveniente. Utilizando la corrección de continuidad y la aproximación normal para encontrar la probabilidad de que X asuma un valor comprendido entre x_a=7.5 y x_b=8.5. Convirtiendo a valores z, entonces

$n= 20,\ p= 0.3\ \ \mu= np = 6.$

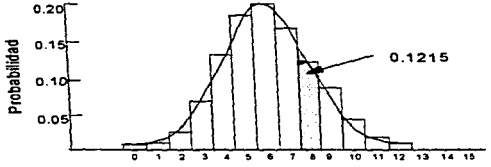
$z_a= 7.5-6/\ (20)(0.3)(0.7)= 1.5/2.05= 0.73$
$z_b= 8.5-6/\ (20)(0.3)(0.7)= 2.5/2.05= 1.22$

De acuerdo a tablas el valor buscado es 0.1215 que es aproximado a 0.1144 y el área bajo la curva normal, correspondiente a $P(7.5\leq X \leq 8.5)$ aparece sombreada en la figura 11b.



P(X=8) usando probabilidades binomiales

Figura a.



P(7.5 = X = 8.5) usando aproximación normales

Figura b.

Figura 11. Aproximación normal de la binomial con n=20, p=0.3 y μ=np= 6, en la que se muestra P(X=8).

CAPITULO 3

CLASIFICACION DE PATRONES MEDIANTE FUNCIONES LIKELIHOOD (DE VEROSIMILITUD)

CAPITULO 3
CLASIFICACION DE PATRONES MEDIANTE FUNCIONES LIKELIHOOD (DE VEROSIMILITUD)

3.1 INTRODUCCION

Este capítulo comienza con el estudio de la aproximación estadística para el Reconocimiento de Patrones. Esta aproximación involucra en su estimación las propiedades estadísticas de las clases de patrones para realizar un diseño de clasificación, está basada en información de la distribución adoptada de las características. El enfoque estadístico se emplea en los casos en que los vectores de algunas clases presentan una dispersión significativa, con lo que es posible derivar una regla de clasificación que tendrá una baja probabilidad de cometer errores al clasificar un patrón.

Se explica el teorema de Bayes y los dos tipos de clasificadores estadísticos que se derivan de él: el clasificador estadístico "a posteriori" y el clasificador estadístico "a priori", que como se verá es denominado **clasificador Bayesiano**.

El desarrollo principal en este capítulo es todo lo referente al clasificador Bayesiano ya que es la aproximación que se usa comúnmente. En el diseño de este clasificador es primordial la estimación de las funciones de densidad de probabilidad (FDP) $p(X|W_i)$ que son conocidas como funciones de verosimilitud, puesto que para clasificar cualquier patrón dentro de alguna clase se basa en estas FDP.

El clasificador Bayesiano empleado se aplica para patrones normales ya que son los más comunes en los casos prácticos. En este caso las clases siguen una distribución normal ó de Gauss donde se indican las medias (m) y las desviaciones estándar (típicas) σ, que determinan el comportamiento de una variable aleatoria Gaussiana, esto es, su fdp para poder clasificar un objeto (patrón).

Se dice, en general, que en el Reconocimiento de Patrones estadístico, la distribución de probabilidad de las características y el teorema de Bayes, se usan para generar reglas de asignación ó decisión óptimas para que haya el mínimo de errores en la clasificación.

3.2 INTRODUCCION GENERAL AL CLASIFICADOR BAYESIANO

La función de un clasificador de patrones es encontrar una decisión óptima para clasificar correctamente un patrón de entrada dado dentro de categorías ó clases.

El clasificador Bayesiano a través de consideraciones estadísticas puede derivar una regla de clasificación óptima que produce la más baja probabilidad de cometer errores de clasificación, por lo que representa la medida óptima de funcionamiento. Esto quiere decir, que minimiza la pérdida total esperada con respecto a todas las decisiones.

El clasificador Bayesiano se basa en el **Teorema de Bayes** el cual puede expresarse de la siguiente forma:

$$p(W_i \mid X) = \frac{p(X \mid W_i)\ p(W_i)}{p(X)} \qquad (3\text{-}1)$$

donde:

$p(W_i \mid X)$ es la probabilidad condicional **a posteriori** de que dado un vector de características X sea miembro de la clase W_i.

$p(X \mid W_i)$ es la probabilidad condicional de que dada una clase W_i el valor de la variable aleatoria sea X. Es decir, es la función densidad de probabilidad (fdp) de la clase W_i considerada como una variable aleatoria.

Este término es llamado con frecuencia **función likelihood** (función de verosimilitud) de la clase W_i.

$p(W_i)$ es la probabilidad **a priori** de que exista un elemento de la clase W_i, es decir, la probabilidad de ocurrencia de la clase W_i.

$p(X)$ es la probabilidad **a priori** de que exista un patrón a clasificar con un vector de características igual a X (un vector numérico concreto). Este elemento es un factor de escala que puede eliminarse por no tener información discriminante, ya que tiene el mismo valor para un conjunto de M clases, $W_1, W_2, \ldots, W_M$ compitiendo por el vector X a clasificar. Esto significa que es independiente de la pertenencia de X a una u otra clase ya que no depende de i.

Por lo tanto la ecuación (3-1) queda de la siguiente forma:

$$p(W_i \mid X) = p(X \mid W_i)\ p(W_i) \qquad (3\text{-}2)$$

CLASIFICACION ESTADISTICA A POSTERIORI

En la ecuación (3-2) se observa que el primer miembro proporciona la solución para encontrar a qué clase $W_1, W_2, \ldots, W_M$ pertenece un vector X dado.

La clasificación sería de acuerdo a lo siguiente:

$$X \in W_i \quad \text{si} \quad p(W_i \mid X) > p(W_j \mid X) \qquad (3\text{-}3)$$
$$\forall\, i \neq j, \quad i = 1,2,\ldots M.$$

Cuando el clasificador se basa en el primer miembro de la expresión (3-2) es llamado **"clasificador estadístico a posteriori"**. Esto se debe a que se intenta calcular ó estimar la probabilidad a posteriori de que X sea miembro de la clase W_i. Para calcular las probabilidades de este clasificador es necesario un proceso de aprendizaje, debido a que en este caso, las características estadísticas de patrones en cada clase tales como: $p(W_i)$ y $p(X|W_i)$ son parcialmente conocidas; se diseña un sistema de reconocimiento de patrón que tiene la capacidad de estimar la información desconocida durante su operación. Las decisiones son hechas en base a la información estimada.

Cuando la información estimada se aproxima gradualmente a la información real, entonces las decisiones basadas en la información estimada se aproximarán eventualmente a la decisión óptima.

En la siguiente figura se representa este tipo de clasificación.

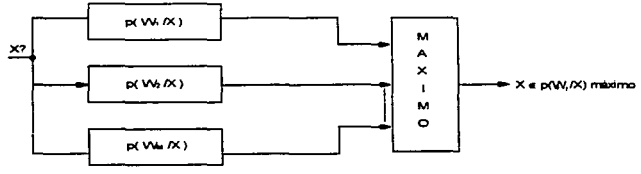


Figura 1. Estructura del clasificador estadístico a posteriori.

CLASIFICACION ESTADISTICA A PRIORI

Cuando el clasificador se basa en el segundo miembro de la expresión (3-2) es llamado **"clasificador estadístico a priori"** debido a que se trata de la probabilidad a priori de que

cumpliéndose la hipótesis de ser miembro de la clase W_i, el patrón tenga el valor numérico X. Por lo que se tiene otra forma de clasificar un vector X de acuerdo a lo siguiente:

$$X \in W_i \quad \text{si} \quad p(X|W_i)\ p(W_i) > p(X|W_j)\ p(W_j) \qquad (3\text{-}4)$$
$$\forall\ i \neq j, \quad i = 1,2,...M.$$

La principal dificultad para el diseño de este tipo de clasificador es la estimación estadística de las funciones de densidad de probabilidad (FDP) de las clases $p(X|W_1)$, $p(X|W_2)$, ..., $p(X|W_M)$ a partir de un conjunto de muestras ó características físicas de las clases W_1, W_2, ..., W_M que se consideran como variables aleatorias.

En la siguiente figura se representa este tipo de clasificación.

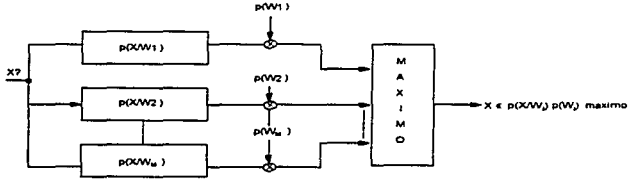


Figura 2. Estructura del clasificador estadístico a priori.

Del tema de funciones de decisión visto en el capítulo 1 se observa que las expresiones (3-3) y (3-4) son las implementaciones de las funciones de decisión:

$$d_i(X) = p(X|\ W_i)\ p(W_i), \quad i = 1,2,....M \qquad (3\text{-}5)$$

en la cual un patrón X se asigna a la clase W_i si para ese patrón $d_i(X) > d_j(X)$ para todo $i \neq j$.

Como los 2 clasificadores estadísticos vistos anteriormente están relacionados entre sí por el teorema de Bayes se les puede llamar clasificadores bayesianos. *"No obstante, el reconocedor a posteriori (así denominado por diseñarse a partir de la estimación mediante el aprendizaje ó entrenamiento supervisados de la fdp a posteriori) se obtiene en la práctica como una fdp, que es una aproximación lineal a la fdp real p(W_i|X)"* [1] ; por otro lado, el reconocedor a priori *"(denominado así por diseñarse a partir de la*

[1] Maravall Darío, "Reconocimiento de formas y visión artificial", p. 94

estimación estadística de la fdp a priori) se obtiene en la práctica como una estimación insesgada de la fdp real $p(X|W_i)$" [2]. Esto quiere decir, que lo que se obtiene en la práctica es una versión aproximada de los verdaderos reconocedores estadísticos. En el caso de los clasificadores a posteriori ó que se basan en el aprendizaje, se diseñan con hipótesis menos aproximadas a la realidad que en los clasificadores a priori en lo que se refiere a las distribuciones en probabilidad reales de las clases W_1, W_2, ..., W_M. Debido a esto, **a los clasificadores estadísticos a priori es a los que se les denomina clasificadores bayesianos**, ya que en la práctica los clasificadores estadísticos a posteriori son una aproximación menos verdadera al primer miembro del teorema de Bayes.

El diseño de un clasificador Bayesiano usando la ecuación (3-2) se puede observar en la figura 3.

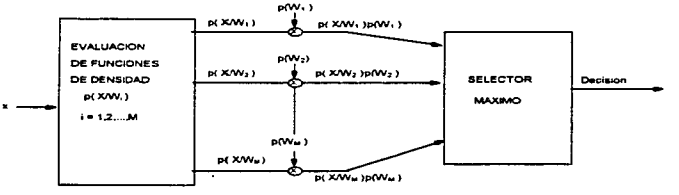


Figura 3. Clasificador Bayesiano.

3.3 CLASIFICADOR BAYESIANO PARA PATRONES NORMALES

Se les llama patrones normales a aquéllos que se ajustan a los criterios dados por la desviación estándar (σ) y la varianza (σ^2). En este caso las funciones de densidad de probabilidad $p(X|W_i)$ son normales multivariables (Gaussianas) y son las que se dan en la mayoría de los casos prácticos.

Comenzaremos con el caso más sencillo de función de densidad de probabilidad univariable para una variable aleatoria X, esto es, cuando se trabaja con una sola característica discriminante se tiene la siguiente ecuación:

[2] Ibidem p. 94.

$$p(X|W_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \quad \exp\left[- \tfrac{1}{2} (X-m_i / \sigma_i)^2 \right] \qquad (3\text{-}6)$$

Donde:

-- m_i es la medida ó valor esperado de X.

-- σ_i es la desviación estándar de la clase W_i.

Cuando se tiene un vector de características bidimensional, es decir, cuando se manejan dos características discriminantes X_1, X_2 la ecuación queda de la siguiente forma:

$$p(X|W_i) = \frac{1}{2\pi\,\sigma_{i1}\,\sigma_{i2}} \exp -1/2\,[(X_1 - m_{i1})^2/\sigma^2_{i1} + (X_2 - m_{i2})^2/\sigma^2_{i2}] \qquad (3\text{-}7)$$

Donde:

-- m_{i1} y m_{i2} son las medidas ó valores esperados de X_1 y X_2.

-- σ_{i1} y σ_{i2} son sus desviaciones estándar.

En el último caso se considera un vector de características n-dimensional con M clases de patrones, la función densidad de probabilidad queda de la siguiente forma:

$$p(X|W_i) = \frac{1}{(2\pi)^{n/2}\,|C_i|^{1/2}} \; e^{\left[-1/2\,(X-m_i)^T\, 1/C_i\,(X-m_i) \right]} \quad i = 1,2 \ldots M \qquad (3\text{-}8)$$

Donde:

-- n es la dimensionalidad de los vectores patrón.

-- cada densidad es especificada por su vector medio ó esperanza matemática (m_i) y la matriz de covarianza (C_i), definidas como sigue:

$$m_i = E_i\{X\}$$

$$C_i = E_i\{ (X-m_i)\,(X-m_i)^T \}$$

-- C_i^{-1} es la matriz inversa de la covarianza.

-- $|C_i|$ es el determinante de la matriz de covarianza C_i.

De esta manera se obtiene:

$$m_i = E_i(X) = E\begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_n \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ \cdot \\ \cdot \\ m_n \end{pmatrix} \qquad (3\text{-}9)$$

$$C_i = E_i\{(X-m_i)(X-m_i)^T\} \qquad (3\text{-}10)$$

$$C_i = E\begin{pmatrix} X_1 - m_1 \\ X_2 - m_2 \\ \cdot \\ \cdot \\ X_n - m_n \end{pmatrix} (X_1 - m_1 \quad X_2 - m_2 \quad --- \quad X_n - m_n)$$

$$C_i = E\begin{pmatrix} (X_1 - m_1)^2 & (X_1 - m_1)(X_2 - m_2) & --- & (X_1 - m_1)(X_n - m_n) \\ (X_2 - m_2)(X_1 - m_1) & (X_2 - m_2)^2 & --- & (X_2 - m_2)(X_n - m_n) \\ -- \; -- & -- \; -- & & -- \; -- \\ -- \; -- & -- \; -- & & -- \; -- \\ (X_n - m_n)(X_1 - m_1) & (X_n - m_n)(X_2 - m_2) & --- & (X_n - m_n)^2 \end{pmatrix}$$

C_i es una matriz simétrica respecto a la diagonal principal. Los elementos de la diagonal principal que llamaremos C_{kk} se denominan varianzas del k-ésimo elemento del vector patrón representadas como $\sigma_i^2 = E(X_i - m_i)^2$, i=1,2,..., n. Los demás elementos fuera de la diagonal principal que llamaremos C_{jk}, se denominan covarianzas de las variables X_j y X_k y cuando X_j y X_k son estadísticamente independientes $C_{jk} = 0$. Esto quiere decir que las matrices de covarianza son diagonales puras:

$$C_i = \begin{pmatrix} (X_1 - m_1)^2 & 0 \\ 0 & (X_2 - m_2)^2 \end{pmatrix}$$

La función de decisión para la clase W_i de acuerdo a la ecuación (3-10) puede ser $d_i(X)=p(X|W_i)\,p(W_i)$. Debido a la forma exponencial en la ecuación (3-8) que es la función de decisión normal, es mejor trabajar con el logaritmo natural de ésta; quedando de la siguiente manera:

$$d_i(X)= \ln[p(X|W_i)\,p(W_i)] = \ln p(X|W_i) + \ln p(W_i) \qquad (3\text{-}11)$$

Sustituyendo la ecuación (3-8) en la ecuación (3-11) se obtiene:

$$d_i(X)= \ln p(W_i) - n/2 \ln 2\pi - 1/2 \ln |C_i| - 1/2\,[(X-m_i)^T\, C_i^{-1}\, (X-m_i)] \qquad (3\text{-}12)$$
$$i = 1,2,..., M$$

El término $n/2 \ln 2\pi$ de la expresión (3-12) puede ser eliminado ya que no depende de i, quedando la siguiente ecuación:

$$\mathbf{d_i(X)= \ln p(W_i) - 1/2 \ln |C_i| - 1/2\,[(X-m_i)^T\, C_i^{-1}\, (X-m_i)]} \qquad (3\text{-}13)$$
$$\mathbf{i = 1,2,..., M}$$

La ecuación (3-13) representa la función de decisión de Bayes para patrones normales y es una función de decisión hipercuadrática ya que no aparecen términos mayores al 2o. grado en ella.

Por lo tanto:

Lo mejor que puede hacer un clasificador de Bayes para patrones normales es colocar una superficie de decisión general de 2o. orden entre cada par de clases de patrón. Si las poblaciones de patrones son caracterizadas exactamente mediante densidades normales, no obstante, ningunas otras superficies producirán mejores resultados en una base típica[3]

-- Si todas las matrices de covarianza son diferentes: $C_1 \neq C_2 \neq \ldots \neq C_M$ y no contienen elementos nulos, *"las funciones discriminantes son no lineales, ya que aparecen todas las combinaciones cuadráticas posibles de las componentes del vector de características"*[4] :

$$X_1^2, X_2^2, \ldots X_M^2, X_1X_2, X_1X_3, \ldots, X_{M-1}X_M \qquad (3\text{-}14)$$

-- Si todas las matrices de covarianza son iguales: $C_1 = C_2 = \ldots = C_M = C$ y no contienen elementos nulos, quiere decir que el vector de características (patrón) tiene un comportamiento estadístico parecido en todas las clases. Por lo tanto se pueden eliminar los términos independientes de índice i ya que no tienen un carácter discriminante, de esta manera, la ecuación (3-13) queda de la siguiente forma:

$$d_i(X) = \ln p(W_i) + x^T C^{-1} m_i - 1/2\, m_i^T C^{-1} m_i, \qquad (3\text{-}15)$$
$$i = 1,2,\ldots, M$$

La ecuación (3-15) representa un conjunto de funciones de decisión lineales.

Si además se tiene que $C=\bar{I}$, siendo I la matriz identidad y $p(W_i)=1/M$, donde $i=1,2,\ldots, M$ queda una ecuación de la siguiente forma:

$$d_i(X) = X^T m_i - 1/2\, m_i^T m_i, \qquad (3\text{-}16)$$

3.4 ESTIMACION DEL VECTOR MEDIO Y MATRIZ DE COVARIANZA

Una vez considerada la naturaleza estadística de las clases de objetos a clasificar y una vez admitida la hipótesis de que siguen distribuciones normales ó de Gauss, el aspecto práctico central y casi único para el diseño del correspondiente reconocedor automático es la obtención de los dos parámetros que determinan las funciones discriminantes: es decir, la matriz de covarianza y el vector medio.[5]

La estimación del vector medio ó esperanza matemática de la clase se puede expresar de la siguiente manera:

$$\mathbf{m = E(X) = 1/N \sum_{j=1}^{N} X_j} \qquad (3\text{-}17)$$

[3] Tou Julius T. y Gonzalez Rafeel C., "Pattern Recognition Principles", p. 121.
[4] Maravall Darío, op. cit., p. 118.
[5] Ibidem, p. 121.

donde:

N = número de elementos en la clase W_i.

X = muestra de la clase W_i.

Expresando de manera explícita la ecuación (3-17) se obtiene lo siguiente:

$$m = E\begin{pmatrix} m_1 \\ m_2 \\ \cdot \\ \cdot \\ m_n \end{pmatrix} = 1/N \begin{pmatrix} X_{11} + X_{12} + \ldots\ldots + X_{1N} \\ X_{21} + X_{22} + \ldots\ldots + X_{2N} \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ X_{n1} + X_{n2} + \ldots\ldots + X_{nN} \end{pmatrix} \qquad (3\text{-}18)$$

La matriz de covarianza se expresa como:

$$C = E\{(X-m)(X-m)^T\} = E\{X X^T\} - m\, m^T \qquad (3\text{-}19)$$

Por lo que su estimación será la siguiente:

$$C = 1/N \sum_{j=1}^{N} X_j X_j^T - m\, m^T \qquad (3\text{-}20)$$

Donde la matriz de covarianza está dada por:

$$C = \begin{pmatrix} C_{11} & C_{12} & \ldots\ldots & C_{1N} \\ C_{21} & C_{22} & \ldots\ldots & C_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ C_{n1} & C_{n2} & \ldots\ldots & C_{nN} \end{pmatrix}$$

3.5 PARAMETROS DE RIESGO ASOCIADOS AL CLASIFICADOR BAYESIANO

Después de haber visto el clasificador Bayesiano para patrones normales, veremos aquí la probabilidad de error relacionada a las clasificaciones de este reconocedor. En la figura 4 está representada la distribución de una característica única, se puede observar que las poblaciones se traslapan. Esto quiere decir que algunos de los datos de W_1 caerán cerca de los de W_2 y viceversa; por lo que algunas veces la decisión del clasificador sería incorrecta ya que clasificaría a x como perteneciente a W_1 cuando en realidad pertenece a W_2 (suponiendo que es seleccionado W_1 si $X \leq A$ y W_2 si $X > A$.

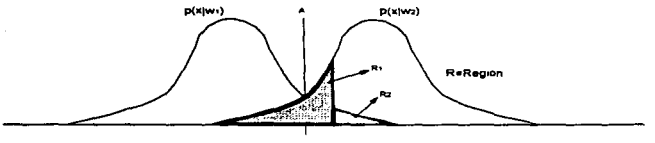


Figura 4. Representación gráfica de los errores que se pueden presentar en la clasificación de un objeto unidimensional entre dos clases W_1 y W_2.

En la figura 4, están representados los errores producidos en la clasificación de un objeto al asociarlo a una de las dos clases W_1 ó W_2. Como los errores son áreas se pueden expresar como integrales:

$$\epsilon_1 = \int_{R_1} p(X|W_2)\, dx = \int_{-\infty}^{x_t} p(X|W_2)\, dx$$

$$\epsilon_2 = \int_{R_2} p(X|W_1)\, dx = \int_{x_t}^{\infty} p(X|W_1)\, dx$$

Por lo tanto la probabilidad total de una clasificación errónea queda de la siguiente forma:

$$P\epsilon_T = p(W_2) \int_{R_1} p(X|W_2)\, dx + p(W_1) \int_{R_2} p(X|W_1)\, dx$$

$$P\epsilon_T = \epsilon_1\, p(W_1) + \epsilon_2\, p(W_2)$$

La integral es minimizada cuando $R_1 = \{X | p(W_1)\, p(X|W_1) - p(W_2)\, p(X|W_2) > 0\}$. Por lo que la regla de Bayes óptima asigna el objeto con vector característico X a la clase W_1 si:

$$p(X|W_1)/p(X|W_2) > p(W_2)/p(W_1) \qquad (3\text{-}21)$$

y de otro modo asigna X a W_2.

El término del lado izquierdo en la ecuación (3-21) es denominado con frecuencia *likelihood ratio (razón de verosimilitud)*:

$$l_{12}(X) = p(X|W_1)/p(X|W_2)$$

la cual es la razón de dos funciones de verosimilitud.

"El error mínimo es el criterio óptimo más simple que puede ser usado con la regla de Bayes. Es posible asignar pesos de varias clases a los varios tipos de errores y obtener reglas de decisión óptimas para situaciones más complejas."[6]

Regresando a la ecuación (3-21), ésta puede ser usada de la forma: $\log p(X|W_1) - \log p(X|W_2) > \log p(W_2) - \log p(W_1)$, ya que los cálculos son más fáciles.

[6] Nadley Morton y Smith Erick P., "Pattern Recognition Engineering", p. 349.

CAPITULO 4

CLASIFICADORES ENTRENABLES. UNA APROXIMACION DETERMINISTICA

CAPITULO 4
CLASIFICADORES ENTRENABLES. UNA APROXIMACION DETERMINISTICA

4.1 INTRODUCCION

Hasta ahora las aproximaciones para el diseño del clasificador de patrones han estado basadas en cálculos directos, en el sentido que las fronteras de decisión generadas por esta aproximación son derivadas de muestras de patrones, las cuales determinan los coeficientes de cálculo.

Un clasificador de patrones entrenable es un clasificador que puede perfeccionar su ejecución en respuesta a información que recibe.

En este capítulo se comienza con el estudio de clasificadores cuyas funciones de decisión son generadas de patrones entrenados por medio de algoritmos de "aprendizaje" iterativos. La idea de llamar aprendizaje viene de los primeros días del reconocimiento de patrones, cuando las primeras pruebas estuvieron hechas para tener un reconocimiento de programas desarrollado con lógica propia para aprender las clasificaciones de objetos sujetas a ciertas características requeridas para llevar a cabo un reconocimiento.

En el capítulo 1 un tipo de función de decisión fue especificado, el problema es la determinación de los coeficientes. Los algoritmos que se presentan en este capítulo son capaces de aprender la solución de coeficientes de los conjuntos entrenados, siempre que este conjunto de patrones entrenados sean separables por la funciones de decisión especificadas.

Si se dan dos conjuntos de patrones correspondiendo respectivamente a las clases W_1 y W_2, se desea encontrar un vector de peso solución W con la propiedad de que $W^TX > 0$ para todos los patrones de W_1 y $W^TX < 0$ para todos los patrones de W_2. Si los patrones de W_2 se multiplican por -1 se obtiene la condición equivalente $W^TX > 0$ para todos los patrones. N representa el número total de muestras de patrones aumentados en ambas clases, se puede expresar el problema de encontrar un vector W, tal que el sistema de desigualdades

$$X_W > 0 \qquad (4\text{-}1)$$

donde

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \cdots \\ X_n' \end{pmatrix}$$

$W = (W_1, W_2, \ldots, W_n, W_{n+1})'$, y 0 es el vector cero. Si los patrones están bien distribuidos, X satisface la condición de que cada submatriz (n+1) x (n+1) de X es de orden n+1.

Si existiera una W que satisfaga la expresión (4-1), las desigualdades serán consistentes; de otra manera son inconsistentes. En la terminología del reconocimiento de patrones se dice que las clases son separables o inseparables, respectivamente. La expresión (4-1) asume que todos los patrones han sido multiplicados por -1, y todos los patrones han sido argumentados.

Se puede tomar la aproximación estadística o la aproximación determinística para la solución de la expresión (4-1). La aproximación determinística forma las bases para los algoritmos que se desarrollan en este capítulo, estos algoritmos son desarrollados sin hacer alguna suposición referente a las propiedades estadísticas de las clases de patrón.

4.2 APROXIMACION MEDIANTE EL PERCEPTRON

En el aprendizaje, en los distintos modelos de redes neuronales se incluyen las redes de propagación hacia atrás, las máquinas de Boltzmann y los perceptrones; de los cuales uno de los más utilizados es definido como **algoritmo del perceptrón** por ser uno de los primeros modelos de redes neuronales. Este algoritmo se originó de los primeros esfuerzos hechos sobre conceptos biológicos aplicados a máquinas electrónicas con fines de aprendizaje en animales y máquinas.

A mediados de los años cincuentas y principios de los sesentas una clase de máquinas construidas por Rosenblatt (1957) llamadas perceptrones, parecían ofrecer un modelo natural y poderoso de aprendizaje de máquinas. Por esta época mucha gente especuló con la posibilidad de construir sistemas inteligentes utilizando a los perceptrones como elementos constituyentes, a partir de bloques construidos con éstos. "*Un perceptrón imita una neurona tomando una suma ponderada de sus entradas y envía a la salida un 1 si la suma es más grande que algún valor umbral ajustable (si ocurre de otro modo devuelve 0)*".[1]

El modelo básico de un perceptrón que es capaz de clasificar un patrón en una de dos clases se muestra en la figura 1. La máquina está compuesta de un arreglo S de unidades sensoriales, donde dichas unidades están conectadas a un segundo arreglo A de unidades asociativas. En donde estas unidades dan como resultado una única salida si es que hay las suficientes unidades sensoriales activadas.

[1] Rich Elaine, "Inteligencia artificial", p. 543

Se puede ver a estas unidades sensoriales como el significado por el cual la máquina recibe estímulos de su medio ambiente exterior, es decir, es el dispositivo de medición. Las unidades asociativas son consideradas como la entrada a la máquina.

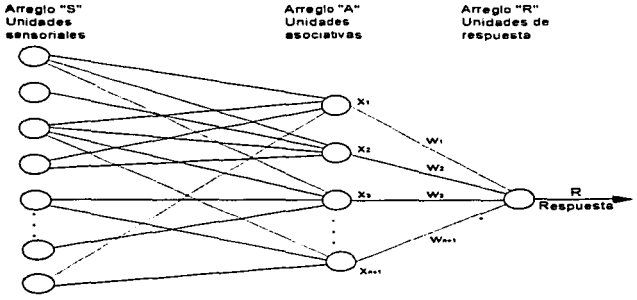


Figura 1. Modelo básico del Perceptrón.

Entonces se puede ver que la respuesta de la máquina es proporcional a la suma de las respuestas del arreglo asociativo, entonces la respuesta de las unidades de respuesta está dada por:

$$R = \sum_{i=1}^{n+1} W_i X_i = W^T X$$

$$R = W^T X$$

en donde:

X_i son las respuestas de la i-ésima unidad asociativa.

W_i es el peso correspondiente.

Si $R > 0$, el patrón visto por las unidades sensoriales pertenecerá a la clase W_1, y si $R < 0$ entonces pertenecerá a la clase W_2. El modelo básico del perceptrón es visto como una implementación de funciones lineales.

Algunos perceptrones se pueden combinar para calcular funciones más complejas, como se ve en la figura 1, que se puede extender para el caso multiclase, aumentando el número de unidades en el arreglo R (unidades de respuesta).

4.2.1 ALGORITMO DEL PERCEPTRON Y PRUEBA DE CONVERGENCIA

El algoritmo de entrenamiento ó aprendizaje del perceptrón es un algoritmo de búsqueda que empieza en un estado inicial aleatorio y acaba encontrando un estado solución, es un esquema simple para la determinación iterativa del vector de peso W. Consiste en incrementar ó disminuir el vector de peso W (ó de coeficientes) proporcionalmente a la muestra X que ha producido una decisión incorrecta. *"El espacio de búsqueda simplemente consiste en todas las posibles asignaciones de valores reales a los pesos del perceptrón y la estrategia de búsqueda es un descenso por el gradiente".*[2]

El esquema denominado como algoritmo del perceptrón se declara como sigue:

Teniendo dos conjuntos de patrones muestra que pertenecen a las clases W_1 y W_2 respectivamente, en donde W(1) representa el vector de peso inicial pudiendo ser escogido arbitrariamente. Entonces en el k-ésimo paso de entrenamiento se tiene:

* Si $X(k) \in W_1$ y $W^T(k)X(k) \leq 0$, sustituir W(k) por
  $W(k+1) = W(k) + cX(k)$ (4-2)

* Si $X(k) \in W_2$ y $W^T(k)X(k) \geq 0$, sustituir W(k) por
  $W(k+1) = W(k) - cX(k)$ (4-3)

* De otra manera dejamos a W sin cambios
  $W(k+1) = W(k)$ (4-4)

En donde c es una constante de corrección, y es un factor vital en la velocidad y fiabilidad de la convergencia del algoritmo hacia la solución del problema; un coeficiente c grande hace que sea rápido el aprendizaje pero puede llegar a desestabilizar la convergencia del algoritmo.

"El algoritmo del perceptrón es claramente un procedimiento de castigo-recompensa, donde la recompensa se da por la ausencia de castigo, esto es, si se clasificó correctamente el vector de peso no cambia".[3] De otro modo el castigo se da en forma proporcional a X(k).

[2] Ibidem, p. 550
[3] Tou Julius T. y González Rafael C., "Pattern Recognition Principles", p. 162

Por ejemplo: Sea la distribución biclase de la figura 2, en donde se pretende encontrar una solución del vector de peso. Como las dos clases de patrones son linealmente separables, el algoritmo tendrá éxito; existen dos clases y el objetivo será aprender una función discriminante lineal (una recta, ya que existen dos características: X_1 y X_2).

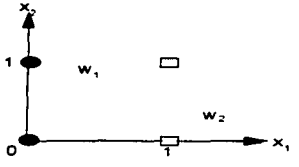


Figura 2. Patrones correspondientes a dos clases.

$$W_1 : \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \qquad W_2 : \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

Antes de que se aplique el algoritmo, los patrones son aumentados. Las clases son W_1 :{(0,0,1)' , (0,1,1)' } y W_2 : { (1,0,1)' , (1,1,1)' }.

donde c = 1 y W(1) = 0

Ahora se va a aplicar el algoritmo del perceptrón, presentando los patrones en el siguiente orden:

Primero es necesario inicializar la función de decisión fd, lo que es conocido como hiperplano de salida *"La velocidad y la fiabilidad del aprendizaje pueden llegar a ser muy dependientes de la posición inicial en ciertas distribuciones de las clases"*[4].

$$W_1 (1) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

El proceso de aprendizaje terminará cuando llegue a una fd que acierte en la clasificación de todas las muestras de entrenamiento.

---

[4] Maravall Darío, "Reconocimiento de formas y visión artificial", p. 53

$$fd(X) = W^T \cdot X = [W_1\ W_2 W_3] \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad W(k+1) = W(k) \pm cX(k)$$

$$W^T(1)X(1) = (0,0,0)\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0\ , \qquad W(2) = W(1) + X(1) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(2)X(2) = (0,0,1)\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = 1\ , \qquad W(3) = W(2) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(3)X(3) = (0,0,1)\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = 1\ , \qquad W(4) = W(3) - X(3) = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(4)X(4) = (-1,0,0)\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = -1\ , \qquad W(5) = W(4) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Hasta aquí se han agotado los datos de entrenamiento, pero como el clasificador aún no ha convergido a una solución, el proceso de aprendizaje de la máquina continuará haciendo la segunda iteración.

donde: X(5) = X(1); X(6) = X(2); X(7) = X(3); X(8) = X(4)

$$W^T(5)X(5) = 0 \qquad W(6) = W(5) + X(5) = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(6)X(6) = 1 \qquad W(7) = W(6) + X(6) = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(7)X(7) = 0 \qquad W(8) = W(7) + X(7) = \begin{bmatrix} -2 \\ 0 \\ 0 \end{bmatrix}$$

$$W^T(8)X(8) = -2 \qquad W(9) = W(8) + X(8) = \begin{bmatrix} -2 \\ 0 \\ 0 \end{bmatrix}$$

Dado que ocurrieron dos errores en esta iteración, los patrones se presentan nuevamente.

$$W^T(9)X(9) = 0\ , \qquad W(10) = W(9) + X(9) = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(10)X(10) = 1\,, \qquad W(11) = W(10) + X(10) = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(11)X(11) = -1\,, \qquad W(12) = W(11) + X(11) = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$W^T(12)X(12) = -2\,, \qquad W(13) = W(12) + X(12) = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

Siguiendo el proceso cuando t=9 se obtiene la solución correcta, se verifica fácilmente que en la siguiente iteración todos los patrones son clasificados correctamente, entonces el vector solución es W = (-2, 0, 1).

La función de decisión correcta para la distribución de clases es $fd(X) = -2X_1 + 1$, la cual, cuando el conjunto es igual a cero, se convierte en la ecuación de la frontera de decisión y es precisamente la recta $X_1 = ½$ ó mediatriz del segmento que une los centroides de ambas clases como se ve en la figura 3.

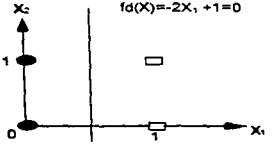


Figura 3. Frontera de decisión mediante entrenamiento.

El algoritmo del perceptrón se puede expresar multiplicando los patrones aumentados de una clase por -1. Así multiplicando arbitrariamente los patrones de W_2 por -1, se puede escribir el algoritmo del perceptrón como sigue:

$$W(k+1) = \begin{cases} W(k) & \text{si } W^T(k)X(k) > 0 \\ W(k) + cX(k) & \text{si } W^T(k)X(k) \le 0 \end{cases} \qquad (4\text{-}5)$$

en donde c es una corrección positiva aumentada.

PRUEBA DE CONVERGENCIA

La convergencia en el algoritmo ocurre cuando un vector de peso clasifica correctamente todos los patrones.

$X_1, X_2, \ldots, x_N$ representan un conjunto de patrones entrenados correspondientes a dos clases en donde los patrones de la clase W_2 se multiplican por -1, se garantiza que el perceptrón encontrará un estado solución, aprendiendo a clasificar cualquier conjunto de entradas linealmente separables, es decir, si las clases de patrones son linealmente separables el algoritmo de aprendizaje de la ecuación (4-3) produce un vector de peso solución W^*, con la propiedad:

$$W^{*T}X_i > 0, \quad i = 1, 2, \ldots, N$$

Es posible generalizar la expresión introduciendo un umbral no negativo T, tal que si las clases son linealmente separables,

$$W^{*T}X_i > T, \qquad i = 1, 2, \ldots, N$$

Así, el algoritmo de la ecuación (4-5) es

$$W(k+1) = \begin{cases} W(k) & \text{si } W^T(k)X(k) > T \\ W(k) + X_i(k) & \text{si } W^T(k)X_i(k) \le T \end{cases} \qquad (4\text{-}6)$$

En donde W(1) es arbitraria y se asume que c = 1, es decir, si c toma cualquier otro valor podría ser absorbido en los vectores patrón como una constante normalizada.

Si las clases son linealmente separables, el algoritmo de la ecuación (4-6) finalizará después de un número finito de pasos, si dejamos fuera los valores de k los cuales corresponden a patrones clasificados correctamente, entonces readaptando la anotación indicada, se puede expresar como:

$$W(k+1) = W(k) + X_i(k) \qquad (4\text{-}7)$$

y

$$W^T(k)X_i(k) \le T \qquad (4\text{-}8)$$

La convergencia del algoritmo significa que, después algún valor k_m infinito indicado

$$W(k_m) = W(k_m + 1) = W(k_m + 2) = \ldots = W(k_m + n)$$

Con las simplificaciones anteriores, la prueba de convergencia es, de la ecuación (4-7)

$$W(k+1) = W(1) + X_i(1) + X_i(2) + \ldots + X_i(k) \qquad (4\text{-}9)$$

tomando el producto interior de W^* con ambos lados de la ecuación (4-9) se obtiene

$$W^T(k+1)W^* = W^T(1)W^* + X_i^T(1)W^* + \ldots + X_i^T(k)W^* \qquad (4\text{-}10)$$

de la expresión (4-7), cada término $X_i(j)W^*$, $j = 1, \ldots, k$, es menor que T

$$W^T(k+1)W^* \geq W^T(1)W^* + kT \qquad (4\text{-}11)$$

Utilizando la desigualdad $\|a\|^2 \|b\|^2 \geq (a'b)^2$, resulta

$$[W^T(k+1)W^*] \leq \|W(k+1)\|^2 \|W^*\|^2 \qquad (4\text{-}12)$$

donde $\|a\|^2$ indica la magnitud de a cuadrada y la ecuación (4-12) se escribe como:

$$\|W(k+1)\|^2 \geq \{ [W^T(k+1)W^*]^2 \} / \|W^*\|^2 \qquad (4\text{-}13)$$

sustituyendo la expresión (4-11) en la (4-13) se tiene

$$\|W(k+1)\|^2 \geq \{ [W^T(1)W^* + kT]^2 \} / \|W^*\|^2 \qquad (4\text{-}14)$$

Una línea alternativa de razonamiento considerando $\|W(k+1)\|^2$ de la ecuación (4-7)

$$\|W(j+1)\|^2 = \|W(j)\|^2 + 2W^T(j)X_i(j) + \|X_i(j)\|^2 \qquad (4\text{-}15)$$

ó

$$\|W(j+1)\|^2 = \|W(j)\|^2 + 2W^T(j)X_i(j) + \|X_i(j)\|^2 \qquad (4\text{-}16)$$

Utilizando la expresión (4-7) y haciendo $Q = \max\|X_i(j)\|^2$ resulta

$$\|W(j+1)\|^2 - \|W(j)\|^2 \leq 2T + Q \qquad (4\text{-}17)$$

agregando estas desigualdades para $j = 1, 2, \ldots, k$ resulta la desigualdad

$$\|W(j+1)\|^2 \leq \|W(1)\|^2 + (2T + Q)k \qquad (4\text{-}18)$$

Comparando las expresiones (4-14) y (4-18), éstas establecen un conflicto de límites sobre $\|W(k+1)\|^2$ para k suficientemente grande. k no puede ser más grande que k_m, que es una solución a la ecuación.

$$\{ [W^T(1)W^* + k_mT]^2 \} / \|W^*\|^2 \} = \|W^*\|^2 + (2T + Q)k_m \qquad (4\text{-}19)$$

en donde k_m es finito, esto implica que el algoritmo del perceptrón converge en un número finito de pasos provisto de clases que son linealmente separables.

4.3 DERIVACION DE ALGORITMOS DE CLASIFICACION DEL PERCEPTRON

Como ya se vió, el desarrollo del algoritmo del perceptrón está basado en el concepto de Castigo-Recompensa. El algoritmo del perceptrón es sólo uno de una familia de esquemas iterativos los cuales se pueden derivar fácilmente y es utilizado por el concepto del gradiente.

4.3.1 LA TECNICA DEL GRADIENTE

El esquema del gradiente suministra una herramienta que sirve para poder encontrar el mínimo de una función. Del análisis del vector, el gradiente de una función f(y) con respecto al vector $y = (y_1, y_2, \ldots, y_n)$ se define como

$$\text{grad } f(y) = [\, df(y) \,] / dy = \begin{pmatrix} \partial f/\partial y_1 \\ \partial f/\partial y_2 \\ \cdots \\ \partial f/\partial y_n \end{pmatrix} \qquad (4\text{-}20)$$

De esta ecuación vemos que *"el gradiente de una función escalar de un vector argumento es un vector y que cada componente del gradiente da el porcentaje de cambio de la función en la dirección de ese componente"*[5]

Una de las principales propiedades del vector gradiente es que señala en el sentido del incremento proporcional máximo de la función f cuando el argumento aumenta. Contrariamente el negativo del gradiente señala en el sentido del incremento proporcional máximo de f, en base a esta propiedad se pueden idear esquemas iterativos para encontrar el mínimo de una función, a continuación sólo se considerarán a las funciones con un mínimo único. Si la función se escoge a fin de lograr el mínimo valor cuando $W^T X_i > 0$, en donde X_i es la i-ésima fila del NX(n + 1) de la matriz X del sistema de desigualdades dada en la expresión (4-1) entonces se encuentra el mínimo de la función para todo i, i = 1, 2, ..., N siendo equivalente para solucionar el sistema de desigualdad lineal.

Por ejemplo: Considerando la función criterio.

$$J(W,X) = (\, |W^T X| - W^T X) \qquad (4\text{-}21)$$

en donde $| W^T X |$ es el valor absoluto de $W^T X$ y el mínimo de la función es J(W,X) = 0 y el resultado mínimo se da cuando $W^T X > 0$, excluyéndose el caso cuando W = 0.

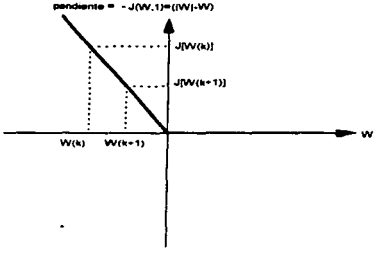
Es decir, W(k) representa el valor de W en el k-ésimo paso, el algoritmo general de descenso por el gradiente se escribe como:

[5] Tou Julius T. y Gonzalez Rafael C., op. cit., p. 169

$$W(k+1) = W(k) = c[[\partial J(W,X)] / \partial W]_{W=W(k)} \qquad (4\text{-}22)$$

donde W(k+1) es el nuevo valor de W, y c > 0 dictada la magnitud de la corrección, no se hacen correcciones en W cuando $(\partial J / \partial W) = 0$ la cual es la condición para un mínimo.

La ecuación (4-22) se puede interpretar geométricamente con la ayuda de la figura 4 que a continuación se ilustra.



$W(k+1)=W(k)-c[\partial J/\partial W]_{W=W(k)}$; pero $\partial J/\partial W = -2$ si $W \le 0$

$= 0$ si $W > 0$

por lo tanto, $W(k+1)=W(k) + 2c$ si $W \le 0$

$=W(k)$ si $W > 0$

Figura 4. Ilustración geométrica del algoritmo de descenso por el gradiente.

De este simple caso escalar vemos que si $(\partial J / \partial W)$ es negativo para el k-ésimo paso, W se incrementa en la dirección de el mínimo de J. De la figura 4 el esquema descendente guiará eventualmente a una W positiva, y, consecuentemente, al valor mínimo de J. Esta figura es una representación de la ecuación (4-21) para el vector X = 1.

Si las desigualdades son consistentes y una apropiada J(W,X) es seleccionada del algoritmo, resultará una solución. De otro modo, solo oscilará hasta que el procedimiento sea parado.

4.3.2 ALGORITMO DEL MINIMO ERROR CUADRATICO MEDIO Y PRUEBA DE CONVERGENCIA

Para los reconocedores deterministas, las funciones ó índice funcional del error que suelen manejarse, son el módulo del error ($J_1 = |e|$) y el cuadrado del error $J_2 = e^2$, utilizando la J para representar la función. Aquí únicamente se considerará al segundo, ya que da lugar a algoritmos de aprendizaje muy simples.

"Para minimizar (o en el límite, eliminar) el error es obligado primeramente a obtener una expresión matemática del error o, al menos ciertas propiedades analíticas que posibiliten su minimización".[6]

El algoritmo del perceptrón y sus variaciones convergen cuando las clases bajo consideración son separadas por la superficie de decisión especificada.

En la siguiente derivación se usará la fórmula dada en la expresión (4-1). En lugar de expresar el problema para encontrar un vector W tal que $X_W > 0$ es satisfecho, se buscará para vectores W y b, tal que

$$X_W = b \qquad (4\text{-}23)$$

en donde los componentes de $b = (b_1, b_2, \ldots, b_N)'$ todos son positivos. Esto es, las dos fórmulas son mutuamente equivalentes.

Considerando la función criterio

$$J(W,X,b) = 1/2 \sum_{i=1}^{N} (W^T X_i - b_i)^2 = 1/2 \, \| X_W - b \|^2 \qquad (4\text{-}24)$$

en donde $\| X_W - b \|$ es la magnitud del vector $(X_W - b)$.

La función J(W, X, b) logra su mínimo valor cuando la ecuación (4-23) se satisface. De esta función depende W y b, no hay razón por la que ambas variables no puedan ser usadas en el procedimiento de minimización. El término $(W^T X_i - b_i)^2$ ó $\| X_W - b \|^2$ expresa el error cuadrático entre las dos cantidades en el argumento. La suma de estos errores es proporcional a un promedio ó valor medio, el algoritmo resultante es llamado **algoritmo del mínimo error cuadrático medio (LMSE).**

[6] Meravell Dario, op. cit., p. 75

En vista de que J será minimizado con respecto a W y b, la aproximación debe tomar necesariamente diferencias escasas de el algoritmo general de la ecuación (4-22). Los gradientes que se asocian con el problema son los que se listan a continuación.

$$\partial J / \partial W = X'(X_W - b) \qquad (4\text{-}25)$$

y

$$\partial J / \partial b = - (X_W - b) \qquad (4\text{-}26)$$

Como W no es restringido de ningún modo, se puede establecer $\partial J / \partial W = 0$ y obtener

$$W = (X'X)^{-1} X'b = X^{\#}b \qquad (4\text{-}27)$$

Donde $X^{\#}$ se le llama el inverso generalizado de X. Debido a que todos los componentes de b son obligados a ser positivos. Este vector se debe variar de tal manera que no se pueda quebrantar esta obligación. Esto se establece por:

$$b(k + 1) = b(k) + \delta b(k) \qquad (4\text{-}28)$$

donde

$$\delta b, (k) = \begin{cases} 2c\,[X_W(k) - b(k)]; & \text{si } [X_W(k) - b(k)] > 0 \\ 0 & \text{si } [X_W(k) - b(k)] \leq 0 \end{cases} \qquad (4\text{-}29)$$

En las ecuaciones (4-28) y (4-29), k expresa el índice iterativo i, que es el índice de los vectores componentes y c es un incremento de corrección positivo para ser determinado después.

La ecuación (4-29) se puede escribir en forma de vector, es decir:

$$\delta b(k) = c[X_W(k) - b(k) + | X_W(k) - b(k) |] \qquad (4\text{-}30)$$

donde

$| X_W(k) - b(k) |$ es el valor absoluto de cada componente del vector $[X_W(k) - b(k)]$. De las ecuaciones (4-27) y (4-28), obtenemos

$$\begin{aligned} W(k+1) &= X^{\#}b(k+1) \\ &= X^{\#}[b(k) + \delta b(k)] \\ &= X^{\#}b(k) + X^{\#}\delta b(k) \\ &= W(k) + X^{\#}\delta b(k) \end{aligned} \qquad (4\text{-}31)$$

Dejando

$$e(k) = X_W(k) - b(k) \qquad (4\text{-}32)$$

tenemos el siguiente algoritmo:

$$W(1) = X^{\#}b(1),\ b(1) > 0,\ \text{pero de otra forma}$$
$$e(k) = X_W(k) - b(k)$$
$$W(k+1) = W(k) + cX^{\#}[e(k) + |e(k)|]$$
$$b(k+1) = b(k) + c[e(k) + |e(k)|] \qquad (4\text{-}33)$$

En la ecuación (4-33), $|e(k)|$ denota el vector cuyas componentes son el valor absoluto de $e(k)$, $W(k+1)$ también se puede calcular utilizando la relación $W(k+1) = X^{\#}b(k+1)$.

Cuando las desigualdades $X_W > 0$ tienen una solución, este algoritmo converge para $0<c\leq 1$. Si todos los componentes de $e(k)$ dejan de ser positivos (pero no todos son cero), en cualquier paso de la iteración, ésto indica que las clases no son separables. Por supuesto, que si $e(k) \geq 0$, y $b(k)$ es un vector positivo. Esta prueba de separabilidad es una importante característica del algoritmo.

PRUEBA DE CONVERGENCIA

La convergencia del algoritmo LMSE se aplica cuando las clases son linealmente separables y la corrección se incrementa satisfaciendo $0<c\leq 1$. La clave para demostrar la convergencia es para mostrar que el vector error $e(k) = X_W(k) = b(k) > 0$ indicando una solución para la ecuación (4-23).

De la ecuación (4-33) tenemos

$$e(k) = X_W(k) - b(k)$$

como $W(k) = X^{\#}b(k)$, esta ecuación se puede escribir de la siguiente forma

$$e(k) = (XX^{\#} - I)b(k) \qquad (4\text{-}34)$$

entonces

$$e(k+1) = (XX^{\#} - I)b(k+1) \qquad (4\text{-}35)$$

usando la expresión $b(k+1) = b(k) + c[e(k) + |e(k)|]$ producimos

$$e(k+1) = (XX^{\#} - I)\{\, b(k) + c[e(k) + |e(k)|]\,\} \qquad (4\text{-}36)$$

De todas estas ecuaciones obtenemos

$$\|e(k+1)\|^2 = \|e(k)\|^2 + 2ce^{T}(k)(XX^{\#} - I)[e(k) + |e(k)|] + \|c(XX^{\#} - I)[e(k) + |e(k)|]\|^2 \qquad (4\text{-}37)$$

La notación en la ecuación (4-37) se puede ver más claramente definiendo

$$e^{+}(k) = e(k) + |e(k)| \qquad (4\text{-}38)$$

Entonces la ecuación llega a ser

$$\|e(k+1)\|^2 = c + 2ce'(k)(XX^{\#} - I)e^{*}(k) + \|c(XX^{\#} - I)e^{*}(k)\|^2 \qquad (4\text{-}39)$$

Estas ecuaciones se pueden simplificar, primero $(XX^{\#})' (XX^{\#}) = XX^{\#}$ y $W(k) = X^{\#}b(k)$ por lo tanto

$$\begin{aligned} XX^{\#}e(k) &= XX^{\#}[(X_W(k)-b(k)] \\ &= XX^{\#}[\, XX^{\#}b(k)-b(k)] \\ &= 0 \end{aligned}$$

$XX^{\#}$ es simétrica, $e'(k)XX^{\#} = 0$ la ecuación (4-39) llega a ser

$$\|e(k+1)\|^2 = \|e(k)\|^2 - 2ce'(k)e^{*}(k) + \|c(XX^{\#} - I)e^{*}(k)\|^2 \qquad (4\text{-}40)$$

si $e'(k)e^{*}(k) = 1/2\,\|e(k)\|^2$, tenemos

$$\|e(k+1)\|^2 = \|e(k)\|^2 - c\|e^{*}(k)\|^2 + \|c(XX^{\#} - I)e^{*}(k)\|^2 \qquad (4\text{-}41)$$

Como $XX^{\#}$ es simétrica y $(XX^{\#})' (XX^{\#}) = XX^{\#}$, el último término en la ecuación (4-41) se expresa en la siguiente forma

$$\begin{aligned} \|c(XX^{\#} - I)e^{*}(k)\|^2 &= c^2e^{*}(k)(\, XX^{\#} - I)'(XX^{\#} - I)e^{*}(k) \\ &= c^2\|e^{*}(k)\|^2 - c^2e^{*}(k)XX^{\#}\, e^{*}(k) \end{aligned}$$

Sustituyendo esta relación en la ecuación (4-41) se produce la siguiente ecuación.

$$\|e(k)\|^2 - \|e(k+1)\|^2 = c(1 - c)\|e^{*}(k)\|^2 + c^2e^{*}(k)XX^{\#}\, e^{*}(k) \qquad (4\text{-}42)$$

De estas ecuaciones se puede probar la convergencia en el caso separable. Primero, como $XX^{\#}$ es semidefinido como positivo, se tiene que $c^2e^{*}(k)XX^{\#}\, e^{*}(k) \geq 0$. Por lo tanto si $0<c\leq1$, el lado derecho de la ecuación (4-42) es más grande que ó igual a cero. Por consiguiente

$$\|e(k)\|^2 \geq \|e(k+1)\|^2 \qquad (4\text{-}43)$$

y la secuencia $\|e(1)\|^2$, $\|e(2)\|^2$, ... es monótonamente decreciente.

El algoritmo no será determinante en el caso separable hasta que e(k) llegue a ser 0. Del teorema de estabilidad para sistemas discretos se sabe que:

$$\lim_{k\to\infty}\|e(k)\|^2 = 0 \qquad (4\text{-}44)$$

Por consiguiente, estas pruebas de convergencia del algoritmo en el caso separable para k infinito muestra la convergencia para k finito, se puede ver que $X_W(k) = b(k) + e(k)$. En donde b_{min} denota el mínimo componente de b(1) y recalcando que b(k) nunca disminuirá, si e(k) converge para 0 para k infinito, debe entrar a la hiperesfera $\|e(1)\| = b_{min}$ en k infinito, a cualquier punto $X_W(k)>0$. Estos completan la prueba.

La prueba dada arriba no indica el número exacto de pasos requeridos para la convergencia. Implementando el algoritmo, por consiguiente, es necesario comprobar el procedimiento para la ocurrencia de una solución. Una manera de hacerlo es examinar $X_W(k)$ y el vector error después de cada iteración.

Si $X_W(k)>0$ ó e(k) llega a ser 0, una solución ha sido obtenida. De otro modo, si e(k) no llega a ser positivo, las clases no serán linealmente separables y el algoritmo se termina.

4.4 CLASIFICACION MULTICATEGORIA

Hay tres configuraciones multiclase que han sido consideradas; en el primer caso, cada una de las M clases de patrón es separable del resto o una única decisión superficial. Cada una de las M funciones de decisión requeridas para resolver el problema puede ser determinada con la ayuda de alguno de los algoritmos de entrenamiento que se trataron en este capítulo. Por ejemplo para determinar la función de decisión para la i-ésima clase de patrón simplemente se toman en cuenta las dos clases de problemas ω_i y $\omega_{\bar{i}}$, donde $\omega_{\bar{i}}$ denota todas las clases excepto ω_i.

En el segundo caso, cada clase es separable de cada una de las otras clases. Aquí el problema es determinar M(M-1)/2 funciones de decisión. Estas funciones se pueden determinar aplicando cualquiera de los algoritmos presentados anteriormente para todo par de clases de patrones bajo consideración.

En el tercer caso, se asume que existen M funciones de decisión con la propiedad de que $x \in \omega_i$, entonces

$$d_i(X) > d_j(X) \text{ para todo } j \neq i$$

El algoritmo que puede ser usado para este caso es una generalización del algoritmo del perceptrón, y se describe como sigue.

Considerando M clases de patrón $\omega_1, \omega_2, \ldots, \omega_M$ se asume que, en el k-ésimo paso iterativo durante el entrenamiento, un patrón X(k) perteneciente a la clase ω_i es presentado a la máquina. Las M funciones de decisión $d_j[(k)] = W_j^T(k)X(k)$, j= 1, 2, 3, ..., M, son evaluados.
Entonces, si $d_i[X(k)] > d_j[X(k)]$ $i = 1, 2, \ldots, M$; $j \neq i$[7]

Los vectores de peso no son ajustados, es decir ,

$$W_j(k+1) = W_j(k), \quad j=1, 2, \ldots, M$$

De otro modo, para algún

$$d_i[X(k)] \leq d_j[X(k)]$$

Bajo esta condición los siguientes pesos ajustados son hechos:

$$W_i(k+1) = W_i(k) + cX(k)$$
$$W_j(k+1) = W_j(k) - cX(k)$$
$$W_j(k+1) = W_j(k), \; j=1, 2, \ldots, M; \quad j \neq i, \; j \neq j$$

donde c es una constante positiva.

Si las clases son separables bajo el caso 3, se puede demostrar que este algoritmo converge en un número finito de iteraciones para vectores de peso inicialmente arbitrarios $W_i(1)$; i = 1, 2, ..., M.

[7] Tou Julius T. y González Rafael C., op. cit., p. 181

CAPITULO 5

CLASIFICADORES ENTRENABLES. UNA APROXIMACION ESTADISTICA

CAPITULO 5
CLASIFICADORES ENTRENABLES. UNA APROXIMACION ESTADISTICA

5.1 INTRODUCCION

En contraste con el capítulo anterior en donde los algoritmos de clasificación de patrones se derivaron de aproximaciones deterministicas, en este capítulo se emplearán aproximaciones estocásticas para la derivación de algoritmos de clasificación de patrones estadísticos es decir, que todos los algoritmos derivados son el resultado de consideraciones estadísticas.

Los clasificadores diseñados mediante esta aproximación, tienen la capacidad de estimar la información desconocida durante su operación. Las decisiones se toman en base a la información estimada.

Si gradualmente la información estimada se aproxima a la información real, entonces poco a poco se aproximará a la decisión óptima como si toda la información requerida fuera conocida. El proceso en el cual adquiere información necesaria para una decisión durante los sistemas de operación y que a la vez mejoran los sistemas de ejecución es generalmente llamado "aprendizaje supervisado".

La aproximación estocástica es usada para estimar sucesivamente (aprender) parámetros desconocidos en una forma dada de distribución de caracteristicas de cada clase. La formulación estadística de los algoritmos de la clasificación de patrones se centran en la regla de clasificación de Bayes. Como se vió en el capítulo 3, las funciones de decisión de Bayes

$$d_i(X) = p(X|\omega_i)\, p(\omega_i), \quad i = 1,2,\ldots, N \qquad (5\text{-}1)$$

producen una baja probabilidad de error al clasificar un patrón. Cuando es usada la expresión

$$p(X|\omega_i) = p(X|\omega_i)\, p(\omega_i)/p(\omega_i)$$

la ecuación (5-1) se convierte en $d_i(X) = p(\omega_i|X)\, p(X)$, de donde se elimina el término $p(X)$ porque no depende de i, quedando lo siguiente:

$$d_i(X) = p(\omega_i|X), \quad i = 1,2,\ldots, N \qquad (5\text{-}2)$$

donde $p(\omega_i|X)$ es la probabilidad "a posteriori" de que dado un vector de características X pertenezca a la clase ω_i.

La estimación de las densidades $p(\omega_i|X)$ para la implementación de las funciones de decisión de la ecuación (5-2) puede ser formulada en un esquema de aprendizaje iterativo.

En general los pasos para el diseño de un clasificador estadístico mediante aprendizaje ó entrenamiento son los siguientes:

1.- Aproximación de las funciones de densidad de probabilidad (fdp) a posteriori $p(\omega_i|X)$, i=1,2, ..., N de cada clase, a una función discriminante lineal:

$$p(\omega_i|X) = W_i^T X$$

2.- Como hay que entrenar al clasificador, debe existir un conjunto de muestras de pertenencia conocida.

3.- Minimizar un índice de error de clasificación.

4.- La aplicación de un algoritmo recursivo de aprendizaje con el fin de hacer mínimo el índice de error anterior.

5.2 METODOS DE APROXIMACION ESTOCASTICA

"La aproximación estocástica es una técnica recursiva que ha sido desarrollada como una técnica de optimización para ambientes aleatorios. Esta aproximación puede ser usada para estimaciones sucesivas de un parámetro desconocido, cuando debido a la naturaleza estocástica del problema las mediciones están teniendo ciertos errores." [1] Esta técnica garantiza la convergencia del algoritmo, aunque los vectores no sean linealmente separables.

En el capítulo 4 tratamos con funciones criterio deterministicas y en la aproximación estocástica trataremos con funciones estadísticas que generalmente se llaman **funciones regresión**. Se utilizarán los métodos de aproximación estocástica para encontrar la raíz (root) de la función regresión. *"Si esta función regresión representa la derivada de una función criterio formulada correctamente, encontrar la raíz de esta función derivada produce el mínimo de la función criterio."* [2] Se pueden generar algoritmos de aprendizaje iterativo seleccionando ciertos tipos de funciones criterio.

[1] Sankar K. Pal, "Fuzzy mathematical approach to pattern recognition", p. 15
[2] Tou Julius T. y Gonzalez Rafael C., "Pattern Recognition Principles", p. 218

5.2.1 EL ALGORITMO ROBBINS-MONRO

Siendo $g(\omega)$ una función de ω que tiene una raíz sola ω de modo que

$$g(\omega) = 0 \tag{5-3}$$

$$g(\omega) > 0 \quad \text{para } \omega > \omega \tag{5-4}$$

$$g(\omega) < 0 \quad \text{para } \omega < \omega \tag{5-5}$$

Las condiciones (5-3) y (5-4) son asumidas con poca pérdida de generalidad ya que la mayoría de las funciones con una raíz sola no satisfacen estas condiciones, pueden ser establecidas para multiplicar la función por -1. Si $h(\omega)$ representa el valor de ruido de la función regresión $g(\omega)$ tenemos que $g(\omega) - h(\omega)$ representa el error entre el valor verdadero y las observaciones de ruido aleatorio en un valor particular de ω. Lo anterior lo podemos ver en la siguiente figura:

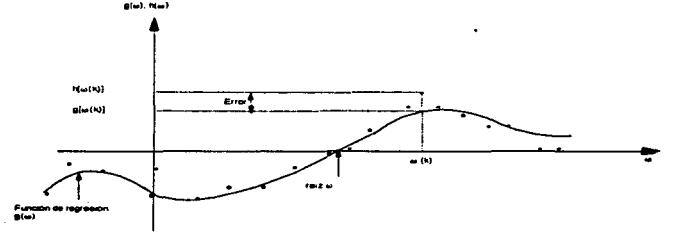


Figura 1. Observaciones de ruido de una función regresión.

Asumimos que:

(a) Las variables aleatorias $h(\omega)$ son imparciales, esto es,

$$E\{h(\omega)\} = g(\omega) \tag{5-6}$$

Lo anterior quiere decir que si hay numerosas observaciones $h(\omega)$ para una determinada ω, el promedio de esas observaciones, dado por el valor esperado debería aproximar $g(\omega)$ en ese punto a medida que el tamaño de la muestra se incrementa.

(b) La varianza de las observaciones aleatorias $h(\omega)$ desde $g(\omega)$ es finita para todos los valores de ω; representando esta varianza mediante:

$$\sigma^2(\omega) = E\{[g(\omega) - h(\omega)]^2\} \quad (5\text{-}7)$$

$$\sigma^2(\omega) < L, \text{ para toda } \omega \quad (5\text{-}8)$$

donde L es una constante positiva finita.

Bajo estas condiciones, el algoritmo Robbins-Monro se puede emplear para estimar sucesivamente la raiz ω de la función $g(\omega)$. *"Si $\omega(1)$ representa la estimación inicial arbitraria de ω, y $\omega(k)$ la estimación en el k-ésimo paso iterativo, el algoritmo Robbins-Monro (R-M) actualiza la estimación de acuerdo a la relación."*[3]

$$\omega(k+1) = \omega(k) - \alpha_k\, h[\omega(k)] \quad (5\text{-}9)$$

donde α_k es miembro de una secuencia de números positivos que satisface las siguientes condiciones:

$$\lim_{k\to\infty} \alpha_k = 0 \quad (5\text{-}10)$$

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad (5\text{-}11)$$

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty \quad (5\text{-}12)$$

Una secuencia que satisface las condiciones anteriores es la secuencia armónica:

$$\{1/k\} = \{1,\ 1/2,\ 1/3,\ \ldots\}$$

Estas tres condiciones (5-10, 5-11 y 5-12) aseguran que el algoritmo de la ecuación (5-9) converge a ω en el sentido del medio cuadrado, representándose de la siguiente forma:

$$\lim_{k\to\infty} \{E\,[(\omega(k) - \omega)^2]\} = 0 \quad (5\text{-}13)$$

La ecuación anterior indica que como el número de iteraciones se aproxima a infinito, la varianza de las estimaciones $\omega(k)$ desde la raiz ω se aproximará a cero, es decir, $\omega(k)$ se aproximará a ω.

[3] Ibidem, p. 220

El algoritmo R-M es extendible directamente al caso multidimensional, teniendo la notación del vector $W = (\omega_1, \omega_2, \ldots, \omega_n, \omega_{n+1})^T$, deseamos encontrar la raíz de una función de regresión $g(\omega)$ de las observaciones de ruido $h(\omega)$. quedando la siguiente relación multidimensional para actualizar la estimación:

$$W(k+1) = W(k) - \alpha_k h[W(k)].$$

5.2.2 ACELERACION DE CONVERGENCIA

Aunque el algoritmo R-M converge hacia la raíz rápidamente existen casos en que no es así ya que k aumenta, los factores de corrección α_k tienen el efecto de disminuir la magnitud de los ajustes con iteraciones sucesivas.

El algoritmo R-M es por lo regular lento para converger, debido a que cualquier secuencia $\{\alpha_k\}$ satisface las ecuaciones (5-10, 5-11 y 5-12), debe disminuir con el incremento de k.

Existe un método eficaz de acelerar la convergencia del algoritmo R-M, el cual consiste en conservar α_k constante durante los pasos en que $h[\omega(k)]$ permanece con el mismo signo. En la siguiente tabla se ilustra este método para $\alpha_k = 1/k$.

k:	1	2	3	4	5	6	7	8	9	10
Signo de $h[\omega(k)]$	+	+	+	-	-	+	-	+	-	-
α_k normal	1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10
α_k acelerada	1	1	1	1/2	1/2	1/3	1/4	1/5	1/6	1/6

5.3 DERIVACION DE ALGORITMOS DE CLASIFICACION DE PATRONES

En este tema se continuará con el mismo formato como en el capítulo anterior, en donde se derivaron los algoritmos determinísticos. En los subtemas siguientes se establecerá el método de aproximación estadística como una aproximación general para la derivación de los algoritmos de clasificación de patrones estadísticos. Se deriva un algoritmo parecido al algoritmo de la técnica de el gradiente de la ecuación (4-22) del capítulo 4, en donde serán comparados los dos algoritmos.

También se derivarán los algoritmos estadísticos de incremento-corrección y el mínimo error cuadrático medio. la derivación de los algoritmos estadísticos por medio de métodos generales

desarrollados posteriormente, estarán limitados por algunas habilidades para especificar criterios de funciones significantes.

5.3.1 ESTIMACION DE FUNCIONES POR METODOS DE APROXIMACION ESTOCASTICA

En este tema aproximaremos la función de decisión de un vector aleatorio X de las muestras X_1, X_2, ..., utilizando métodos de aproximación. "*El aspecto clave en el diseño de los reconocedores estadísticos con aprendizaje, estriba en partir de una expresión concreta para las probabilidades a posteriori $p(\omega_i/x)$ que se van a estimar (ó aprender) basándose en las muestras de entrenamiento*"[4]. El tema principal de este capítulo es la estimación de patrones entrenados de las densidades $p(\omega_i/X) = p(X/\omega_i)$, $p(W_i)/p(X)$ para la implementación de las funciones de decisión de Bayes $d_i(X) = p(\omega_i/X)$, $i = 1, 2, \ldots, M$ clases. La aproximación que se tomará es para expandir estas funciones en un conjunto de funciones básicas conocidas de acuerdo a la relación

$$d_i(X) = p(\omega_i/X) \approx \sum_{j=1}^{k+1} W_i^T \varphi_j(X) = W_i^T \varphi(X) \qquad (5\text{-}14)$$

donde $W_i = (W_{i1}, W_{i2}, \ldots, W_{in}, W_{i,n+1})^T$ es el vector de peso de la i-ésima clase de patrón, y $\varphi(X) = [\varphi_1(X), \varphi_2(X), \ldots, \varphi_k(X), 1]^T$ "*$\{\varphi_i(X), i = 1, \ldots, n\}$ es un conjunto de funciones independientes conocidas de X poseyendo momentos arriba de el sexagesimo momento, y la matriz $E\{p(X)\varphi^T(X)\}$ es definida positiva. El número n denote la memoria necesitada para almacenar la aproximación*"[5]. Se van a considerar las aproximaciones lineales (funciones de densidad a posteriori), es decir:

$$d_i(X) = p(\omega_i/X) \approx W_i^T X \qquad (5\text{-}15)$$

en donde $W_i = (W_{i1}, W_{i2}, \ldots, W_{in}, W_{i,n+1})^T$ y $X_i = (X_{i1}, X_{i2}, \ldots, X_{in}, 1)^T$

Para cada clase se define una clasificación aleatoria, una variable $r_i(X)$, con la siguiente propiedad:

$$r_i(X) = \begin{cases} 1 & \text{si } X \in \omega_i \\ 0 & \text{otro} \end{cases} \qquad (5\text{-}16)$$

Los valores 0 y 1 para $\{p(X)\varphi^T(X)\}$ se seleccionan arbitrariamente.

[4] Maravall Darío, "Reconocimiento de formas y visión artificial", p. 136
[5] Mendel M. Jerry, "A prelude to Neural NetWorks: Adaptative and Learning Systems", p. 364

Aunque no se puede observar a $p(\omega_i/X)$ durante el entrenamiento, aún así, se pueden conocer los valores de $r_i(X)$ durante la fase de entrenamiento.

En el capitulo anterior se desarrolló el criterio de funciones determinísticas que se sustituyeron en el algoritmo de el gradiente para obtener algoritmos de clasificación. En esta sección se seguirá la misma aproximación, excepto que las funciones criterio, serán estadísticas y el algoritmo general es el algoritmo Robbins-Monro. Como una introducción a esta aproximación, tomando en cuenta la función criterio $J(W_i,X) = E\{|r_i(X) - W_i^T(X)|\}$, el mínimo de esta función es cero, y ocurre cuando $W_i^T(X)=r_i(X)$, esto es, el mínimo ocurre cuando el patrón X se clasificó correctamente, esto se da debido a que $r_i(X)$ es una variable de clasificación que se conoce durante el entrenamiento. Por consiguiente, si $W_i^T(X)=r_i(X)$ para todos los patrones de el conjunto entrenado, W_i es capaz de clasificar todos estos patrones correctamente.

Así, $E\{r_i(X)\} = E\{p(\omega_i/X)\}$, $J(W_i,X)$, también se puede expresar como $J(W_i,X) = E\{|p(\omega_i/X) - W_i^T(X)|\}$. Esta expresión muestra que encontrar el mínimo de $J(W_i,X)$ es equivalente a encontrar una aproximación promedio a $p(\omega_i/X)$. *"En otras palabras, la aproximación es tal que el valor esperado de la diferencia entre la función $p(\omega_i/X)$ y su aproximación es cero"*[6]

El principal interés es encontrar el mínimo de una función J(W,X) que es el valor esperado de alguna otra función f(W,X), para esto, se encuentra la raíz de su derivada, así,

$$J(W,X)\ E\{f(W,X)\} \qquad (5\text{-}17)$$

Aplicando la derivada parcial a J(W,X) con respecto a W, se obtiene

$$\partial J(W,X) / \partial W = E\{\partial f(W,X) / \partial W\} \qquad (5\text{-}18)$$

La raíz de $\partial J(W,X) / \partial W$ se puede estimar ahora sucesivamente referenciando al algoritmo R-M con

$$h[W(k)] = \{\partial f(W,X) / \partial W\}_{W=W(k)} \qquad (5\text{-}19)$$

usando la ecuación:

$$W(k+1) = W(k) - \alpha_k h[W(k)]$$

se obtiene la ecuación:

$$W(k+1) = W(k) - \alpha_k\{\partial f(W,X) / \partial W\}_{W=W(k)} \qquad (5\text{-}20)$$

[6] Tou Julius T. y González Rafael C., op. cit., p. 228.

en donde W(1) puede tomar cualquier valor.

Comparando la ecuación (5-20) con el algoritmo general de el gradiente determinístico que se obtuvo en la ecuación (4-22) de el capitulo 4. se tiene la ecuación:

$$W(k+1) = W(k) - c\{\partial f(W,X) / \partial W\}_{W=W(k)}$$

Claramente se ve que tiene varias diferencias con respecto a la ecuación (5-20), como son los incrementos de corrección α_k y c, en los términos de las derivadas parciales. La función criterio J(W,X) aparece directamente en el algoritmo determinístico, debido a que J(W,X) se observa directamente en el caso determinístico pero no es observable en el caso estadístico, el algoritmo de la ecuación (5-20) emplea la función f(W,X). Otra diferencia importante es que el algoritmo estadístico buscará una aproximación para el clasificador de Bayes, mientras que la parte determinística no tiene esa capacidad. El algoritmo estadistico convergerá a la aproximación sin hacer caso de si las clases son ó no estrictamente separables, mientras que el algoritmo determinístico simplemente oscila en situaciones no separables. Se garantiza la convergencia de el algoritmo estadístico, pero tiene su gran desventaja, la lentitud con la que generalmente se alcanza esta convergencia.

5.3.2 ALGORITMO INCREMENTO-CORRECCION

Este algoritmo es parecido al algoritmo de el perceptrón que se puede derivar considerando la función criterio mencionada en el tema anterior.

$$J(W_i,X) \; E\{|r_i(X) - W_i^T(X)|\} \qquad (5\text{-}21)$$

donde

$$r_i(X) = \begin{cases} 1 & \text{si } X \in \omega_1 \\ 0 & \text{otro} \end{cases}$$

Cuando los patrones son clasificados correctamente se logra el mínimo de $J(W_i,X)$ con respecto a W_i.

La derivada parcial de J con respecto a W_i, esta dada por

$$\partial J / \partial W_i = E\{-X \; sgn[r_i(X) - W_i^T(X)]\} \qquad (5\text{-}22)$$

en donde

sgn es la función signo

$$sgn(\phi) = \begin{cases} 1 & \text{si } \phi > 0 \\ 0 & \text{si } \phi < 0 \end{cases}$$

Así $h(W_i) = -X\ \mathrm{sgn}[r_i(X) - W_i^T(X)]$ y sustituyendo en el algoritmo general de la ecuación (5-20) se genera la ecuación:

$$W_i(k+1) = W_i(k) + \alpha_k X(k)\ \mathrm{sgn}\{r_i[X(k)] - W_i^T(k)X(k)\} \qquad (5\text{-}23)$$

en donde $W_i(1)$ toma cualquier valor. Utilizando la definición de la función signo, la ecuación (5-23) se expresa de la siguiente forma equivalente

$$W_i(k+1) = \begin{cases} W_i(k) + \alpha_k X(k) & \text{si } W_i^T(X)X(k) < r_i[X(k)] \\ W_i(k) - \alpha_k X(k) & \text{si } W_i^T(X)X(k) \geq r_i[X(k)] \end{cases} \qquad (5\text{-}24)$$

Este algoritmo hace un ajuste en el vector de peso en cada paso, en contraste con el algoritmo del perceptrón, en donde sólo se hace una corrección cuando el patrón se clasifica mal. El algoritmo de la ecuación (5-23) ó (5-24) deriva su nombre de el hecho que las correcciones son proporcionales al incremento α_k.

El procedimiento iterativo de la ecuación (5-23) ó (5-24) se dice que ha convergido para una solución libre de error cuando todos lo patrones entrenados de ω_i, $i = 1,2,\ldots, M$, han sido clasificados correctamente.

Cuando las clases consideradas no son las funciones de decisión especificadas, se asegura que la solución convergerá en el límite para el valor absoluto aproximado de $p(\omega_i|X)$, como se indicó por la función criterio en la ecuación (5-21).

En el caso de dos clases, se puede evaluar directamente el vector de peso de la superficie de separación. Para este caso, la ecuación (5-23) se expresa como

$$W(k+1) = W(k) + \alpha_k X(k)\ \mathrm{sgn}\{r[X(k)] - W^T(k)X(k)\} \qquad (5\text{-}25)$$

En donde $W(1)$ toma cualquier valor, cuando se utiliza la ecuación (5-25) se hace la suposición que W es el vector de peso de la clase ω_1, así que $r[X(k)] = 1$ si $X(k) \in \omega_1$ y $r[X(k)] = 0$ si $X(k) \in \omega_2$. De la siguiente ecuación

si $p(\omega_1|X) > 1/2$, asigna X a ω_1
si $p(\omega_1|X) < 1/2$, asigna X a ω_2

Por lo tanto se tiene la regla de decisión:

si $p(\omega_1|X) = W^T X > 1/2$, asigna X a ω_1

si $p(\omega_1|x) = W^T X < 1/2$, asigna X a ω_2 (5-26)

en donde $W^T X$ representa una aproximación a $p(\omega_1|X)$.

De la expresión (5-26) el algoritmo de dos clases ha convergido para una solución libre de error cuando $W^T X > 1/2$ para todos los patrones de ω_1 y $W^T X < 1/2$ para todos los patrones de ω_2. Esto también es válido en el caso de usar algoritmos multiclase para obtener dos funciones de decisión, $d_1(X) = W_1^T X$ y $d_2(X) = W_2^T X$. Entonces quiere decir que se puede obtener una sola función de decisión definiendo $d(X) = d_1(X) - d_2(X)$.

Por ejemplo. Se va a plantear el aprendizaje de una sola función discriminante explicando el algoritmo de incremento-corrección, que se aplicó en esta sección para la estimación de la alternativa de las funciones de decisión $d_i(X) = p(\omega_i|X)$. Se utilizará una secuencia no alternada de las muestras de aprendizaje, es decir, se presentarán en primer lugar todas las muestras de ω_1 y en seguida las de ω_2, después las de ω_1, y así sucesivamente hasta obtener una solución que clasifique correctamente todas las muestras de aprendizaje.

ω_1 : {(0,0,0,1)', (1,0,0,1)', (1,0,1,1)', (1,1,0,1)'}
ω_2 : {(0,0,1,1)', (0,1,0,1)', (0,1,1,1)', (1,1,1,1)'}

$W(2) = W(1) + \alpha_1 X(1)\ \text{sgn}\{r[X(1)] - W^T(1)X(1)\}$

$$= 0 + X(1) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

En el siguiente paso, $X(2) = (1,0,0,1)'$, $\alpha_2 = 1/2$, y como $X(2) \in \omega_1$, $r[X(2)] = 1$. Por lo tanto,
$W(3) = W(2) + \alpha_2 X(2)\ \text{sgn}\{r[X(2)] - W^T(2)X(2)\}$

$$= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + 1/2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \text{sgn}\{0\}$$

$$= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - 1/2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1/2 \\ 0 \\ 0 \\ 1/2 \end{pmatrix}$$

$X(3) = (1,0,1,1)'$, $\alpha_3 = 1/3$, y $r[X(3)] = 1$, así que

$$W(4) = W(3) + 1/3\, X(3)\ \text{sgn}\ \{1\}$$

$$= \begin{pmatrix} -1/2 \\ 0 \\ 0 \\ 1/2 \end{pmatrix} + \begin{pmatrix} 1/3 \\ 0 \\ 1/3 \\ 1/3 \end{pmatrix} = \begin{pmatrix} -1/6 \\ 0 \\ 0 \\ 5/6 \end{pmatrix}$$

continuando de este modo y probando después de cada iteración para ver si el nuevo vector de peso se clasificó correctamente, se encuentra que el algoritmo converge en k=15 produciendo el vector de peso

$$W = \begin{pmatrix} 0.233 \\ -0.239 \\ -0.216 \\ 0.619 \end{pmatrix}$$

Para calcular la ecuación de la frontera de decisión se debe cumplir que $W^TX>0.5$ ó $W^TX<0.5$. Así, la frontera está dada por $W^TX=0.5$ ó $W^TX-0.5=0$, entonces la función definitiva tendrá la forma

$$0.233X_1 - 0.239X_2 - 0.216X_3 + 0.119=0$$

esta frontera de decisión se ilustra en la figura 2.

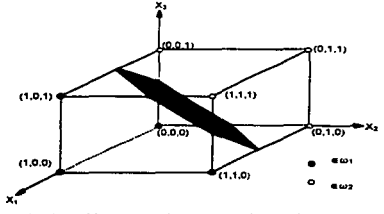


Figura 2. Frontera de decisión determinada por el algoritmo incremento-corrección.

5.3.3 ALGORITMO DEL MINIMO ERROR CUADRATICO MEDIO

Se pueden definir varios índices del error, por ejemplo el índice del error en módulo ó el índice del error cuadrático. Puesto que se están manejando variables aleatorias, tanto r, como la aproximación $W_i^T X$ de $p(\omega_i|X)$ son variables aleatorias. El criterio del mínimo error cuadrático medio (LMSE) se emplea para derivar otros algoritmos entrenados. Considerando la función criterio

$$J(W_i, X) = 1/2\ E\{[r_i(X) - W_i^T X]^2\} \qquad (5\text{-}27)$$

La minimización de los índices anteriores es similar a la minimización de los índices de error que se manejaron en la hipótesis determinística.

Tomando la derivada parcial de J con respecto a W_i se tiene:

$$\partial J/\partial W_i = E\{-X[r_i(X) - W_i^T X]\} \qquad (5\text{-}28)$$

Como el argumento de E es constante para X(k), puede eliminarse del operador esperanza matemática, por lo tanto, la ecuación $h(W_i) = -X[r_i(X) - W_i^T X]$ se sustituye en el algoritmo general de la ecuación (5-20), obteniéndose la actualización de los coeficientes de las funciones discriminantes ó frontera de decisión, resultando la ecuación

$$W_i(k+1) = W_i(k) + \alpha_k X(k)\{r_i[X(k)] - W_i^T(k)X(k)\} \qquad (5\text{-}29)$$

donde $W_i(1)$ es arbitraria y $r_i[X(k)]=1$ ó 0, ésta depende de si X(k) pertenece ó no a la clase ω_i. Este algoritmo también hace una corrección en W_i en cada paso de la iteración, y las magnitudes de la corrección son diferentes a las de el algoritmo derivado en el tema anterior por los factores $\{r_i[X(k)] - W_i^T(k)X(k)\}$. *"El algoritmo LMSE converge para una solución que minimiza la ecuación (5-27) si las siguientes condiciones se satisfacen:*

1. *α_k satisface la ecuación (5-17).*
2. *$E\{XX^T\}$ y $E\{XX^T)^2\}$ debe existir y ser definido positivo.*
3. *$E\{X\, p(\omega_i|X)\}$ y $E\{XX^T X\, p(\omega_i|X)\}$ debe existir."*[7]

Para el caso de dos clases la ecuación (5-29) se expresa de la forma

$$W(k+1) = W(k) + \alpha_k X(k)\{r[X(k)] - W^T(k)X(k)\} \qquad (5\text{-}30)$$

en donde W(1) es arbitraria.

[7] Ibidem, p. 233.

Antes de realizar un ejemplo es conveniente hacer un comentario. En primer lugar, *"se trata de un proceso de aprendizaje multiclase; es decir, se actualizan simultáneamente todas las funciones discriminantes W_1, W_2, ... W_N para cada muestra de entrenamiento."*[8]

Por ejemplo: Utilizando el ejemplo del tema anterior, y aplicando el algoritmo LMSE. Tenemos que los patrones aumentados son:

ω_1 : {(0,0,0,1)', (1,0,0,1)', (1,0,1,1)', (1,1,0,1)'}
ω_2 : {(0,0,1,1)', (0,1,0,1)', (0,1,1,1)', (1,1,1,1)'}

La decisión del clasificador resulta ser:
$W(1) = 0$, $\alpha_k = 1/k$ y $X(1) = (0,0,0,1)'$

y usando el algoritmo de la ecuación (5-17) se produce,

$$W(2) = W(1) + \alpha_1 X(1)\,[1 - W^T(1)X(1)]$$

$$= 0 + X(1) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

En el siguiente paso, $X(2) = (1,0,0,1)'$, $\alpha_2 = 1/2$, y $r[X(2)] = 1$. Por lo tanto,

$$W(3) = W(2) + \alpha_2 X(2)\,[1 - W^T(2)X(2)]$$

$$= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + 0/2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Se verifica fácilmente que $W(5)=W(4)=W(3)$. En el siguiente paso $X(5)=(0,0,1,1)'$, $\alpha_5=1/5$, y $X(5) \in \omega_2$, $r[x(5)]=0$.

$$W(6) = W(5) + \alpha_5 X(5)\,[0 - W^T(5)X(5)]$$

$$= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - 1/5 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -1/5 \\ 4/5 \end{pmatrix}$$

[8] Marsvall Dario, op. cit., p. 143.

y así sucesivamente, el algoritmo LMSE convergerá en W=19 que da lugar al vector de peso

$$W = \begin{bmatrix} 0.135 \\ -0.238 \\ -0.305 \\ 0.721 \end{bmatrix}$$

La función discriminante ó frontera de decisión está dada por $W^T X - 0.5 = 0$ ó

$$0.135X_1 - 0.238X_2 - 0.305X_3 + 0.221 = 0$$

Los patrones y su superficie de decisión se muestran en la figura 3.

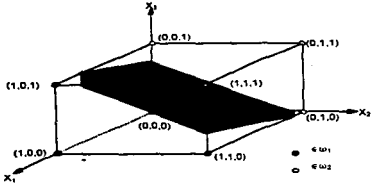


Figura 3. Frontera de decisión determinada por el algoritmo LMSE.

CAPITULO 6

APLICACION

CAPITULO 6
APLICACION

6.1 DESCRIPCION DE LA APLICACION

En el presente capítulo se representa por medio de un programa un ejemplo citado en forma general en el libro "Pattern Recognition Principles", en el cual se aplica la fórmula del clasificador bayesiano que representa la función de decisión de Bayes para patrones normales vista en el capítulo 3. Su propósito es clasificar un vector de entrada X compuesto de 3 elementos X_1, X_2 y X_3 que representan una coordenada en el plano tridimensional en alguna de las dos clases existentes representada por W_1 y W_2.

Para poder clasificar una entrada se tienen que obtener las funciones de decisión de cada clase dadas mediante la fórmula del clasificador bayesiano:

$$d_i(X) = \ln p(W_i) - \tfrac{1}{2} \ln |C_i| - \tfrac{1}{2} [(X-m_i)^T C_i^{-1} (X-m_i)] \qquad i = 1, 2, \ldots, M$$

Los elementos de la fórmula anterior son:

- $p(W_i)$ »» Probabilidad de que el elemento pertenezca a la clase W_i
- $|C_i|$ »» Determinante de la matriz de covarianza
- X »» Vector de entrada
- m_i »» Vector medio de la clase W_i
- C_i^{-1} »» Matriz de covarianza inversa
- T »» Representa la traspuesta de un vector ó una matriz

El programa se realizó en Visual Basic y lleva por nombre CLASIFI.VBP. Funciona de la siguiente manera: primero se despliega una pantalla en la que el usuario debe insertar el vector de entrada X compuesto por 3 caracteres numéricos del 0 al 9, si no se introducen así, se despliega un mensaje de error y se tiene que capturar nuevamente la información. Posteriormente se introduce el valor de probabilidad para la clase 1, ésta debe ser mayor que 0 y menor ó igual que 99, enseguida se introduce la probabilidad para la clase 2; tomando en cuenta que la suma de ambas deben ser menor ó igual que 100, si no se cumplen estas condiciones se despliega un mensaje de error y se deben insertar los valores otra vez.

El programa tiene almacenadas 4 muestras de cada clase que servirán para determinar a qué clase pertenece el vector de entrada. Dichas muestras están representadas en un dibujo tridimensional que se localiza en la parte superior derecha de la pantalla.

Para que se empiecen a realizar los cálculos necesarios para obtener las funciones de decisión $d_1(X)$ y $d_2(X)$ el usuario tiene que dar un click con el ratón en el botón CLASIFICAR. Entonces se desplegarán los valores de d_1 y d_2 y aparecerá un letrero que indica a qué clase pertenece el vector de entrada, dependiendo ésta del valor que sea mayor, esto es, si $d_1 > d_2$ se señalará por medio de un indicador (circulo relleno) y pertenecerá a la clase 1; si $d_2 > d_1$ se señalará por medio de un indicador (circulo relleno) y pertenecerá a la clase 2. De acuerdo al resultado obtenido el usuario puede verificar en el dibujo tridimensional si éste concuerda con la posición del vector de entrada.

Si el usuario no sabe cómo utilizar el programa existe un botón llamado AYUDA en el que se despliega una ventana que contiene las instrucciones de su funcionamiento, para salir de esta ventana se selecciona el botón llamado SALIR y automáticamente regresa a la ventana principal para que comience a capturar los valores.

Para abandonar el programa se debe seleccionar el botón llamado SALIDA.

A continuación se anexa el código fuente del programa, una impresión de las ventanas en las que se capturan los valores y un ejemplo que contenga los resultados.

6.2 PROGRAMA FUENTE

FRMCLASIFICADOR

```
Dim W11( ) As Integer
Dim W12( ) As Integer
Dim W13( ) As Integer
Dim W14( ) As Integer
Dim W21( ) As Integer
Dim W22( ) As Integer
Dim W23( ) As Integer
Dim W24( ) As Integer
Dim X( ) As Integer

Private Sub A_Change (Index As Integer)
        Picture1.Visible = False
        Picture2.Visible = False
        If Index < 2 Then
                A(Index + 1) = ""
                A(Index + 1).SetFocus
        Else
                Picture1.Visible = False
                Picture2.Visible = False
                prob1.SetFocus
        End If
End Sub

Private Sub A_Click (Index As Integer)
        FRMCLASIFICADOR.Cls
        A(0).Text = ""
        A(1).Text = ""
        A(2).Text = ""
        prob1.Text = ""
        prob2.Text = ""
End Sub

Private Sub A_KeyPress (Index As Integer, KeyAscii As Integer)
        FRMCLASIFICADOR.Cls
End Sub

Private Sub CMDAYUDA_Click ( )
        FRMCLASIFICADOR.Hide
        FRMAYUDA.Show
End Sub

Private Sub CMDCLASIFICA_Click ( )
        Cls

        Dim m0( ) As Double, m00( ) As Double
        Dim ca( ), cb( ), cc( ), cd( ), adic( ), m1( ), c1( ), cofc1( ), invc1( ), resta1( ), prod1( )
        Dim m2( ), c2( ), cofc2( ), invc2( ), resta2( ), prod2( )
```

```
        ReDim X(2), m0(2), m00(2)
        ReDim W11(2), W12(2), W13(2), W14(2)
        ReDim W21(2), W22(2), W23(2), W24(2)
        ReDim ca(2,2), cb(2,2), cc(2,2), cd(2,2), adic(2,2), m1(2,2), c1(2,2), cofc1(2,2), invc1(2,2), resta1(2), prod1(2)
        ReDim m2(2,2), c2(2,2), cofc2(2,2), invc2(2,2), resta2(2), prod2(2)

    Dim suma As Double, mult As Double, detc1 As Double, detc2 As Double, det1 As Double, det2 As Double,
suma1 As Double
    Dim suma2 As Double, res1 As Double, res2 As Double, pw1 As Double, pw2 As Double, d1 As Double, d2 As
Double, p1 As Double, p2 As Double

        Const Ni = 0.25
        d1 = 0
        d2 = 0
        p1 = 0
        p2 = 0

        W11(0) = 0                          'SE CARGAN LAS 4 MUESTRAS DE LA CLASE 1.
        W11(1) = 0
        W11(2) = 0
        W12(0) = 1
        W12(1) = 0
        W12(2) = 1
        W13(0) = 1
        W13(1) = 0
        W13(2) = 0
        W14(0) = 1
        W14(1) = 1
        W14(2) = 0

        W21(0) = 1                          'SE CARGAN LAS 4 MUESTRAS DE LA CLASE 2.
        W21(1) = 1
        W21(2) = 1
        W22(0) = 0
        W22(1) = 0
        W22(2) = 1
        W23(0) = 0
        W23(1) = 1
        W23(2) = 1
        W24(0) = 0
        W24(1) = 1
        W24(2) = 0

            'ALMACENAR ENTRADAS DEL USUARIO

    Cls
    Dim prueba
    For j = 0 To 2
            prueba = 0
            prueba = A(j).Text
            If IsNumeric(prueba) Then
                    X(j) = prueba
            Else
                    MsgBox "sólo se permiten números del 0 al 9 en el vector de entrada"
            End If
    Next j

    p1 = prob1.Text
    p2 = prob2.Text
```

```
If (p1 <= 0 Or p1 > 99) Or (p2 <= 0 Or p2 > 99) Or (p1 + p2 > 100) Then
        MsgBox "Las probabilidades que dió son incorrectas"
Else

        'CALCULAR VECTOR MEDIO DE LA CLASE 1

For j = 0 To 2
        suma = 0
        suma = W11(j) + W12(j) + W13(j) + W14(j)
        m0(j) = Ni * suma
Next j

        'CALCULAR VECTOR MEDIO DE LA CLASE 2

For j = 0 To 2
        suma = 0
        suma = W21(j) + W22(j) + W23(j) + W24(j)
        m00(j) = Ni * suma
Next j

        'CALCULAR MATRIZ DE COVARIANZA DE LA CLASE 1

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W11(j) * W11(i)
                ca (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W12(j) * W12(i)
                cb (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W13(j) * W13(i)
                cc (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W14(j) * W14(i)
                cd (j, i) = Ni * mult
        Next i
Next j
```

```
For j = 0 To 2
        For i = 0 To 2
                adic(j, i) = ca(j, i) + cb(j, i) + cc(j, i) + cd(j, i)
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                m1(j, i) = m0(j) * m0(i)
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                c1(j, i) = adic(j, i) - m1(j, i)
        Next i
Next j

        'CALCULAR MATRIZ DE COVARIANZA DE LA CLASE 2

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W21(j) * W21(i)
                ca (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W22(j) * W22(i)
                cb (j, i) = Ni *.mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W23(j) * W23(i)
                cc (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                mult = 0
                mult = W24(j) * W24(i)
                cd (j, i) = Ni * mult
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                adic(j, i) = ca(j, i) + cb(j, i) + cc(j, i) + cd(j, i)
        Next i
Next j
```

```
For j = 0 To 2
        For i = 0 To 2
                m2(j, i) = m00(j) * m00(i)
        Next i
Next j

For j = 0 To 2
        For i = 0 To 2
                c2(j, i) = adic(j, i) - m2(j, i)
        Next i
Next j
```

'CALCULAR EL DETERMINANTE DE LA CLASE 1

```
detc1 = (c1(0, 0) * c1(1, 1) * c1(2, 2)) - (c1(0, 0) * c1(1, 2) * c1(2, 1)) - (c1(0, 1) * c1 (1, 0) * c1 (2, 2)) + (c1(0, 1) * c1(1, 2) * c1(2, 0)) + (c1(0, 2) * c1(1, 0) * c1(2, 1)) - (c1(0, 2) * c1(1, 1) * c1(2, 0))
```

'CALCULAR EL DETERMINANTE DE LA CLASE 2

```
detc2 = (c2(0, 0) * c2(1, 1) * c2(2, 2)) - (c2(0, 0) * c2(1, 2) * c2(2, 1)) - (c2(0, 1) * c2 (1, 0) * c2 (2, 2)) + (c2(0, 1) * c2(1, 2) * c2(2, 0)) + (c2(0, 2) * c2(1, 0) * c2(2, 1)) - (c2(0, 2) * c2(1, 1) * c2(2, 0))
```

'CALCULAR MATRIZ DE COFACTORES DE LA CLASE 1

```
cofc1(0, 0) = (c1(1, 1) * c1(2, 2)) - (c1(2, 1) * c1(1, 2))
cofc1(0, 1) = - ((c1(1, 0) * c1(2, 2)) - (c1(2, 0) * c1(1, 2)))
cofc1(0, 2) = (c1(1, 0) * c1(2, 1)) - (c1(2, 0) * c1(1, 1))
cofc1(1, 0) = - ((c1(0, 1) * c1(2, 2)) - (c1(2, 1) * c1(0, 2)))
cofc1(1, 1) = (c1(0, 0) * c1(2, 2)) - (c1(2, 0) * c1(0, 2))
cofc1(1, 2) = - ((c1(0, 0) * c1(2, 1)) - (c1(2, 0) * c1(0, 1)))
cofc1(2, 0) = (c1(0, 1) * c1(1, 2)) - (c1(1, 1) * c1(0, 2))
cofc1(2, 1) = - ((c1(0, 0) * c1(1, 2)) - (c1(1, 0) * c1(0, 2)))
cofc1(2, 2) = (c1(0, 0) * c1(1, 1)) - (c1(1, 0) * c1(0, 1))
```

'CALCULAR MATRIZ DE COFACTORES DE LA CLASE 2

```
cofc2(0, 0) = (c2(1, 1) * c2(2, 2)) - (c2(2, 1) * c2(1, 2))
cofc2(0, 1) = - ((c2(1, 0) * c2(2, 2)) - (c2(2, 0) * c2(1, 2)))
cofc2(0, 2) = (c2(1, 0) * c2(2, 1)) - (c2(2, 0) * c2(1, 1))
cofc2(1, 0) = - ((c2(0, 1) * c2(2, 2)) - (c2(2, 1) * c2(0, 2)))
cofc2(1, 1) = (c2(0, 0) * c2(2, 2)) - (c2(2, 0) * c2(0, 2))
cofc2(1, 2) = - ((c2(0, 0) * c2(2, 1)) - (c2(2, 0) * c2(0, 1)))
cofc2(2, 0) = (c2(0, 1) * c2(1, 2)) - (c2(1, 1) * c2(0, 2))
cofc2(2, 1) = - ((c2(0, 0) * c2(1, 2)) - (c2(1, 0) * c2(0, 2)))
cofc2(2, 2) = (c2(0, 0) * c2(1, 1)) - (c2(1, 0) * c2(0, 1))
```

'CALCULAR MATRIZ INVERSA DE LA CLASE 1

```
detc1 = 1 / detc1
For j = 0 To 2
        For i = 0 To 2
                invc1(j, i) = detc1 * cofc1(j, i)
        Next i
Next j
```

```
'CALCULAR MATRIZ INVERSA DE LA CLASE 2

det2 = 1 / detc2
For j = 0 To 2
        For i = 0 To 2
                invc2(j, i) = det2 * cofc2(j, i)
        Next i
Next j


'CALCULAR FORMULA DEL CLASIFICADOR BAYESIANO PARA LA CLASE 1

For i = 0 To 2
        resta1(i) = X(i) - m0(i)
Next i

For j = 0 To 2
suma1 = 0
        For i = 0 To 2
                suma1 = suma1 + (invc1(j, i) * resta1(i))
        Next i
                prod1(j) = suma1
Next j

res1 = 0
For i = 0 To 2
        res1 = res1 + (resta1(i) * prod1(i))
Next i

pw1 = p1 / 100
d1 = Log (pw1) - (0.5 * Log (detc1)) - (0.5 * res1)
Print: Print: Print "                                                   d1=  "; d1


'CALCULAR FORMULA DEL CLASIFICADOR BAYESIANO PARA LA CLASE 2

For i = 0 To 2
        resta2(i) = X(i) - m00(i)
Next i

For j = 0 To 2
suma2 = 0
        For i = 0 To 2
                suma2 = suma2 + (invc2(j, i) * resta2(i))
        Next i
                prod2(j) = suma2
Next j

res2 = 0
For i = 0 To 2
        res2 = res2 + (resta2(i) * prod2(i))
Next i

pw2 = p2 / 100
d2 = Log (pw2) - (0.5 * Log (detc2)) - (0.5 * res2)
Print: Print: Print"                                                    d2=  "; d2
```

```
                    'DESPLIEGA RESULTADOS DE CLASIFICACION

        Print: Print
        If d1 > d2 Then
                Picture1.Visible = True
                Picture1.Circle = (15, 15), 6
                Print "                         El patrón de entrada pertenece a la clase 1 porque "
                Print "                         la función de decision d1 > la función de decisión d2"
        Else
                If d2 > d1 Then
                        Picture2.Visible = True
                        Picture2.Circle = (15, 15), 6
                        Print "                         El patrón de entrada pertenece a la clase 2 porque "
                        Print "                         la función de decisión d2 > la función de decisión d1"
                End If
        End If
        End If
        End Sub


Private Sub CMDSALIR_Click( )
        End
End Sub


Private Sub Form_Load ( )
        FRMCLASIFICADOR.Cls
End Sub


Private Sub prob1_Change( )
'       prob2 = ""
'       prob2.SetFocus
End Sub


Private Sub prob1_Click( )
        FRMCLASIFICADOR.Cls
        Picture1.Visible = False
        Picture2.Visible = False
End Sub


Private Sub prob1_KeyPress(KeyAscii As Integer)
        FRMCLASIFICADOR.Cls
        Picture1.Visible = False
        Picture2.Visible = False
End Sub


Private Sub prob2_Click( )
        FRMCLASIFICADOR.Cls
        Picture1.Visible = False
        Picture2.Visible = False
End Sub
```

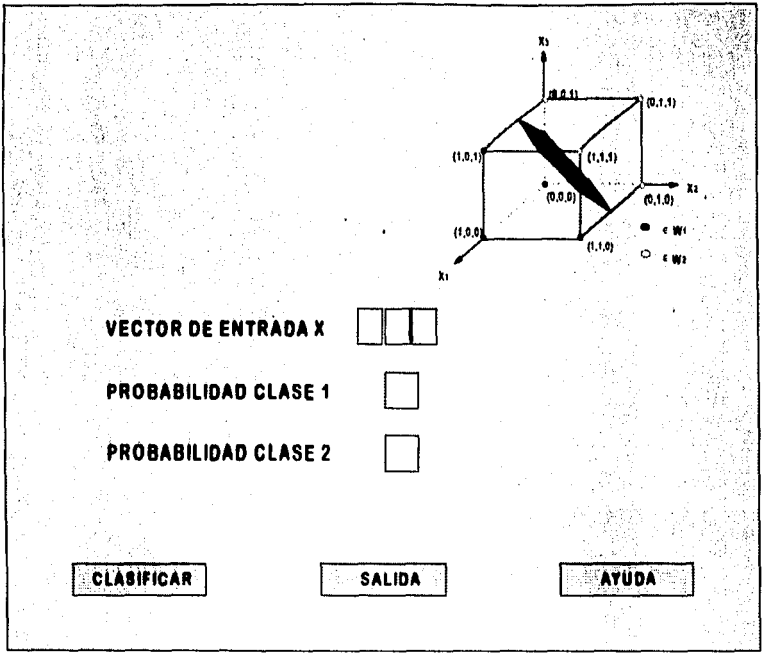
```
Private Sub prob2_KeyPress(KeyAscii As Integer)
        FRMCLASIFICADOR.Cls
        Picture1.Visible = False
        Picture2.Visible = False
End Sub


FRMAYUDA

Private Sub CMDSALIDA_Click( )
        FRMAYUDA.Hide
        FRMCLASIFICADOR.Show
End Sub


Private Sub Form_Activate( )
        Print: Print: Print
        Print "                                                    CLASIFICADOR BAYESIANO"
        Print: Print
        Print "          El programa clasifica a qué clase pertenece un vector de entrada X compuesto de 3 caracteres"
        Print "     numéricos del 0 al 9, si no se teclea correctamente se despliega un mensaje de error y se tiene "
        Print "     que capturar nuevamente la información. Posteriormente se deben introducir las probabilidades para"
        Print "     cada una de las dos clases W1 y W2, estas deben ser mayores que 0 y menores ó iguales que 99 y la"
        Print "     suma de ambas debe ser menor o igual que 100, si no se cumple lo anterior se despliega un mensaje"
        Print "     y se teclearán otra vez."
        Print: Print
        Print "          Para que el programa empiece a realizar los cálculos necesarios y determine a qué clase perte-"
        Print "     nece el vector de entrada, el usuario debe de oprimir el boton llamado CLASIFICA."
        Print: Print
        Print "          La clasificación obtenida se realiza mediante la formula del clasificador bayesiano:"
        Print: Print: Print
        Print "                              di(X) = ln p(Wi) - ½ ln |Ci| - ½ [ (X-mi)' invzCi (X-mi) ],          i= 1, 2, ..., M"
End Sub
```

6.3 FORMATO DE PANTALLAS



PANTALLA PRINCIPAL

CLASIFICADOR BAYESIANO

El programa clasifica a qué clase pertenece un vector de entrada X compuesto de 3 caracteres numéricos del 0 al 9, si no se teclea correctamente se despliega un mensaje de error y se tiene que capturar nuevamente la información. Posteriormente se deben introducir las probabilidades para cada un de las dos clases W1 y W2, éstas deben ser mayor que 0 y menores o igual que 99 y la suma de ambas debe ser menor ó igual que 100; si no se cumple lo anterior se despliega un mensaje y se teclearán otra vez.

Para que el programa empiece a realizar los cálculos necesarios y determine a qué clase pertenece el vector de entrada, el usuario debe oprimir el botón llamado CLASIFICAR.

La clasificación obtenida se realiza mediante la fórmula del clasificador bayesiano:

$$di(X) = \ln p(Wi) - 1/2 \ln |Ci| - 1/2 [(X-mi)' invzCi (X-mi)], \qquad i= 1, 2, ...,$$

SALIR

PANTALLA DE AYUDA

6.4 EJEMPLO

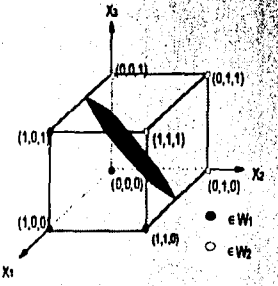
Aquí se muestran dos ejemplos de la ejecución del programa. En el primero se insertan los datos de tal forma que el vector de entrada sea clasificado para la clase 1, y en el segundo caso los datos que se introducen clasifican al vector de entrada para la clase 2.

En las siguientes páginas se muestran las pantallas de la ejecución, la primera corresponde a la clase 1 y la segunda a la clase 2.

d1= 0.579441541679836

d2= -3.42055845845832016

El patrón de entrada pertenece a la clase 1 porque la función decisión d1 > la función de decisión d2



VECTOR DE ENTRADA X 1 0 1

PROBABILIDAD CLASE 1 50

PROBABILIDAD CLASE 2 50

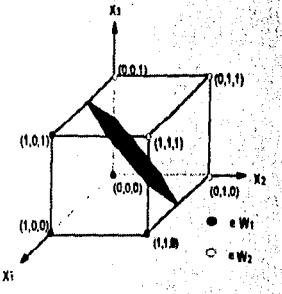
CLASIFICAR SALIDA AYUDA

EJEMPLO PARA LA CLASE 1

d1= -11.933840820862

d2= 0.841805806147327

El patrón de entrada pertenece a la clase 2 porque la función decisión d2 > la función de decisión d1



VECTOR DE ENTRADA X 0 1 1

PROBABILIDAD CLASE 1 30

PROBABILIDAD CLASE 2 65

CLASIFICAR SALIDA AYUDA

EJEMPLO PARA LA CLASE 2

CONCLUSIONES

CONCLUSIONES

Con la investigación realizada pudimos damos cuenta que el Reconocimiento de Patrones es una área relativamente nueva, ya que en las últimas décadas ha tenido un gran auge e importantes avances dentro de la Informática.

El Reconocimiento de Patrones ha llegado a ser reconocido como un factor importante en el diseño de sistemas de información computarizados modernos, prueba de ello son las aplicaciones que se han realizado con éxito en diferentes áreas como: medicina (análisis de tejido celular, análisis de electrocardiogramas "ECG's", etc.), reconocimiento de lenguajes, criminología (identificación de huellas digitales y fotografías, entre otras), robótica, etc. Por lo que dentro de poco tiempo las aplicaciones de éste serán en áreas que no han sido consideradas aún.

Los elementos probabilísticos son un factor importante que intervienen en el Reconocimiento de Patrones para diseñar sistemas que clasifiquen con el mínimo de errores a un patrón de entrada dado, ya que en ciertas aplicaciones como la identificación de huellas digitales de delincuentes ó en los análisis de tejidos celulares infectados por algún virus, se debe tener la mayor precisión y exactitud posible en los resultados puesto que, algún error significativo podría traer graves consecuencias, por ésto se deben practicar continuamente pruebas para comprobar la efectividad del sistema en estos casos.

BIBLIOGRAFIA

BIBLIOGRAFIA

1. Barnett Victor David, Comparative statistical inference, Ed. John Wiley & Sons London.

2. Becker Peter W., Recognition of patterns (using the frequencies of occurrence of binary words), Ed. Springer-Verlag, Wien New York.

3. Black W. J., Intelligent knowledge based systems an introduction, Ed. Van Nostrand Renhold.

4. Box P. George E. and Tiao George C., Bayesian inference in statistical analysis, Ed. Addison Wesley Publishing Company.

5. C. H. Chen, Pattern recognition and artificial intelligence, Ed. Academic Press, Inc.

6. Edward Patrick A., Artificial intelligence with statistical pattern recognition, Ed. Prentice Hall.

7. Feller William, Introducción a la teoría de probabilidades y sus aplicaciones, Vol. I, Ed. Limusa.

8. Lipschutz Seymur Ph. D, Teoría y problemas de probabilidad, Ed. Mc Graw Hill, Serie Schaum.

9. Maisel Louis, Probabilidad y estadística, Ed. Fondo Educativo Interamericano, S.A.

10. Maravall Darío y Gómez-Allende, Reconocimiento de formas y visión artificial, Ed. Addison Wesley Iberoamericana.

11. Meisel William S., Computer-oriented approaches to pattern recognition, Ed. Academic Press, Inc.

12. Mendel Jerry M., A prelude to neural networks: Adaptive and learning systems, Ed. PTR Prentice Hall.

13. Mendel Jerry M. and K. S. Fu, Adaptive, learning and pattern recognition system. Theory and applications, Ed. Academic Press.

14. Merrck Uhr Leonard, Pattern recognition, learning and thought, Ed. Prentice Hall, Inc.

15. Michalski Ryszard S., Carbonell Jaime G. and Mitchell Tom M., Machine learning (an artificial intelligence approach), Ed. Morgan Kaufmann Publishers, Inc.

16. Mode Elmer B., Elementos de probabilidad y estadística, Ed. Reverté Mexicana, S.A.

17. Moharir P. S., Pattern recognition transforms, Ed. Research Studies Press Ltd. John Wiley & Sons Inc.

18. Moreno Bonnet Alberto y Jauffred M. Francisco Javier, Elementos de probabilidad y estadística, Ed. Alfa Omega.

19. Nadley Morton and Smith Eric P., Pattern recognition engineering, Ed. John Wiley & Sons Inc.

20. Nigrin Albert, Neural networks for pattern recognition, Ed. A Breadford Book, The Mit Press.

21. Pal Sankar K.and Dufta Majumber Dwuijesh K., Fuzzy-Mathematical approach to pattern recognition, Indian Statistical Institute Calcutta, India, Ed. John Wiley & Sons Inc.

22. Pao Yoh-Han, Adaptive pattern recognition and neural networks, Ed. Addison Wesley Publishing Company, Inc.

23. Rascón CH. Octavio A., Introducción a la teoría de probabilidades, Universidad Nacional Autónoma de México.

24. Rich Elaine y Knight Kevin, Inteligencia artificial, Ed. Mc Graw Hill.

25. Shapiro Stuart C., Encyclopedy of artificial intelligence, Vol. 2. M-Z, Ed. John Wiley & Sons Inc., Wiley Interscience Publication.

26. Simpson Patrick K., Artificial neural systems. foundations, paradigms, applications and implementations, Ed. Pergamon Press.

27. Sklansky Jack, Pattern classifiers and trainable machines, Ed. Springer-Verlag.

28. Sklansky Jack, Pattern recognition. Introduction and foundations, Ed. Dowden, Hutchinson & Ross, Inc.

29. Tou Julius T. y González Rafael C., Pattern recognition principles, Ed. Addison Wesley Publishing Company.

30. Walpole Ronald E. y Myers Raymond H., Probabilidad y estadística para ingenieros, Ed. McGraw-Hill.

31. Wayne Daniel W., Estadística en aplicaciones a las ciencias sociales y a la educación, Ed. McGraw-Hill Interamericana.

OTRAS OBRAS

1. Readings in knowledge acquisition and learning, Edited by Bruce G., Buchanan & Wilkins C. David.