CAPITULO I

1. ANALISIS DISCRIMINANTE

En el sentido más amplio uno de los objetivos del análisis discriminante es inferir a cual de un grupo de k clases predeterminadas pertenece una instancia conociendo previamente las clases a la que pertenecen un grupo de n instancias anteriores, donde cada instancia solamente pertenece a una clase. Cada instancia es determinada por un conjunto de p variables que pueden ser nominales o numéricas.

La tabla I muestra un subconjunto de 8 registros provenientes de una base de datos hipotética, en este caso podríamos decir que cada registro es una instancia. Si se supone que el riesgo de enfermar de osteoporosis[1] o si la persona ya está enferma de osteoporosis se determina según un examen que se le realiza al paciente conocido como densitometría[2], tenemos que un paciente cualquiera puede estar en uno de los cuatro casos listados.

TABLA I

EJEMPLO DE REGISTROS PARA UN ANÁLISIS DISCRIMINANTE

No.	No. ID	Edad	Sexo	Raza predominante	Practica deporte	Dieta Balanceada	Nivel de Osteoporosis
1	11023	52	M	Negra	N	N	Riesgo Medio
2	15890	35	F	Blanca	S	S	Riesgo Bajo
3	22560	18	M	Asiática	S	N	Riesgo Bajo
4	15000	65	F	Blanca	N	N	Enfermo
5	12200	42	F	Negra	N	N	Riesgo Medio
6	14890	29	M	Indígena	S	N	Riesgo Medio
7	11534	39	M	Blanca	N	N	Riesgo Alto
8	12667	49	F	Negra	N	N	??

Cada instancia en la tabla I está compuesta por una clase (nivel de osteoporosis) y 6 variables que se supone determinan la clase a la que pertenece cada instancia (paciente), aunque la variable categórica No. ID simplemente es una identificación por lo tanto no ejerce ninguna influencia.

[1] La osteoporosis es una enfermedad que consiste en la pérdida constante de la densidad mineral ósea lo cual puede llevar a fracturas.
[2] La densitometría es un examen que mide la densidad mineral de los huesos.

Dada la muestra de los 7 primeros pacientes de la tabla I sería útil conocer a que clase pertenecería la instancia # 8 de la cual se conocen sus variables pero no su nivel de osteoporosis.

Al igual que el caso de la tabla I existen una serie de situaciones en las que el análisis discriminante es procedente.

Si se considera el caso particular de n instancias determinadas por 2 variables numéricas X y Y que pertenecen a una de 2 clases diferentes, se tiene que el par ordenado (X_i,Y_i) pertenece a la clase C_1 o a la clase C_2. La Figura 1.1 representa esta situación gráficamente.

Pudo haber ocurrido que al intentar determinar la clase (el nivel de osteoporosis) a la que pertenece la instancia del registro No. 8 del ejemplo anterior se cometa un error. El análisis discriminante no está exento de la mala clasificación tal como se aprecia en la Figura 1.1 donde la curva discriminante falla en algunos de los puntos. Aunque los errores de mala clasificación pueden ocurrir, al discriminar se intenta minimizar la probabilidad de una mala clasificación (o maximizar la probabilidad de clasificar correctamente).

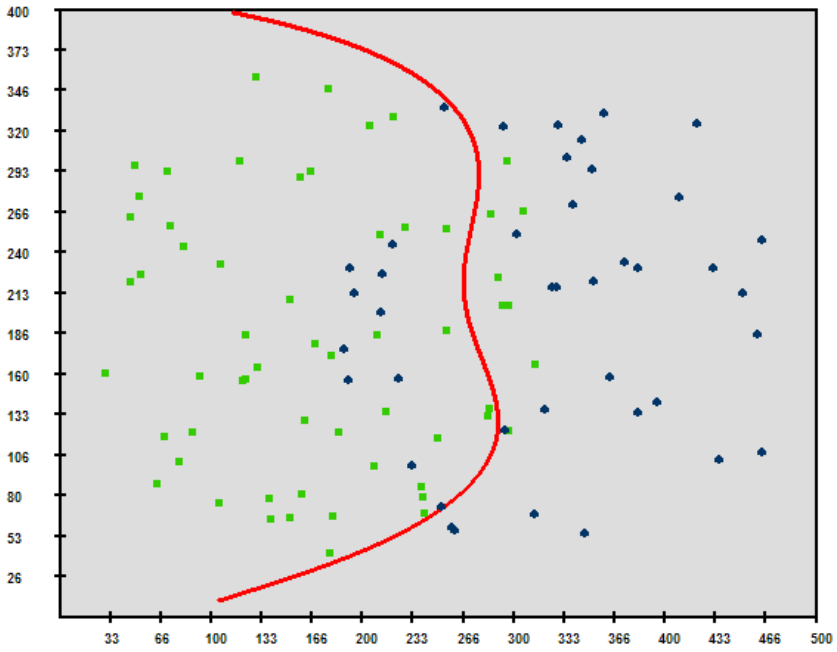

FIGURA 1.1 ILUSTRACIÓN DE UNA CURVA DISCRIMINANTE

A este respecto es útil el conocido teorema de Bayes:

$$P(A \setminus B) = \frac{P(B \setminus A) \cdot P(A)}{P(B)}$$

El cual da la probabilidad de un evento A dado que ha ocurrido otro evento B. En esta expresión P(A) se conoce como probabilidad a priori del evento A, pues es la probabilidad de que ocurra el evento A sin considerar que ya ha ocurrido el evento B, mientras que P(A\B) se conoce como probabilidad a posteriori.

Para el caso que se mencionó en la tabla I bien podríamos expresar la probabilidad P(A\B) del teorema de Bayes de la siguiente forma:

p=Prob(N= R. Alto \ B) =P(N=R. Alto\ X_1=49;X_2=F;X_3=Negra;X_4=N;X_5=N)

donde:

N: Nivel de osteoporosis

X_1: Edad

X_2: Sexo

X_3: Raza predominante

X_4: Practica deporte

X_5: Dieta balanceada

B: Evento en el que X_1=49; X_2=F; X_3=Negra; X_4=N y X_5=N a la vez.

Siguiendo el teorema de Bayes se tiene que:

$$p = \frac{P(X1 = 49; X2 = F; X3 = Negra; X4 = N; X5 = N \setminus N = Alto) \cdot P(N = Alto)}{P(X1 = 49; X2 = F; X3 = Negra; X4 = N; X5 = N)}$$

Donde $P(X_1,X_2,X_3,X_4,X_5)$ es la función de probabilidad conjunta.

Si se tuviesen las funciones para calcular las probabilidades de la expresión anterior se obtuviese un valor para Prob(N= Riesgo Alto \ B). De la misma forma se obtuviese las probabilidades Prob(N=R Medio \ B) Prob(N= R Bajo \ B) y Prob(N=Enfermo \ B). De estas 4 probabilidades se escogería la más alta es decir, aquel nivel de osteoporosis (o clase) que tiene la más alta probabilidad de ocurrir dado B.

Más adelante se verá el clasificador Naive Bayes que añade una suposición a este enfoque para producir resultados.

1.1 Minería de datos

Se podría decir que la minería de datos se encarga de la extracción de conocimiento a partir de datos. Los datos pueden provenir de grandes bases transaccionales de sistemas informáticos. El aprendizaje a partir de los datos no es una tarea sencilla e involucra algunos factores a considerar.

La minería de datos puede realizar algunas tareas con los datos tales como la discriminación, la categorización, la regresión, entre otras. De igual manera existen algunos métodos para realizar estas tareas. En el presente trabajo se mencionan 5 métodos para realizar la misma tarea: la discriminación.

Puesto que no todos los métodos proporcionan los mismos resultados, existen métodos que son más apropiados para un cierto conjunto de datos antes que otros. Cada método tiene sus

fortalezas y debilidades basadas en una serie de características que permiten al analista decidir cual o cuales de estos métodos utilizar. Algunas de estas características son las siguientes:

Precisión: En el caso del análisis discriminante es la capacidad que tiene el método de clasificar correctamente un nuevo elemento. Una forma natural de selección sería escoger el método que discrimine mejor para una determinada muestra. Para el análisis discriminante una medida de precisión de un método es el porcentaje de malas clasificaciones. Aquí, es necesario establecer la diferencia entre error de entrenamiento y error de prueba.

Error de entrenamiento**:** Un determinado método opera en base a una muestra de datos que se le proporciona, ya sea explícita o implícitamente basará sus respuestas en estos datos los cuales son conocidos como datos de entrenamiento. A partir de este conjunto de entrenamiento el método intentará generalizar para datos futuros o no proporcionados. En el caso de discriminación un método no siempre logrará separar completamente los datos de entrenamiento, en la Figura 1.2 se muestra que el método ha clasificado incorrectamente 3 puntos azules. El error de entrenamiento será el porcentaje de malas clasificaciones en los datos de entrenamiento, en el caso de la figura será 8/40=20%=0.2

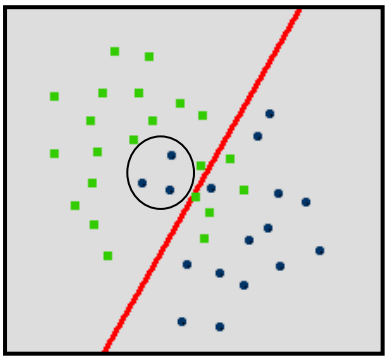

FIGURA 1.2 ERROR DE ENTRENAMIENTO

Aunque el error de entrenamiento es una medida del posible error del método no será adecuado utilizarlo porque por lo regular este error estará subestimado pues está ajustado a los datos.

Error de prueba: El procedimiento para evaluar el error de un método es tomar de la muestra total un conjunto de datos que el método no tome en cuenta, este conjunto conocido como conjunto de prueba podrá ser utilizado para evaluar el método una vez que éste sea capaz de responder. Será más confiable evaluar el método con este conjunto de prueba puesto que el método no basa sus inferencias en este conjunto. En el caso de discriminación el error de prueba es el porcentaje de malas clasificaciones que realiza el método con los datos de prueba.

Estabilidad: Es la variabilidad que tiene el método a diferentes muestras de la misma población. El método es inestable cuando sus resultados varían significativamente al aplicarlo a una muestra diferente de la misma población.

Expresividad: Es la capacidad que tiene el método para capturar el patrón de los datos. En el caso de la discriminación, es la capacidad que tiene el método de separar los datos según la clase de cada miembro. Las figura 1.3 muestra dos grados diferentes de expresividad para la misma muestra. Es notorio en la figura que la curva discriminante de la izquierda no puede separar completamente los datos debido a su rigidez. Se puede decir entonces que el método aplicado a la derecha es más expresivo que el método de la izquierda.

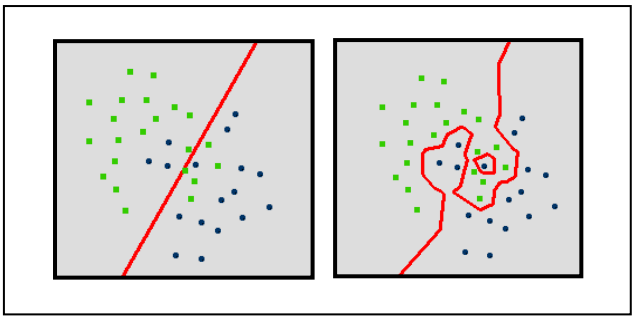

FIGURA 1.3 GRADOS DE EXPRESIVIDAD

Comprensibilidad: Aunque un método pueda realizar una tarea no necesariamente producirá un modelo que se pueda aplicar a futuras observaciones y en caso de que lo produzca éste no

siempre resultará comprensible. En este punto es necesario mencionar que existe una división de los métodos en relación a la forma como proceden para dar resultados.

Métodos anticipativos: Estos métodos emplean largo tiempo en aprender o entrenarse pero poco tiempo en responder. Estos métodos emplean todo el conjunto de datos para generar algún tipo de modelo que sirva para decisiones futuras, en el momento que logran obtener el modelo (*aprender*), los datos ya no son más útiles y pueden ser desechados, debido a que el modelo será suficiente para responder a cualquier observación futura. Estos métodos son globales en el sentido de que los modelos que producen están basados en todo el conjunto de datos. Un método anticipativo, por ejemplo, es el método de regresión lineal que se verá más adelante.

Métodos retardados: Es posible que un método no produzca un modelo general que responda a cualquier observación futura, sino que en el momento que se le pregunte comience a procesar y responda. A diferencia de los anticipativos estos métodos utilizan las observaciones anteriores cada vez que se les pregunta, por lo tanto, el conjunto de datos no puede ser desechado. Estos métodos ocupan muy poco tiempo en entrenarse pero mucho tiempo en responder. Puesto que no producen un modelo general los métodos retardados son incomprensibles.

Se podría decir también que desde el punto de vista humano ciertos métodos son más comprensibles que otros, como se verá más adelante, el sistema de reglas que produce el árbol de decisión es muy comprensible.

Robustez al ruido o a outliers: Esta característica se refiere a la sensibilidad que tiene un método a los datos incorrectos, datos aberrantes o extremos (*outliers*). Se dice que un método es robusto a outliers, si los valores anómalas ejercen muy poca influencia sobre su funcionamiento.

Costo computacional: Esta característica se refiere a la eficiencia del aprendizaje de un método o el tiempo que tarda en responder a observaciones futuras. Aunque el tiempo que tardan los métodos anticipativos en generar un modelo puede ser alto cuando ya lo tienen el tiempo que tardan en responder a una observación futura es casi instantáneo. Sin embargo los métodos retardados casi no emplean ningún tiempo en entrenarse y su mayor tiempo lo emplean en responder a una observación futura. Dependiendo del número de datos el costo computacional de estos

métodos suele ser mayor en relación al costo computacional de los métodos anticipativos.

1.2 Método de los K-Vecinos más cercanos.

El método de los k-vecinos más cercanos (knn de k-nearest neighbors) utiliza un sencillo criterio de clasificación, sin embargo, este método dependiendo del número de instancias puede demandar un considerable costo computacional.

Cuando se quiere inferir la clase a la que pertenece un nuevo ejemplo, como su nombre lo indica, este método busca los k ejemplos más próximos a él (de los cuales, como ya se mencionó, ya se conocen sus clases) y le asigna al nuevo ejemplo la clase que más se repite en estos k ejemplos.

Para el caso particular en que cada ejemplo viene determinado por 2 variables éstos pueden ser graficados como puntos en un plano bidimensional (véase la Figura 1.4). Para un nuevo punto se pueden obtener sus k vecinos más cercanos utilizando la distancia

euclídea entre 2 puntos y obtener la clase que más se repite en estos vecinos.

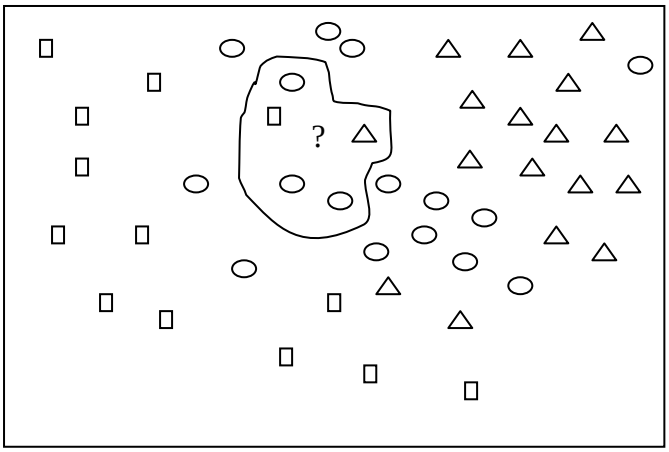

FIGURA 1.4 ILUSTRACIÓN DE LOS 5-VECINOS MÁS CERCANOS

Si se considera k=5 se observa en la figura 1.4 que el nuevo punto denotado por "?" cuya clase es desconocida tiene 1 vecino de clase "triángulo", 1 vecino de clase "cuadrado" y 3 vecinos de clase "elipse". Por lo tanto se inferirá que este punto es de tipo "elipse".

El método de los k vecinos es bastante flexible y es capaz de trazar curvas discriminantes bastante irregulares, sin embargo cuando k=1 este método puede trazar curvas que no clasifiquen tan correctamente nuevos ejemplos puesto que no toma en cuenta la densidad o la región donde se encuentra el ejemplo. En la figura 1.5 se puede observar este hecho. El nuevo punto "?" en la figura

se lo clasifica como “elipse “ a pesar de que se encuentra en una región poblada por puntos de tipo “triángulo”.

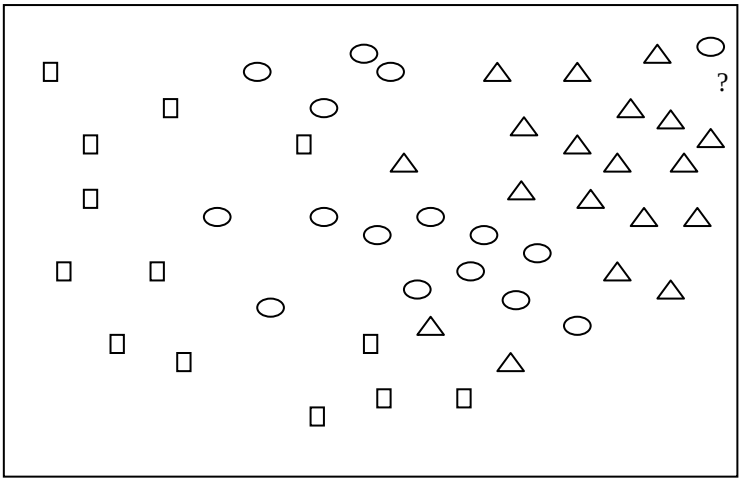

FIGURA 1.5 LIMITACIÓN DE K-VECINOS PARA K=1

1.3 Método de Naive Bayes Kernel

En una sección anterior se expuso el uso del teorema de Bayes en el análisis discriminante. En resumen se mencionó que si se tienen k clases { C_1,C_2,C_3,……,C_k } y n variables { X_1,X_2,…..,X_n } se pueden obtener las probabilidades condicionales de obtener cada clase dado que el vector X=(X_1,X_2,X_3,……,X_n) toma los valores {a_1,a_2,a_3,…..,a_n} de la siguiente forma:

$$P(C_1 \setminus a_1, a_2, a_3,....,a_n) = \frac{P(a_1, a_2, a_3,....,a_n \setminus C_1)P(C_1)}{P(a_1, a_2, a_3,....,a_n)}$$

$$P(C_2 \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_2)P(C_2)}{P(a_1, a_2, a_3, ..., a_n)}$$

$$P(C_3 \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_3)P(C_3)}{P(a_1, a_2, a_3, ..., a_n)}$$

…….

$$P(C_k \setminus a_1, a_2, a_3, ..., a_n) = \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_k)P(C_k)}{P(a_1, a_2, a_3, ..., a_n)}$$

La clase C que produce la probabilidad condicional máxima es la que se elige como clase de pertenencia del elemento $\{a_1,a_2,a_3,\ldots..,a_n\}$. Puesto que el denominador de las expresiones anteriores es el mismo se puede prescindir de éste para calcular la probabilidad máxima, es decir las siguientes 2 expresiones producen el mismo resultado:

$$Max\left\{\frac{P(a_1, a_2, a_3, ..., a_n \setminus C_1)P(C_1)}{P(a_1, a_2, a_3, ..., a_n)}, ..., \frac{P(a_1, a_2, a_3, ..., a_n \setminus C_k)P(C_k)}{P(a_1, a_2, a_3, ..., a_n)}\right\}$$

$$Max\left\{P(a_1, a_2, a_3, ..., a_n \setminus C_1)P(C_1), ..., P(a_1, a_2, a_3, ..., a_n \setminus C_k)P(C_k)\right\}$$

Como ya se mencionó $P(a_1,a_2,a_3,\ldots..,a_n \setminus C_i)$ es la probabilidad conjunta dado que pertenecen a C_i. El método de Naive Bayes supone que las variables $\{X_1,X_2,\ldots\ldots,X_n\}$ son condicionalmente independientes, por lo tanto:

$P(X_1,X_2,X_3,\ldots\ldots,X_n \setminus C_i)=P(X_1 \setminus C_i)P(X_2 \setminus C_i)P(X_3 \setminus C_i)\ldots\ldots P(X_n \setminus C_i)$.

Con esta suposición la expresión anterior se simplifica a:

$$Max\left\{P(C_1)\prod_{i=1}^{n}P(a_i \setminus C_1), P(C_2)\prod_{i=1}^{n}P(a_i \setminus C_2),....,P(C_k)\prod_{i=1}^{n}P(a_i \setminus C_k)\right\}$$

Si se está tratando con valores numéricos no es común que se conozca las funciones de probabilidad de cada variable, por ello para estimar la función de densidad de las variables se hace uso de los estimadores núcleo K(x).

Una función de densidad f(x) se puede estimar haciendo uso de un estimador núcleo K(x) de la siguiente manera:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - x_i}{h}\right)$$

El cuadro 1.1 contiene algunas opciones para las funciones núcleo K(x) donde la función I(u) es como se indica en el cuadro.

El parámetro h de esta estimación requiere una muy especial atención. Si se trata de la función núcleo G(u) del cuadro 1.1 conocida como función núcleo gaussiano un posible valor de h podría ser:. $h_G = (1.06s)\cdot n^{-\ 0.2}$

Donde s es la desviación estándar de la muestra.

Para la función núcleo Q(u) del cuadro 1.1 un valor de h podría ser:

$$h_Q = 2.62(1.06s)\cdot n^{-\ 0.2}$$

$K(u) = \frac{1}{2} I(\lvert u \rvert \le 1)$
$K(u) = (1 - \lvert u \rvert) I(\lvert u \rvert \le 1)$
$K(u) = \frac{3}{4}(1 - u^2) I(\lvert u \rvert \le 1)$
$Q(u) = K(u) = \frac{15}{16}(1 - u^2)^2 I(\lvert u \rvert \le 1)$
$G(u) = K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$
$I(\lvert u \rvert \le 1) = \begin{cases} 1, & si \quad \lvert u \rvert \le 1 \\ 0, & caso \quad contrario \end{cases}$

CUADRO 1.1 ALGUNAS FUNCIONES NÚCLEO K(X) UTILIZADAS

Utilizando la función núcleo para estimar las funciones de densidad la expresión para la máxima probabilidad se podría reescribir así:

$$Max\left\{\hat{P}(C_1)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_1), \hat{P}(C_2)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_2), ..., \hat{P}(C_k)\prod_{i=1}^{n}\hat{f}(a_i \setminus C_k)\right\}$$

$$donde \quad \hat{P}(C_i) = \frac{\#\ de\ elementos\ de\ la\ clase\ C_i}{n}$$

1.4 Método de Regresión lineal

Este método pretende modelar el comportamiento de una variable respuesta en función de un conjunto de n variables explicativas $X_1,X_2,X_3,\ldots..,X_n$. Si se considera el caso simple de una variable respuesta y una sola variable explicativa X, se podría pensar que un modelo para este caso sería $Y=f(X)+Є$ donde Є es el componente aleatorio del modelo. La función f(X) podría ser lineal $f(X)= \beta_0+\beta_1X$, podría ser cuadrática $f(X)= \beta_0+\beta_1X+ \beta_2X^2$ o en general polinómica de grado n con $f(X)= \beta_0+\beta_1X+ \beta_2X^2+\ldots\ldots+ \beta_nX^n$.

Si se tienen más de 1 variable explicativa los módelos fácilmente aumentan de complejidad. En el caso de 2 variables explicativas X_1 y X_2 un modelo cuadrático sería $f(X)= \beta_0+\beta_1 X_1+ \beta_2 X_2+ \beta_3X_1X_2+ \beta_4(X_1)^2+ \beta_5(X_2)^2$ el cual como se aprecia tiene 6 términos y un polinomio de grado 4 tendrá 15 términos.

El objetivo de este método es estimar el vector $\beta=(\beta_0, \beta_1, \beta_2,\ldots, \beta_k)$ tal que la siguiente expresión sea mínima:

$$\sum_{i=1}^{n} \left(y_i - f(x_{i1}, x_{i2}, x_{i3},....,x_{im}) \right)^2$$

En la expresión anterior se considera que existen n valores tanto para la variable respuesta como para cada variable explicativa. También se observa que existen m variables explicativas.

Considerando el caso particular en que solamente existe una variable explicativa o independiente X y se plantea un modelo cuadrático, con n valores para las variables X y Y se tiene un conjunto de n ecuaciones, donde b_i es el estimador de β_i que se desea encontrar

$$\begin{cases} y_1 = b_0 + b_1 x_1 + b_2 (x_1)^2 \\ y_2 = b_0 + b_1 x_2 + b_2 (x_2)^2 \\ y_3 = b_0 + b_1 x_3 + b_2 (x_3)^2 \\ \ldots\ldots\ldots\ldots \\ y_n = b_0 + b_1 x_n + b_2 (x_n)^2 \end{cases}$$

La versión matricial del sistema anterior sería:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \cdots & & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Lo cual podría ser reescrito como Y=XB considerando que Y,X y B son matrices. La solución del sistema matricial que considera el criterio de mínimos cuadrados que se mencionó anteriormente es $B=(X'X)^{-1}X'Y$ donde X' es la matriz transpuesta de X.

Anteriormente se expuso que en análisis discriminante se intenta clasificar un elemento determinado por n variables en una de k clases posibles. Se podría plantear este problema bajo el enfoque de regresión lineal suponiendo que las n variables que determinan un elemento son las variables explicativas y la variable de respuesta Y es una medida que es útil para clasificar de donde proviene el elemento.

Si se considera el caso general en que se tienen k clases posibles y n variables que determinan cada elemento podría resolverse el problema planteando k modelos de regresión del tipo $Y_i = f(X_1, X_2, X_3,\ldots\ldots, X_n)$. Es decir un modelo de regresión por cada clase. Puesto que las clases son valores no cuantitativos, la variable respuesta Y_i del modelo de regresión $f(X_1, X_2, X_3,\ldots\ldots, X_n)$ se establecerá a 1 si es que el elemento $(X_1, X_2, X_3,\ldots\ldots, X_n)$ pertenece a la clase i y se establecerá a 0 en caso contrario. De esta manera

se podrán encontrar las estimaciones de los β's para cada modelo de regresión.

Utilizando el enfoque anterior para clasificar un nuevo elemento con variables (X_1, X_2, X_3,……, X_n) se deberá evaluar los valores de este nuevo elemento en cada una de los k modelos de regresión. Se inferirá que este elemento pertenece a la clase i si su correspondiente variable respuesta Y_i es la máxima dentro de todas las variables respuesta.

Si se considera el caso particular en que un determinado elemento solamente puede pertenecer a 2 clases posibles, se tiene el caso en que la variable respuesta Y es binaria. En este caso se podría generar un solo modelo de regresión en el cual la variable Y se la establecería como 1 si el elemento o vector (X_1, X_2, X_3,……, X_n) pertenece a una clase y 0 si pertenece a la otra. Una política para clasificar un nuevo elemento, dado que se han encontrado las estimaciones de los coeficientes β's del modelo de regresión propuesto es evaluar los valores de las variables del nuevo elemento en el modelo de regresión, si la variable respuesta Y es mayor o igual a 0.5 se inferirá que el nuevo elemento pertenece a la

clase que se codificó como 1 al plantear el problema, caso contrario se inferirá que pertenece a la otra clase.

1.5 Método de regresión logística

Para este método se considerará el caso en el que la variable respuesta es binaria.

Si se desea aplicar regresión lineal a un conjunto de datos en donde la variable respuesta Y es binaria, 1 ó 0, una pregunta que surgiría sería si E[Y]=Prob(Y=1). En el caso de regresión lineal es posible que los valores $\hat{Y}_i$ sean menores que 0 o mayores que 1 lo cual no puede ser interpretado como una probabilidad. La regresión logística es un método que “acomoda” la variable respuesta para que ésta “caiga” en el intervalo [0,1].

En regresión lineal se plantearon modelos del tipo:

$$Y_i=\beta_0+\beta_1X_{1i}+ \beta_2X_{2i}+ \beta_3X_{3i}+\ldots\ldots+ \beta_nX_{mi}$$

Donde i=1,2….,n y hay m variables explicativas.

Para regresión logística se planteará el siguiente modelo:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_n X_{mi}$$

Donde p_i es igual a Prob(Y_i=1). El modelo de regresión logística anterior también puede ser escrito como:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$$

En esta última expresión se puede observar que a diferencia de la regresión lineal en la regresión logística sean cuales sean las estimaciones para $\beta_0,\beta_1,\beta_2,\beta_3,\ldots,\beta_m$ y sean cuales sean los valores de X_1, X_2, X_3,……, X_m la respuesta p_i siempre estará en el intervalo [0,1].

Puesto que p_i es la Prob(Y_i=1), si ésta es mayor que 0,5 el elemento cuyas variables se evaluaron pertenecerá a una clase determinada, si p_i<0.5 pertenecerá a la otra. También se puede estar interesado en graficar la curva discriminante (cuando es posible) o al menos conocer la ecuación algebraica de esta curva. El límite de discriminación, es decir la frontera entre una clase y la otra ocurrirá cuando p_i=0.5. Para hallar la ecuación de la curva discriminante se reemplaza p_i=0.5 en el modelo:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$$

Quedando:

$$0.5 = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$$

Esto ocurrirá cuando $\beta_0+\beta_1X_{1i}+ \beta_2X_{2i}+ \beta_3X_{3i}+\ldots\ldots+ \beta_nX_{mi} = 0$. Siendo ésta la ecuación de la curva discriminante.

Para el caso particular en que existen sólo 2 variables explicativas, la curva discriminante podrá visualizarse en un plano bidimensional como se verá más adelante.

Para estimar los parámetros del modelo logístico se utiliza el criterio general de máximo-verosimilitud lo que desemboca en la siguiente ecuación matricial iterativa:

$B^{t+1}=(X'WX)^{-1}X'Wz$, donde $z=XB^t+W^{-1}(Y-P)$, donde B^t es el estimador del vector de coeficientes β en la iteración t.

La ecuación matricial para hallar B^{t+1} es bastante similar a la expresión $B=(X'X)^{-1}X'Y$ que es la solución para encontrar los estimadores de los coeficientes en el método de regresión lineal que expuso anteriormente.

Una de las diferencias es la presencia de la matriz W, conocida como matriz de ponderación que se define como:

$$W = \begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \cdots & \\ & & & p_n(1-p_n) \end{bmatrix}$$

Una de las suposiciones de la regresión lineal $Y_i = f(X_i) + \epsilon_i$ es que $Var[\epsilon_i]=\sigma^2$, es decir la varianza de los errores se la supone constante para cada observación. Sin embargo, esta suposición no siempre es realista puesto que es posible que esta varianza cambie conforme los valores de X_i cambian. En el modelo de regresión logística se supone que $Var[\epsilon_i]=\sigma_i^2$, lo cual indica que la varianza depende de la observación i. La presencia de la matriz W se justifica debido a esta última suposición puesto que Y_i es una variable aleatoria binomial con media $E[Y_i]=1(pi)+0(1-pi)=p_i$ y varianza $Var[Y_i]=p_i(1-p_i)$, siendo estas varianzas los pesos de la matriz W.

Para encontrar $B^{t+1}=(X'WX)^{-1}X'Wz$, donde $z=XB^t+W^{-1}(Y-P)$ se aplica el siguiente método iterativo:

1) Se inicia con $B^0=0$ y $p_i=0.5$ para todos los i. Con estos valores iniciales se construye la matriz W y se encuentra la matriz respuesta z.
2) Se calcula B^1 con la matriz z anterior.
3) Se utiliza los β's de B^1 para calcular las nuevas probabilidades p_i utilizando el mismo modelo: $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\beta_n X_{mi})}}$
4) Se calcula nuevamente la matriz respuesta z.
5) Se calcula B^2 y se continúa con el proceso iterativo hasta que los valores de B converjan.

Aunque la convergencia no se garantiza, es usual que ocurra.

1.6 Árbol de decisión.

El árbol de decisión es un método que realiza la tarea de clasificación mediante segmentaciones sucesivas del conjunto de datos.

La figura 1.6 muestra el esquema que en general sigue un árbol de decisión al realizar la tarea de discriminación para el caso de 2 clases.

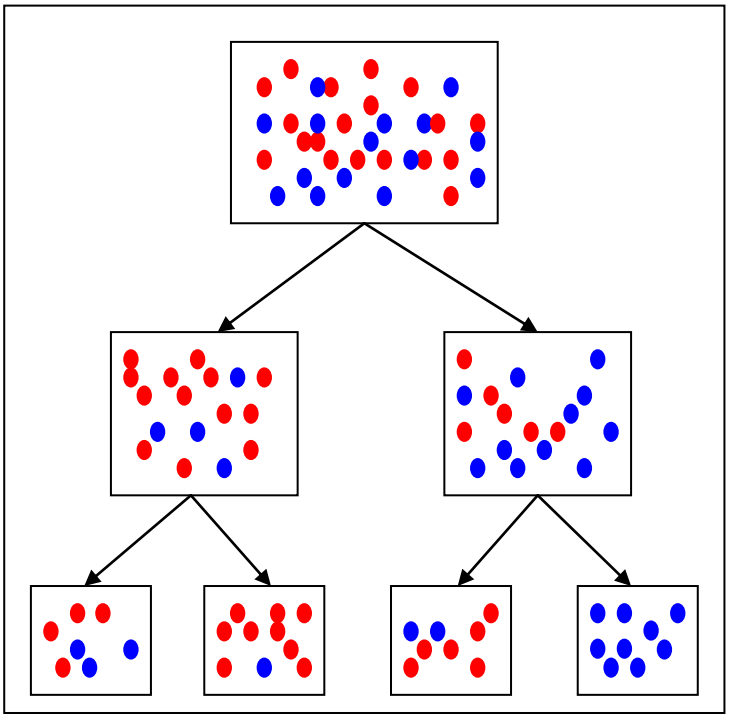

FIGURA 1.6 ESQUEMA DE LA DISCRIMINACIÓN UTILIZANDO UN ARBOL DE DECISIÓN

Del árbol de decisión del esquema anterior se derivan algunas observaciones:

- El nodo raíz contiene todos los elementos.
- Cada segmentación produce hijos que son más puros que el padre, es decir, que tienen menos diversidad. En el esquema anterior se nota que en el hijo izquierdo del nodo raíz el porcentaje de elementos azules es menor que en el caso de su padre, en este sentido es más puro. De igual manera en el hijo derecho el porcentaje de elementos rojos es menor que en el caso de su padre.

- Un nodo cualquiera deja de tener hijos cuando es totalmente puro, es decir tiene el 100% de sus elementos de una clase determinada. En este caso se dice que este nodo es una hoja del árbol. En el esquema anterior se observa que 3 de los 4 nodos finales pueden seguirse segmentando. En la práctica no siempre se exige que todas las hojas del árbol sean 100% puras, sino que se puede establecer una condición de parada, como por ejemplo, que el número de elementos mínimo que tenga un nodo para que pueda tener hijos (ser segmentado) sea un tanto por ciento del número total de elementos.

En el esquema del árbol de decisión presentado anteriormente se supone que cada elemento (punto rojo o azul) tienen una cierta cantidad de atributos o variables que pueden ser nominales o numéricos. Cada segmentación solamente se la hará en base a un atributo a la vez. Aunque el árbol del esquema anterior es binario, una partición en el caso nominal puede generar más de 2 hijos.

Si un atributo Xi es numérico continuo y puede tomar valores en el intervalo [a,b] entonces las 2 condiciones que segmentarán un nodo en base a ese atributo serán las siguientes: $Xi \leq k$ y $Xi > k$

donde k=(v+w)/2 donde v y w son valores consecutivos en todo el conjunto de valores que toma el atributo Xi.

De esta manera si se supone que el árbol del esquema anterior tiene elementos con 2 atributos numéricos continuos X y Y un resultado podría ser el siguiente:

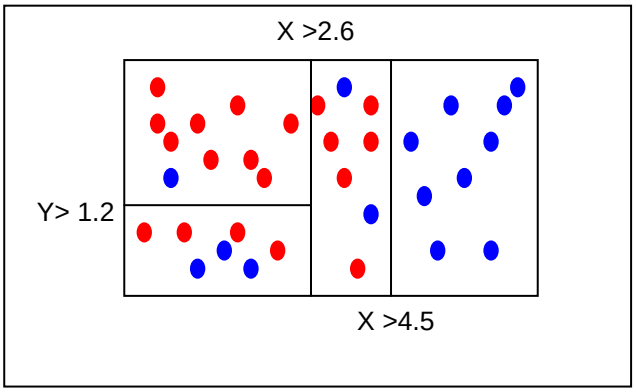

FIGURA 1.7 RESULTADO DE UNA DISCRIMINACIÓN UTILIZANDO UN ÁRBOL DE DECISIÓN

Considerando las 3 particiones de la figura 1.7 entonces el árbol de la figura 1.6 quedaría como se muestra en la figura 1.8.

Suponiendo que los puntos de la figura 1.7 son los datos de entrenamiento, entonces una de las reglas de generalización que producirá el árbol de decisión entrenado será: *si X>4.5 entonces el elemento es de la clase "azul"*. De esta manera si se quisiese

conocer la clase de un nuevo elemento cuyo atributo X=4.8 por ejemplo, entonces utilizando la regla anterior se inferirá que este nuevo elemento es de la clase “azul”.

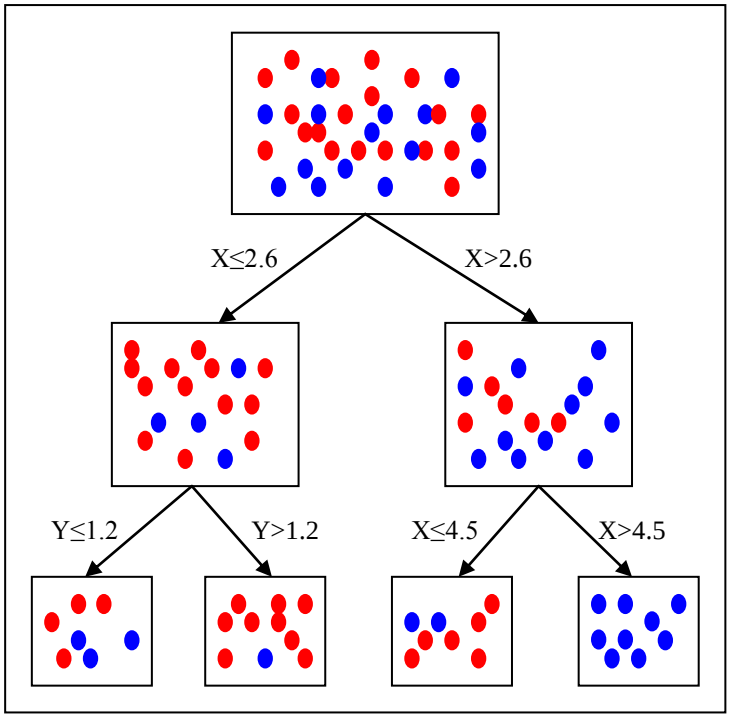

FIGURA 1.8 ESQUEMA DE UN ÁRBOL DE DECISIÓN CON ATRIBUTOS NUMÉRICOS

En el árbol anterior la primera partición se baso en el atributo X y no en el atributo Y. En la condición se escogió la constante 2.6 con la cual comparar el atributo X y no se escogió otra constante como 3.2 por ejemplo. La razón por la cual se lo hizo así es porque la partición que utiliza la condición X>2.6 es la óptima de acuerdo a un criterio determinado.

Existen distintos criterios para seleccionar la “mejor” partición, los cuales utilizan funciones de impureza del tipo $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})$ donde p_{ij} es la proporción de elementos de la clase i en el nodo j. Entre algunos criterios con sus funciones de impureza tenemos los siguientes:

- Error esperado: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})=\min(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj})$
- GINI: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = 1- \sum(p_{ij})^2$
- Entropía: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = \sum p_{ij} \log(p_{ij})$
- DKM: $f(p_{1j}, p_{2j}, p_{3j},\ldots, p_{cj}) = 2 (\prod p_{ij})^{1/2}$

Cuando se trata de atributos numéricos por cada partición se producen 2 nuevos nodos del árbol, según el criterio que se seleccione se podrá calcular la función de impureza para cada nodo. De esta forma para cada posible partición se podrá obtener el promedio ponderado de la impureza de sus 2 hijos, de la siguiente manera:

Impureza(partición)=$p_i f(p_{1i}, p_{2i}, p_{3i},\ldots, p_{ci}) + p_d f(p_{1d}, p_{2d}, p_{3d},\ldots, p_{cd})$

Donde p_i es la proporción de elementos del nodo padre que pertenecen al hijo izquierdo y p_d es la proporción de elementos que pertenecen al hijo derecho. Entonces, de todas las posibles particiones se escogerá aquella que produzca la menor impureza ponderada.

El tratamiento anterior podrá extenderse cuando no se trate de árboles binarios simplemente obteniendo el promedio ponderado de las impurezas de los hijos que se produzcan con cada partición.