CAPITULO II

2. BREVE DESCRIPCIÓN DEL ANÁLISIS DISCRIMINANTE.

2.1 Introducción

La discriminación y la clasificación son técnicas multivariantes concernientes a separar distintos conjuntos de objetos (u observaciones) y con colocar nuevos objetos (observaciones) en grupos previamente definidos. En este capítulo intentaremos básicamente explicar en qué consiste genéricamente la discriminación crediticia que llevaremos a cabo. El análisis discriminante es exploratorio en naturaleza.

Como un procedimiento separativo, a veces es empleado en una base única con el objeto de investigar las diferencias observables cuando las relaciones causales no son bien entendidas. Los

procedimientos clasificatorios son menos exploratorios en el sentido que conllevan hacia reglas bien definidas, que pueden ser usadas para asignar nuevos objetos. La clasificación ordinariamente requiere más estructura del problema que la discriminación.

Entonces, los objetivos inmediatos de la discriminación y clasificación, respectivamente son:

- Describir, sea gráfica (en tres o menos dimensiones) o algebraicamente, las características primarias diferenciales de los objetos (observaciones) de varias colecciones conocidas (poblaciones). Tratamos de encontrar “discriminantes” cuyos valores numéricos sean tales que las colecciones sean separadas tanto como sea posible.

- El sortear objetos (observaciones) en dos o más clases etiquetadas. El énfasis es en derivar una regla que pueda ser usada para asignar óptimamente nuevos objetos hacia las clases predefinidas.

Debemos seguir la convención y usar el término *discriminación* para referirnos al primer objetivo. Esta terminología fue introducida por R. A. Fisher en el primer tratado moderno de problemas separativos. Un término que describe mejor este objetivo es *separación*. Nos referiremos hacia el segundo objetivo como *clasificación* o *colocación*.

Una función que separa objetos puede servir, algunas veces, como un colocador, y, análogamente, una regla que coloca objetos, esto puede sugerir un procedimiento discriminatorio. En la práctica, ambos objetivos se confunden y la distinción entre separación y colocación se torna difusa.

2.2 Separación y clasificación para dos poblaciones

Para fijar las ideas, listemos situaciones en que uno puede estar interesado en (1) separar dos clases de objetos o (2) asignar un nuevo objeto hacia una de dos clases (o ambas). Es conveniente etiquetar las clases como π_1 y π_2.

Los objetos son ordinariamente separados o clasificados en la base de mediciones de, por ejemplo, *p* variables aleatorias asociadas ***X'*** = $[X_1, X_2, \ldots, X_P]$.

Los valores observados de ***X*** difieren en alguna extensión de una clase a otra (sí los valores de ***X*** no fueran muy diferentes para los objetos en π_1 y π_2, no existiría un problema; esto es, las clases serían indistinguibles, y los nuevos objetos podrían ser asignados hacia cualquier clase indiscriminadamente).

Podremos pensar en la totalidad de los valores de la primera clase como ser de la población de ***x*** valores para π_1 y aquellos para la segunda clase como la población de ***x*** valores para π_2.

Estas dos poblaciones pueden ser entonces descritas por las funciones de densidad de probabilidad $f_1(\boldsymbol{x})$ y $f_2(\boldsymbol{x})$, y consecuentemente, podemos hablar de asignar observaciones a poblaciones u objetos a clases indiferentemente.

Para comprender mejor el caso práctico la siguiente lista nos ilustra algunos ejemplos de este planteamiento.

TABLA 2.1 Ejemplos de Clasificación y Discriminación

Poblaciones p_1 ***y*** p_2	***Variables Medidas X***
1. Compañías de seguros solventes y financieramente estresadas	Activos totales, costo de stocks y bonos, valor del stock en el mercado, gastos de pérdidas, ahorros, cantidad de primas contratadas.
2. Discépticos Ulcéricos (con problemas de estómago sensible) y controles (“normales”)	Medidas de ansiedad, dependencia, culpa, perfeccionismo
3. *Documentos Federalistas* escritos por James Madison y aquellos escritos por Alexander Hamilton.	Frecuencia de distintas palabras y largo de las oraciones.
4. Dos especies de hierba	Largo de sépalos y pétalos, profundidad de las grietas de los pétalos, largo de la hoja, diámetro del polen.

5. Compradores de un nuevo producto y aquellos compradores que dubitan al comprar.	Educación, ingreso, tamaño de la familia, cantidad de marcas previamente cambiadas.
6. Estudiantes universitarios exitosos y no exitosos (fallaron al graduarse)	Notas del examen de ingreso, nota promedio de graduación del colegio, número de actividades extracurriculares de la secundaria.
7. Hombres y mujeres	Medidas antropológicas tales como circunferencia y volumen de cráneos antiguos.
8. Buenos y malos sujetos de crédito*	Ingreso, edad, número de tarjetas de crédito, tamaño de la familia.
9. Alcohólicos y no alcohólicos	Actividad de las enzimas de monoaminas oxidadas, actividad de las enzimas de adenilato cicloso.

Vemos del ítem 5, por ejemplo, que los objetos (consumidores) serían separados en dos clases etiquetadas (“compradores” y “no compradores”) en la base de valores observados de presumiblemente variables relevantes (educación, ingreso, etc.). En la terminología de observación y población, queremos identificar una observación de la forma x’ = [x1 (educación), x2 (ingreso), x3 (tamaño familiar), x4 (número de cambios de marca)] como una población $\pi 1$, compradores, o población $\pi 2$, no compradores.

2.2.1 Clasificación de dos poblaciones

La colocación o clasificación tiene reglas que son usualmente desarrolladas de ejemplos de entrenamiento (tal y como ocurre en las redes neuronales).

Las características medidas de objetos, aleatoriamente elegidos sabiendo que provienen de cada una de las dos poblaciones son examinadas para distinguir sus diferencias.

Esencialmente, el conjunto de todos los posibles resultados se divide en dos regiones, R_1 y R_2, tal que si una nueva observación cae en R_1, se coloca en la población π_1, y si cae en R_2, la colocamos en π_2. Por tanto, un conjunto de observaciones favorece a π_1, mientras que el otro conjunto de valores favorece a π_2.

Talvez nos estemos preguntando cómo es que sabemos que algunas observaciones pertenecen a una población en particular, pero dudamos de otras. (Esto, por supuesto, es lo que hace a la clasificación un problema) Algunas condiciones pueden originar esta aparente anomalía:

- *Conocimiento incompleto de un funcionamiento futuro. Ejemplo:* En el pasado los valores extremos de variables financieras eran observados con dos años de anterioridad antes de que ocurra la subsecuente bancarrota de alguna firma. Clasificar otra firma como “estresada” sobre la base de valores observados de estos indicadores liderantes, podrían permitir a los oficiales, el tomar acciones

correctivas, de ser necesario, antes de que sea demasiado tarde. Otro ejemplo; una oficina de aplicaciones para un colegio médico quisiera clasificar un aplicante como un posible o no posible candidato a convertirse en Médico sobre la base de las notas de los exámenes y otros registros universitarios. Aquí, una determinación actual puede ser hecha únicamente al final de varios años de entrenamiento.

- *Información "perfecta" requiere destruir el objeto. Ejemplo:* La vida útil de una batería de calculadora se determina al usarla hasta que falle, y la resistencia de un trozo de madera se obtiene al doblarlo hasta que se quiebra. Los productos fallosos no pueden ser vendidos. Uno quisiera clasificar productos como buenos y malos (que no cumplen con las especificaciones) sobre la base de ciertas mediciones preliminares.

- *Información no disponible o costosa. Ejemplo:* Se asume que algunos de los *Documentos Federalistas* fueron

escritos por James Madison o Alexander Hamilton porque ellos los firmaron. Otros documentos, sin embargo, no fueron firmados y es de interés el determinar cuál de los dos hombre escribieron los papeles sin firmar. Claramente, no podemos preguntarles. Las frecuencias de las palabras y largo de las oraciones podrían ayudar a clasificar los papeles en disputa. Otro ejemplo; muchos de los problemas médicos pueden ser identificados concluyentemente sólo al conducir una costosa operación. Usualmente, uno quisiera diagnosticar una enfermedad a través de fácilmente observables, aunque potencialmente fallidos, síntomas externos. Esta aproximación ayuda a evitar operaciones costosas e innecesarias.

Debe estar claro de estos ejemplos que las reglas de clasificación no pueden usualmente proveer un método de asignación libre de errores.

Esto ocurre porque puede no existir una clara distinción entre las características medidas de las poblaciones; esto es, los

grupos pueden traslaparse. Entonces es posible, por ejemplo, el clasificar incorrectamente a un objeto perteneciente a π_1 en π_2 o perteneciente a π_2 en π_1.

Consideremos el siguiente ejemplo: (Discriminando poseedores de no poseedores de cortadoras de césped)
Consideremos dos grupos en una cuidad: $\pi 1$, los dueños de una cortadora de césped, y π_2, aquellos que no tienen una.

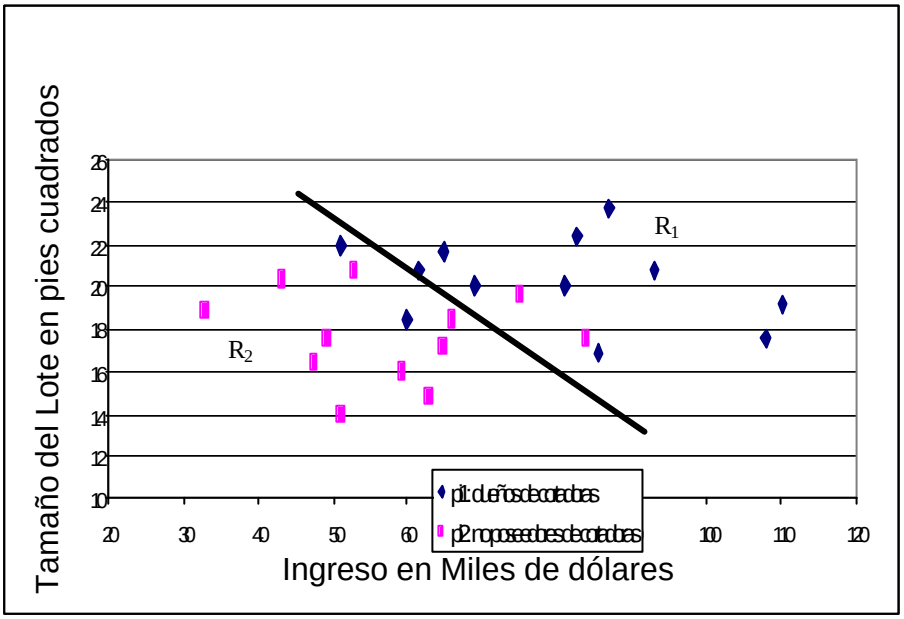
Para poder identificar los mejores prospectos de ventas para una intensa campaña de ventas, un fabricante de cortadoras de césped está interesado en clasificar a las familias como prototipo de poseedores y no poseedores sobre la base de x1 = ingreso y x2 = tamaño del lote. Muestras aleatorias de n1 = 12 dueños actuales y n2 = 12 no poseedores actuales produjeron los valores en la tabla 2.2.

Tabla 2.2 Clasificación de Familias

p_1: *dueños de cortadoras*		p_2: *no poseedores de cortadoras*	
x1 (ingreso en miles $)	x2 (área del lote en pies cuadrados)	x1 (ingreso en miles $)	x2 (área del lote en pies cuadrados)
60	18.4	75	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87	23.6	84	17.6
110.1	19.2	49.2	17.6
108	17.6	59.4	16
82.8	22.4	66	18.4
69	20	47.4	16.4
93	20.8	33	18.8
51	22	51	14
81	20	63	14.8

Estos datos se muestran en la figura 2.1. Vemos que los dueños de estas cortadoras tienden a tener mayores ingresos y más grandes lotes que los que no tienen cortadoras, aunque el ingreso parece ser un mejor “discriminador” que el tamaño del lote.

Figura 2.1 Ingreso y tamaño de lote para poseedores y no poseedores de cortadoras de césped



Por otra parte, existe algún traslapamiento entre ambos grupos. Si, por ejemplo, tuviéramos que colocar aquellos valores de (x1, x2) que caen en la región R1 (como lo determina la línea negreada de la figura) en $\pi 1$, poseedores, y aquellos valores (x1, x2) que caen en R2 para $\pi 2$, no poseedores, cometeríamos errores. Algunos dueños serían clasificados incorrectamente como no poseedores e inversamente algunos no poseedores como poseedores. La idea es crear una regla en las regiones R1 y R2 que minimice los chances de cometer errores.

Un buen procedimiento clasificatorio debe resultar en pocas colocaciones fallidas. En otras palabras, los chances, o probabilidades, de malas clasificaciones deberían ser pequeñas. Cabe mencionar que existen características adicionales que una regla de clasificación “óptima” debe tener.

Podría ocurrir que una clase o población tenga una mayor propensión de ocurrencia que otra debido a que una o dos poblaciones sean relativamente mucho más largas que otras.

Por ejemplo, tiende a haber más firmas financieramente estables de las que no lo están. Como otro ejemplo, una especie de hierba puede ser más prevaleciente que otro. Una regla de clasificación óptima debería tomar estas "probabilidades previas de ocurrencia" en cuenta.

Si realmente creemos que la probabilidad (a priori) de una firma financieramente estresada y finalmente en bancarrota es muy pequeña, entonces debemos clasificar una firma elegida al azar como no en bancarrota a no ser que los datos eminentemente favorezcan a la bancarrota.

Otro aspecto de la clasificación es el costo. Supongamos que clasificar un objeto $\pi 1$ como pertenecedor a $\pi 2$ representa un error mucho más serio que clasificar un objeto $\pi 2$ como pertenecedor a $\pi 1$. Entonces debemos ser cautelosos al hacer futuras asignaciones.

Como un ejemplo, el fallar al diagnosticar una enfermedad potencialmente fatal es substancialmente más “costoso” que concluir que la enfermedad está presente cuando, de hecho, no lo está. Un procedimiento clasificatorio debería, en la medida de lo posible, contar con los costos asociados a las malas clasificaciones.

2.3 Observaciones

2.3.1 Incluyendo variables cualitativas

Hasta ahora la discusión ha sido en torno a que las variables discriminatorias X1, X2, ... , XP tengan unidades naturales para su medición.

Esto es, cada variable puede, en principio, asumir cualquier número real, y estos números pueden ser registrados. A veces, una variable cualitativa o categórica puede ser un discriminador útil (clasificador).

Por ejemplo, la presencia o ausencia de una característica tal como el color rojo puede ser un clasificador justificadamente útil. Esta situación se maneja frecuentemente al crear una variable X cuyos valores numéricos sean uno (1) si el objeto posee la característica y cero (0) si el objeto no posee la característica.

La tabla es luego tratada como las variables medidas usualmente en el proceso discriminatorio – clasificatorio.

Hay muy poca teoría disponible para manejar el caso en que algunas variables son continuas y algunas otras cualitativas. Los experimentos de simulación computacional indican que la función linear de clasificación de Fisher [ver bibliografía] puede responder pobre o satisfactoriamente, dependiendo de las correlaciones entre las variables cualitativas y continuas.

Cuando un número de variables son del tipo 0 – 1, podría ser mejor el considerar una aproximación mejor, llamada

aproximación de regresión logística hacia la clasificación. La probabilidad de pertenencia hacia el primer grupo, p1 (x), se modela directamente como:

$$p_1(x) = \frac{e^{a+b'x}}{1+e^{a+b'x}}$$

Para el problema de dos poblaciones. Él α necesita ser ajustado para acomodar una distribución a priori, y puede no ser fácil el incluir costos. Si las poblaciones son cercanamente normales con matriz de covariancias iguales, la aproximación linear de clasificación es mejor.

En el siguiente capítulo se notará cuán coherente es la utilización de esta función logística o sigmoidea para la tarea discriminatoria – clasificatoria.

2.3.2 Observaciones finales

Hemos querido explicar el problema de clasificación desde la óptica genérica del análisis discriminante. Aunque esta tesis no incluye un análisis discriminante tradicional, si es válido

decir que el proceso clasificatorio que lleva a cabo una red neuronal tiene naturaleza discriminante y la tarea o fin, es exactamente el mismo, aunque los procedimientos sean distintos.

Debido a esto, no se incluye más que una explicación somera del análisis discriminante en este capítulo ya que, entrar en materia más profunda escapa al alcance de esta tesis.

En esta investigación nos hemos dado cuenta que la función sigmoide tiene un uso común, eso lo corroboramos en el tercer capítulo.