ANÁLISIS DISCRIMINATE

DR. JORGE RAÚL PÉREZ GALLARDO

raul.perez@cimat.mx

Unidad Aguascalientes

Métodos Multivariados II

Ingreso a la universidad…

Ante el inminente inicio de un nuevo periodo de ingreso para la universidad, la dirección convoca a una reunión con todos los maestros. El motivo, establecer alguna estrategia para aumentar el porcentaje de estudiantes que terminan satisfactoriamente sus estudios. La conclusión a la que se llega después de terminar la reunión es que hay que desarrollar y establecer un mecanismo desde el proceso de ingreso que permita realizar una mejor selección de aquellos aspirantes que demuestren capacidad para terminar la licenciatura.

¿Qué hacer?

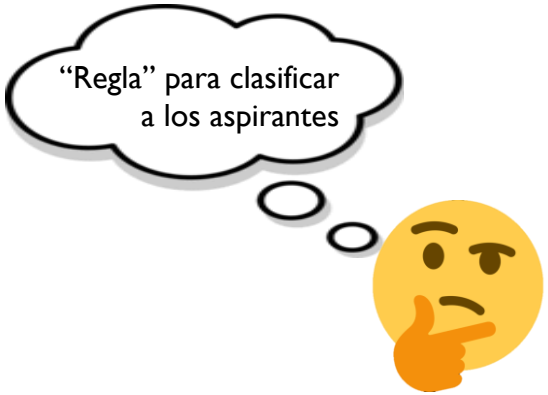
Objetivo

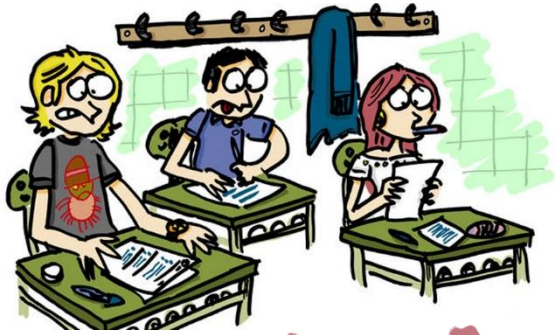
Implementar un mecanismo que mejore la selección de aspirantes de ingreso para aumentar el porcentaje de estudiantes que terminan satisfactoriamente sus estudios.

Ingreso a la universidad…

Establecer una "regla" que permita clasificar a los estudiantes, basado en los puntajes obtenidos en las pruebas de ingreso, en el aquellos que seguramente se titularán, los que no lo terminarán los estudios y aquellos que están en el límite.

La universidad conserva el historial de varias generaciones de los resultados de las evaluaciones de ingreso y de si lograron terminar sus estudios.





No terminan

Titulados

En el límite

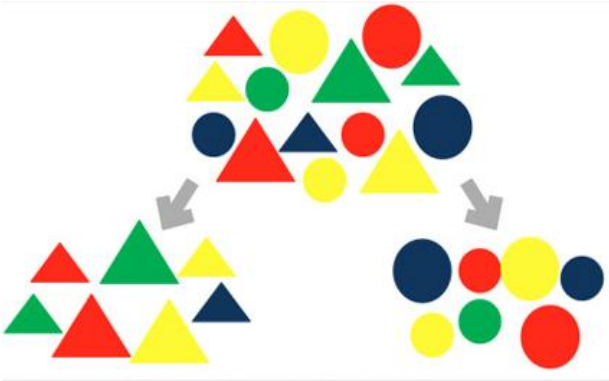
Introducción

*Realizar un **Análisis Discriminante** para clasificar a los aspirantes basado en los resultados de las pruebas.*

Análisis Discriminante

Técnica multivariada capaz de decirnos qué variables permiten **DIFERENCIAR** a *dos o más grupos* y **CREAR** *una "regla"* capaz de *distinguir y clasificar* con precisión *a los miembros* de uno u otro grupo.

Se ***CONOCE***, al inicio del estudio, el número de grupos y las característica de las personas que conforman cada uno de ellos.

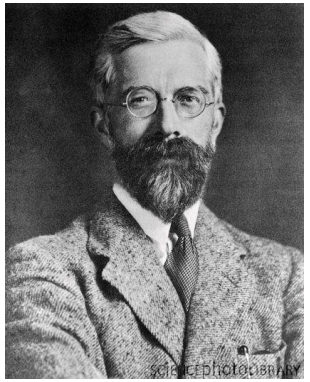


Un poco de historia…

El Análisis Discriminante (DA: *discriminant analysis*) fue desarrollada por Sir Roland Fisher en 1936 gracias a un problema de antropología (los cráneos de las momias egipcias).

En la actualidad, el DA es utilizado en:

- Medicina (distinguir grupos de sujetos patológicos)
- Administración (selección de personal)
- Finanzas (seguros, préstamos)
- Informática (clasificación de e-mail)
- Ingeniería (reconocimiento de patrones)
- Otras disciplinas como educación, antropología, biología, etc.



Sir Roland Fisher

Ejemplos de aplicación

Antes de cualquier intervención quirúrgica es necesario determinar si un anestésico es seguro para una persona.

Con la información del paciente como peso, edad, sexo, raza, presión sanguínea, etc. será posible:

- ¿Construir una regla para separar a pacientes en receptores seguros o inseguros del anestésico?
- ¿Cuáles serían las posibilidades de cometer un error?

Ejemplos de aplicación

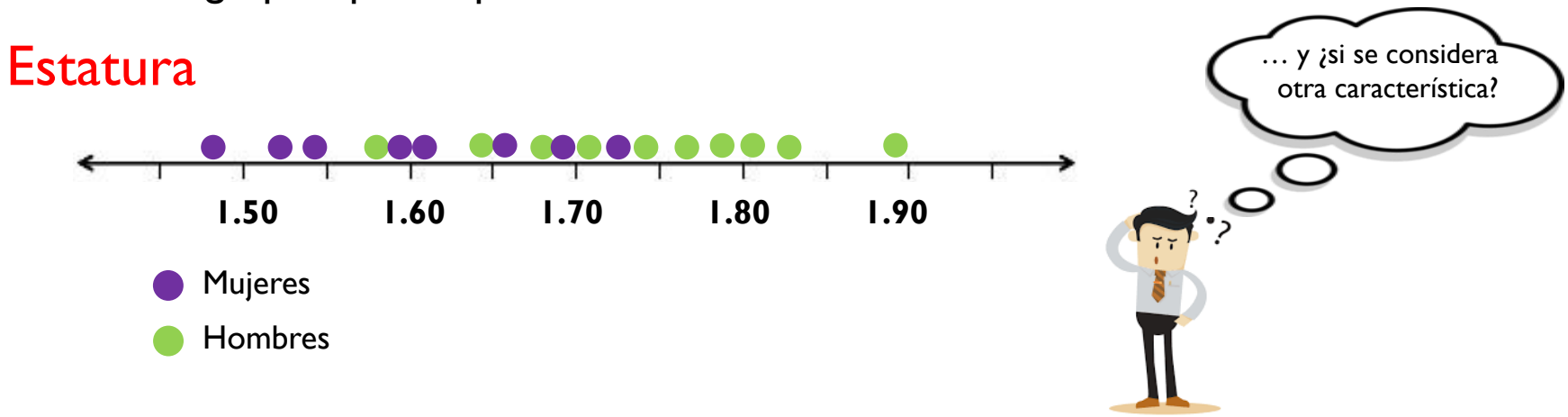
El riesgo de crédito puede provocar mayores pérdidas potenciales a las entidades bancarias, de ahí que sea el que mayor número de crisis financieras ha provocado y al que se dedica mayor atención.

Por lo tanto, los gestores tratan de evitar o por lo menos minimizar cualquier tipo de riesgo al clasificar a sus clientes por su comportamiento financiero.

Principio en que se basa (*caso:"El sexo del paciente"*)

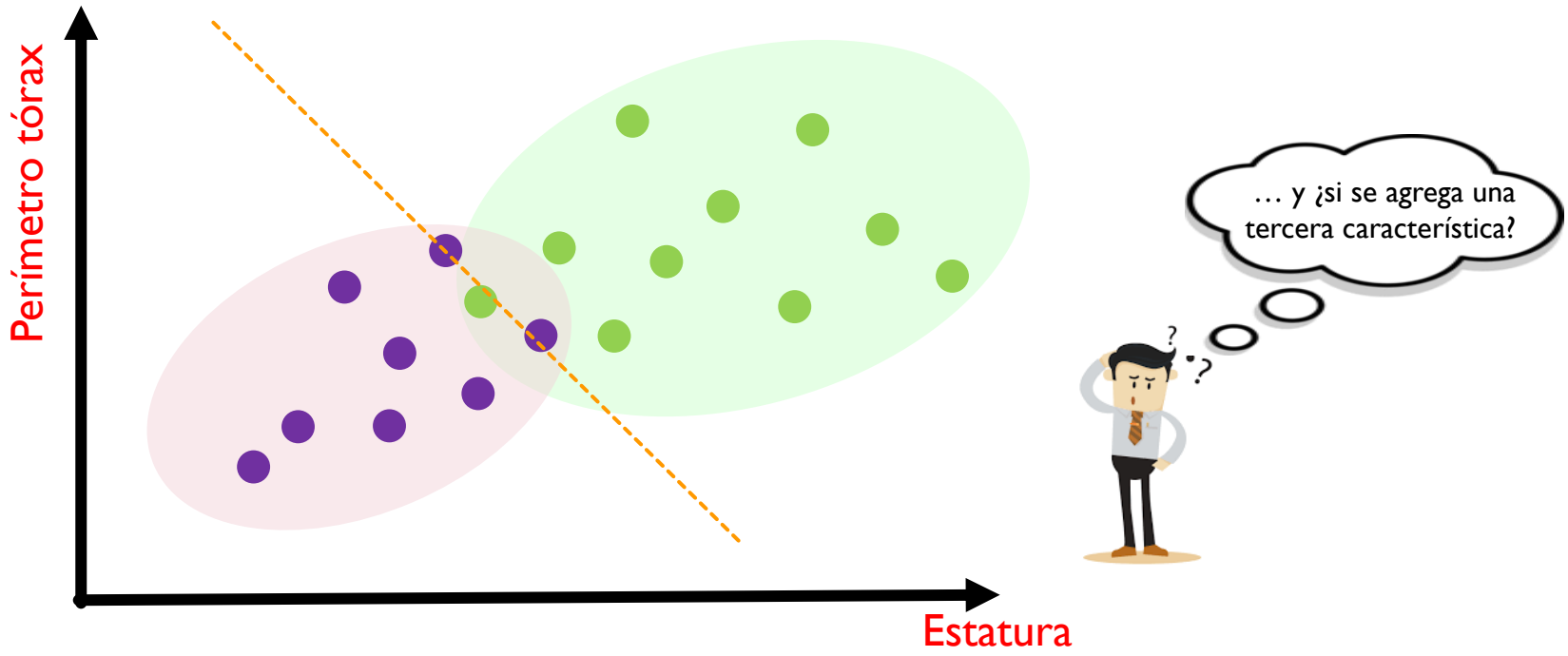
Imagine que debe asignar el sexo a ciertos pacientes que olvidaron marcar dicha opción al momento de su registro y no tiene forma de volverlos a contactar. Considera que la estatura puede servir, por lo que consulta una cierta cantidad de expedientes donde si viene indicado el sexo.

Al graficar los valores observa que no es tan evidente identificar un punto de corte entre ambos grupos que los pueda diferenciar.



Principio en que se basa (*caso:"El sexo del paciente"*)

Se decide agregar el perímetro del tórax. Al realizar la grafica de dispersión ya se observa una mayor separación de ambos grupos



Quizá agregando más variables se pueda establecer de forma clara la región de cada uno de los grupos (hombres/mujeres). Donde se minimice en la medida de lo posible la tasa de error en la clasificación.

Esquema general (*caso: cáncer de pulmón*)

P0. Definir los grupos

Esperanza de vida

1: Menos de un año
2: Entre 1 y 2 años
3: Más de dos años

P2. Establecer una "regla" de separación entre grupos

"Regla"

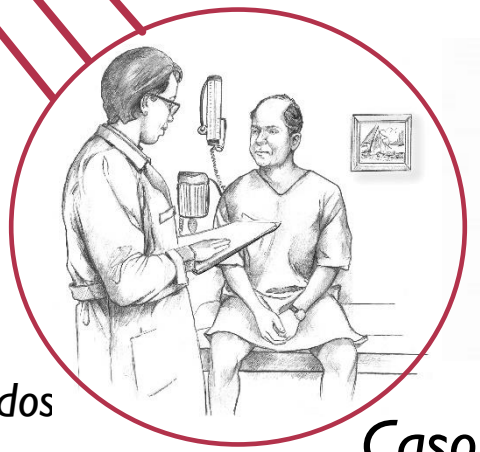
Primera etapa

P1. Selección de variables de interés

- LDH
- Proteínas totales
- Acido úrico
- Hemoglobina
- Leucocitos
- Plaquetas
- …

Segunda etapa

P3. Clasificar nuevo individuos en los grupos establecidos



Caso nuevo

¿Cómo clasificar?

Estadísticamente se distinguen dos enfoques diferentes:

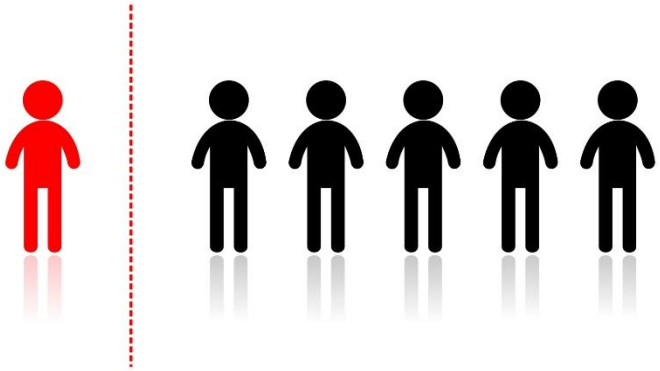
1. *A priori* ***no se conocen los grupos*** y lo que precisamente ***se desea*** es ***establecerlos*** a partir de los datos que poseen. Se aplica

2. Los ***grupos*** están ***bien definidos*** y se trata de ***determinar un criterio para etiquetar cada individuo*** como perteneciente a alguno de los grupos a partir de los valores de una serie limitada de parámetros. Se aplica

¿Discriminar = clasificar?

Discriminar

- Seleccionar excluyendo (RAE).
- Señalar las diferencias entre una cosa y otra
- Separar, diferenciar una cosa de otra.



Clasificar:

- Ordenar o disponer por clases (RAE)
- Acción de organizar o situar algo según determinada directiva, e.g.:
  - Clasificación biológica (taxonomía)
  - Clasificación periódica (tabla periódica)

¿Discriminar = clasificar?

Estadísticamente:

Discriminar

Describir, gráfica o algebraicamente, las características que diferencian a los objetos (observaciones) pertenecientes a una colección (población)

Palabra clave

Clasificar:

Ordenar objetos (observaciones) en dos o más clases mediante una regla previamente establecida.

Palabra clave

Análisis Discriminante

Objetivo:

Estimar la *relación entre* una *ÚNICA variable categórica dependiente y* un *conjunto de variables* métricas *independientes que* mejor *discrimine* los objetos en los grupos definidos a priori.

$$\underset{(categórica)}{Y_1} = \underset{(métricas)}{X_1 + X_2 + X_3 + \cdots + X_n}$$

- La *función discriminante* es la combinación de variables que mejor separen a los grupos.
- El AD es de naturaleza exploratoria.
- También se denomina:

Clasificación supervisada

Supuestos

- Las variables independientes deben seguir una distribución normal multivariante.
- La matriz de varianza-covarianza dentro de cada grupo debe ser aproximadamente iguales (homogénea).
- Ninguna variable discriminante debe de ser combinación lineal de otra variable discriminante, i.e. no debe de existir correlación entre ellas.
- Evitar variables redundantes.
- Los grupos deben ser mutuamente excluyentes.

Restricciones

- Se *requiere* de *al menos dos grupos conocidos y* por cada grupo hay *dos o más casos* (observaciones).
- El más pequeño de los grupos debe tener más casos que variables independientes.
- El *número máximo de variables discriminantes (p)* debe ser menor que el número de objetos u observaciones (n) menos dos:

$$p < (n - 2)$$

- El *número máximo* de *funciones discriminantes* es igual al valor mínimo entre el número de variables discriminantes (p) y el número de grupos (k) menos 1:

$$min\ [p, k - 1]$$

Ejemplo: Las podadoras

Consideremos dos grupos en la ciudad:

π_1: Los propietarios de podadoras
π_2: Los no propietarios de podadoras

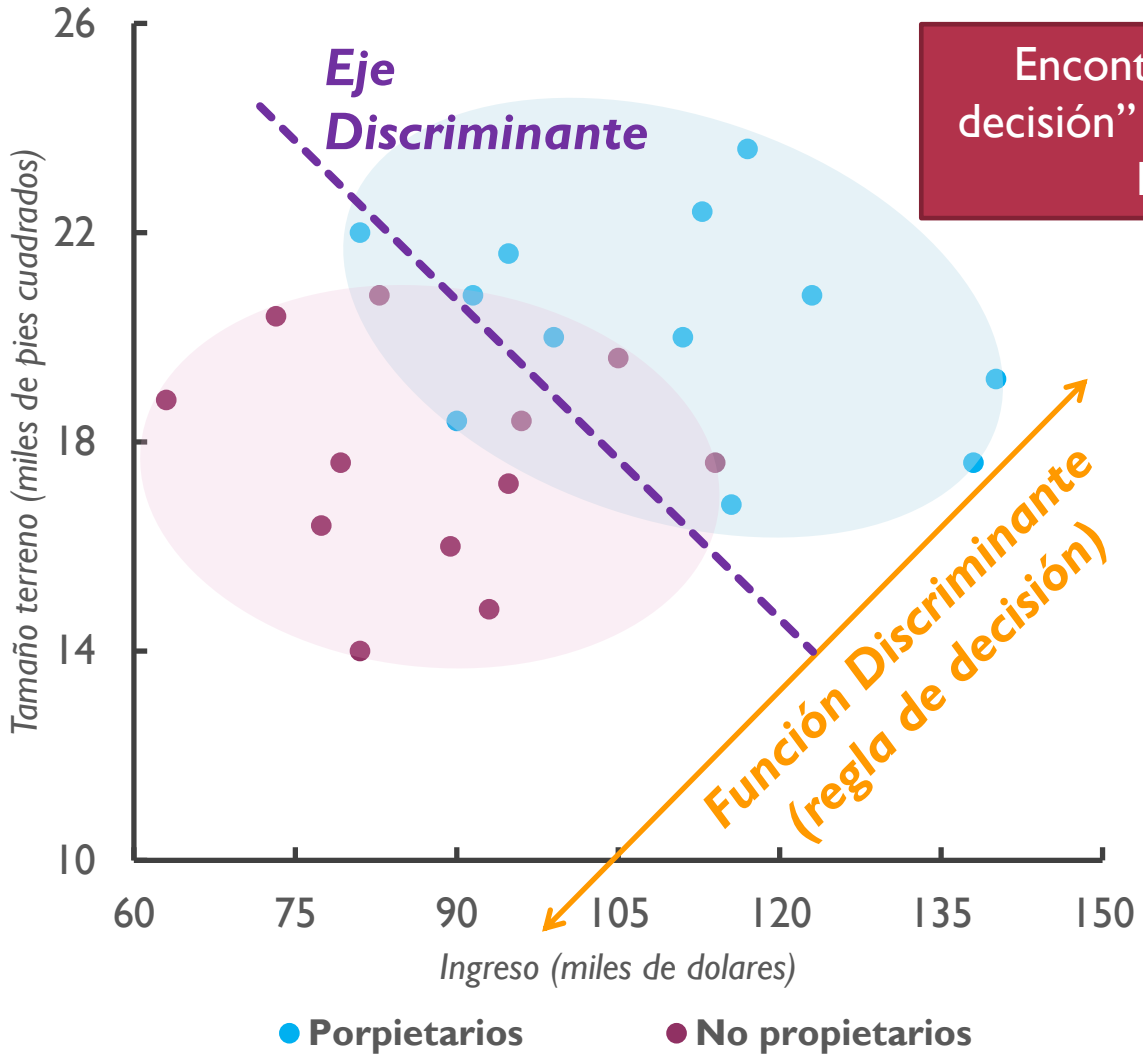
Con el propósito de identificar los mejores prospectos para concentrar los esfuerzos de venta antes de iniciar una campaña intensiva de promoción, un productor de estos tractores esta interesado en clasificarlos en los grupos anteriores basado en:

X_1 = ingreso (miles de dólares)
X_2 = tamaño de terreno (miles de pies cuadrados)

El tamaño e la muestra para cada grupo fue de
$n_1 = n_2 = 12$

Núm. máx. var. discriminante	Funciones discriminantes
$p < (n - 2)$	$min\ [p, k - 1]$
$2 < (12 - 2)$ ✔	$\min [2,\ 2 - 1] = 1$

Ejemplo: Las podadoras



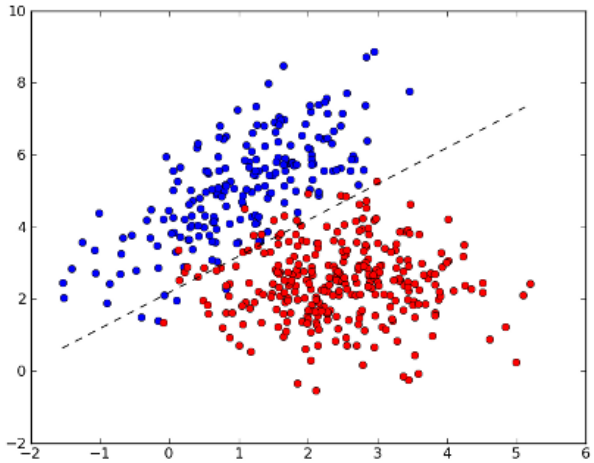
Encontrar la "regla de decisión" que mejor separe las clases

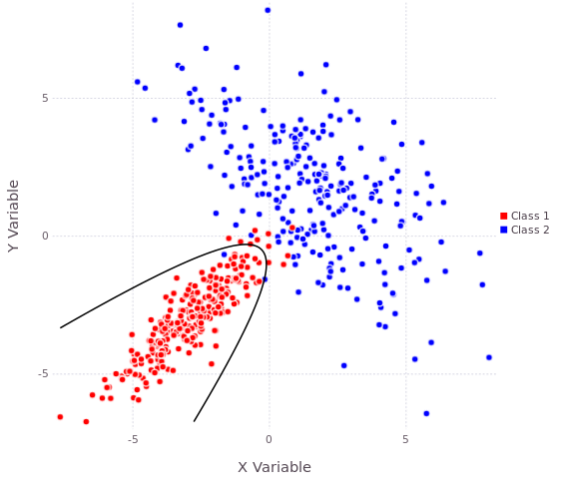
Tipo de análisis discriminante

Análisis Discriminante Lineal (LDA)

Análisis Discriminante Cuadrático (QDA)

2 grupos





3 grupos