UNIVERSIDAD DE PUERTO RICO
Recinto de Río Piedras
Facultad de Administración de Empresas
Instituto de Estadística

ANALISIS DISCRIMINANTE, HERRAMIENTA EN ESTADISTICA GERENCIAL

José C. Vega Vilca, PhD

Presentación en la Escuela Graduada

Marzo 2008

INTRODUCCION

En negocios hay muchas situaciones donde sujetos en estudio pueden ser separados en dos o más grupos bien definidos. Estos sujetos pueden ser personas, ciudades, universidades, países u otros. El propósito del *Análisis Discriminante* es construir un *clasificador* basado en datos multivariados, pertenecientes a grupos bien conocidos por el investigador, para ser usado en clasificación de nuevos sujetos y puedan ser localizados en alguno de estos grupos en estudio.

Según las características (multivariadas) de los nuevos sujetos, podremos dar respuesta a casos tales como:

1.- ¿Comprará, este cliente nuestro producto, o no?

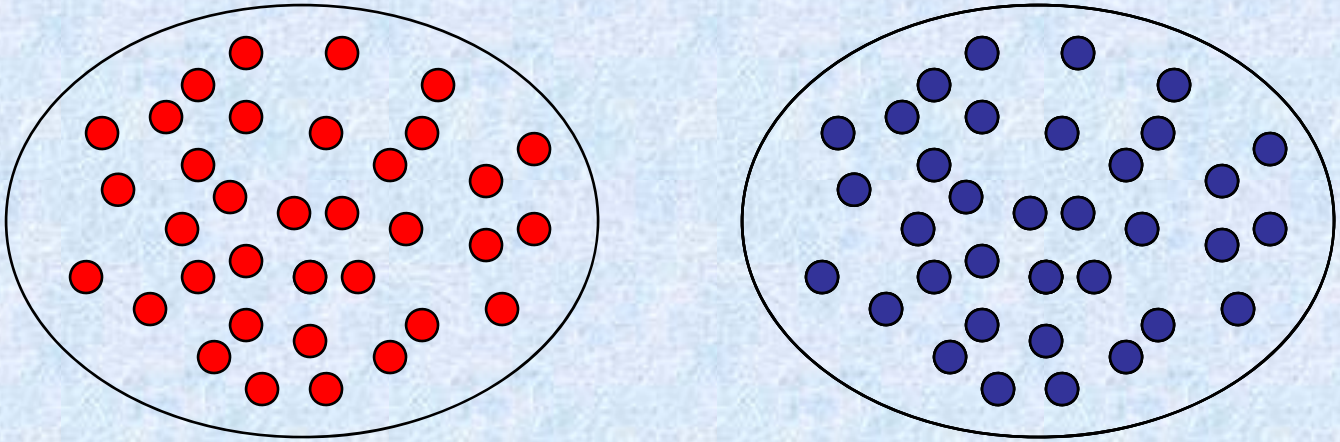
2.- ¿Devolverá, este cliente el crédito, o no?

3.- ¿Se adaptará, este candidato al puesto de trabajo, o no?

EL PROBLEMA GENERAL EN CLASIFICACIÓN

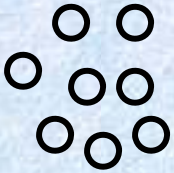
Población 1: π_1

Población 2: π_2

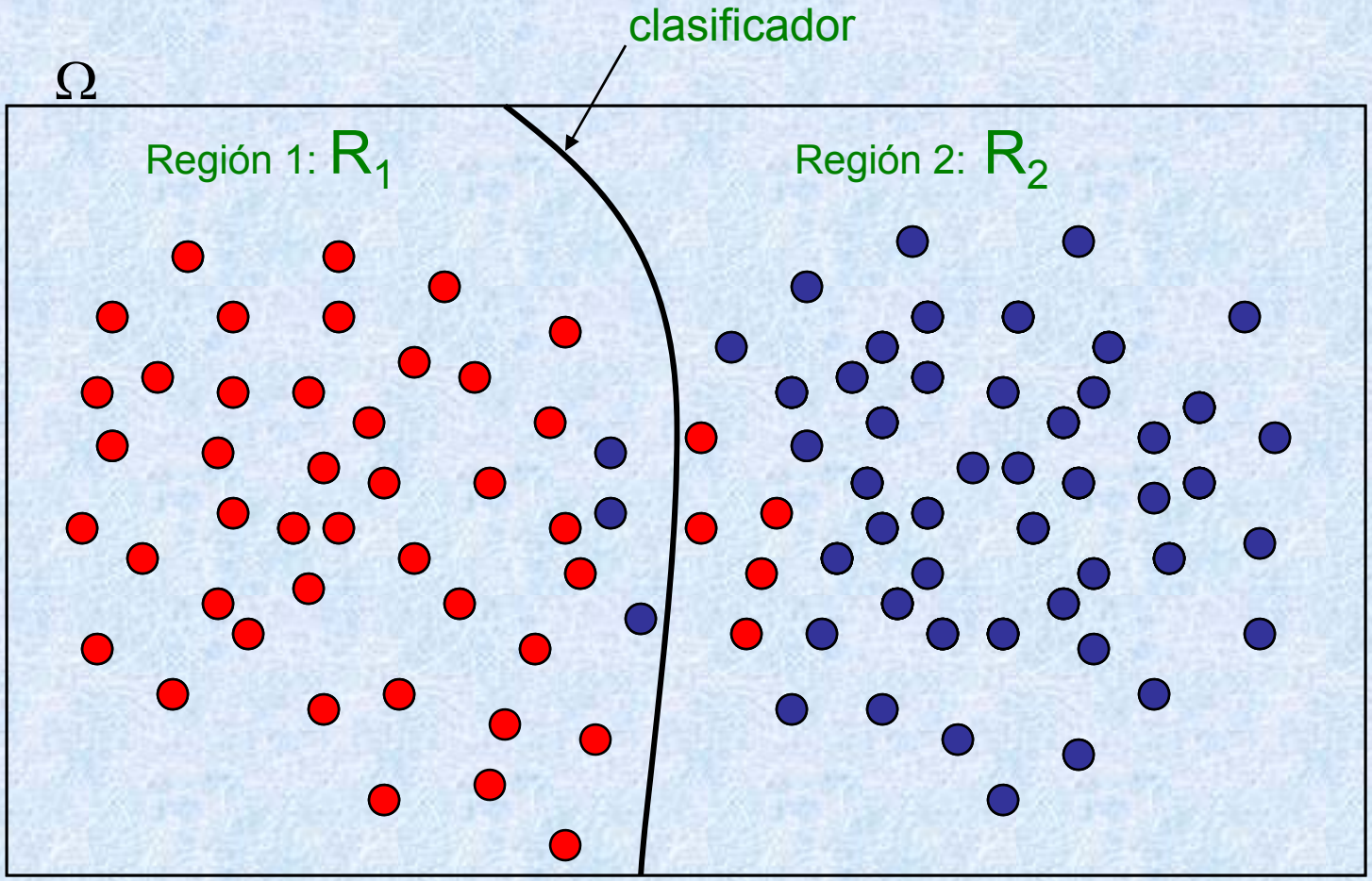


Cada punto representa un sujeto, en el espacio de p-dimensiones

Clasificar estos nuevos objetos, en una de estas dos poblaciones



CASO PRÁCTICO: los objetos se representan dentro de un espacio muestral Ω



CONCEPTOS

R_1 : Región donde los sujetos son clasificados como perteneciente a π_1

R_2 : Región donde los sujetos son clasificados como perteneciente a π_2

Ω : Espacio muestral, $R_1 \cup R_2 = \Omega$ $\quad$ $R_1 \cap R_2 = \phi$

c(i | j) : costo de clasificar un objeto en π_i , cuando realmente pertenece a π_j

$\mathbf{x} = (x_1, x_2, x_3, \cdots, x_p)'$ vector aleatorio : punto en el espacio

$f_i(\mathbf{x})$ función de densidad de población i, i=1,2

$p_1 = P(\mathbf{x} \in \pi_1)$ Probabilidad apriori de pertenecer a población 1

$p_2 = P(\mathbf{x} \in \pi_2)$ Probabilidad apriori de pertenecer a población 2

REGIONES DE CLASIFICACIÓN

$$R_1 \;:\; \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left[\frac{c(1\,|\,2)}{c(2\,|\,1)}\right]\left[\frac{p_2}{p_1}\right]$$

$$R_2 \;:\; \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left[\frac{c(1\,|\,2)}{c(2\,|\,1)}\right]\left[\frac{p_2}{p_1}\right]$$

Generalmente se asume que: $\left[\frac{c(1\,|\,2)}{c(2\,|\,1)}\right] = 1 \;\; y \;\; \left[\frac{p_2}{p_1}\right] = 1$

Regiones de clasificación

$$R_1 \;:\; D(\mathbf{x}) \geq 0$$
$$R_2 \;:\; D(\mathbf{x}) < 0$$

Es la función discriminante

$$D(x) = \ell n \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}$$

DISTRIBUCIÓN NORMAL p-VARIADA

El vector aleatorio $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}'$ tiene distribución normal p-variada, si su función de densidad es :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\}$$

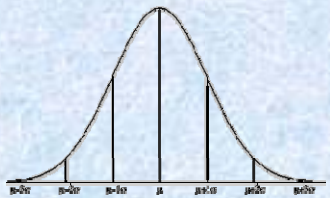
vector de medias

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Matriz de covarianzas

p = 1: univariada

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$



p = 2: bivariada

CLASIFICACIÓN EN DOS POBLACIONES NORMALES

Sean $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ las funciones de densidad correspondientes a las poblaciones en estudio: π_1 y π_2 , respectivamente

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\ |\Sigma_1|^{1/2}} \exp\left\{ -\frac{1}{2}\ (\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \right\}$$

$$f_2(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\ |\Sigma_2|^{1/2}} \exp\left\{ -\frac{1}{2}\ (\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \right\}$$

Regiones de clasificación

$$R_1 : D(\mathbf{x}) \geq 0$$
$$R_2 : D(\mathbf{x}) < 0$$

Es la función discriminante

$$D(\mathbf{x}) = \ell n \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}$$

Caso a) : $\Sigma_1 = \Sigma_2 = \Sigma$ (discriminante lineal)

$$R_1 \;:\; D(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \Sigma^{-1} \mathbf{x} \;-\; \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \;\geq\; 0$$

$$R_2 \;:\; D(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \Sigma^{-1} \mathbf{x} \;-\; \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \;<\; 0$$

Caso b) : $\Sigma_1 \neq \Sigma_2$ (discriminante cuadrático)

$$R_1 \;:\; D(\mathbf{x}) = -\frac{1}{2} \, \mathbf{x}' \left(\Sigma_1^{-1} - \Sigma_2^{-1} \right) \mathbf{x} \;+\; \left(\boldsymbol{\mu}_1' \Sigma_1^{-1} - \boldsymbol{\mu}_2' \Sigma_2^{-1} \right) \mathbf{x} \;-\; k \;\geq\; 0$$

$$R_2 \;:\; D(\mathbf{x}) = -\frac{1}{2} \, \mathbf{x}' \left(\Sigma_1^{-1} - \Sigma_2^{-1} \right) \mathbf{x} \;+\; \left(\boldsymbol{\mu}_1' \Sigma_1^{-1} - \boldsymbol{\mu}_2' \Sigma_2^{-1} \right) \mathbf{x} \;-\; k \;<\; 0$$

donde : $$k \;=\; \frac{1}{2} \, \ell n \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) \;+\; \frac{1}{2} \left(\boldsymbol{\mu}_1' \Sigma_1^{-1} \boldsymbol{\mu}_1 \;-\; \boldsymbol{\mu}_2' \Sigma_2^{-1} \boldsymbol{\mu}_2 \right)$$

CLASIFICACIÓN: "PROBABILIDAD POSTERIOR"

$P(\pi_i \mid \mathbf{x}_0)$ probabilidad de clasificar en π_i , un vector observado $\mathbf{x}_0$
i = 1, 2

Usando probabilidad condicional

$$P(\pi_2 \mid \mathbf{x}_0) = \frac{P(\pi_2 \cap \mathbf{x}_0)}{P(\mathbf{x}_0)} = \frac{P(\pi_2)\,P(\mathbf{x}_0 \mid \pi_2)}{P(\mathbf{x}_0)} = \frac{p_2\ f_2(\mathbf{x}_0)}{P(\mathbf{x}_0)} \quad \ldots \quad (1)$$

Usando probabilidad total

$$P(\mathbf{x}_0) = P(\mathbf{x}_0 \cap \pi_1) + P(\mathbf{x}_0 \cap \pi_2)$$

$$= P(\pi_1)\,P(\mathbf{x}_0 \mid \pi_1) + P(\pi_2)\,P(\mathbf{x}_0 \mid \pi_2)$$

$$P(\mathbf{x}_0) = p_1\ f_1(\mathbf{x}_0) + p_2\ f_2(\mathbf{x}_0) \quad \ldots \quad (2)$$

Reemplazando (2) en (1) y asumiendo que $p_1 = p_2$

$$P(\pi_2 \mid \mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)}{f_1(\mathbf{x}_0) + f_2(\mathbf{x}_0)} = \frac{1}{\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} + 1} = \frac{1}{\exp[D(\mathbf{x}_0)] + 1}$$

$$P(\pi_1 \mid \mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)}{f_1(\mathbf{x}_0) + f_2(\mathbf{x}_0)} = \frac{1}{1 + \frac{f_2(\mathbf{x}_0)}{f_1(\mathbf{x}_0)}} = \frac{1}{1 + \exp[-D(\mathbf{x}_0)]}$$

Regla de clasificación

Si $P(\pi_1 \mid \mathbf{x}_0) > P(\pi_2 \mid \mathbf{x}_0) \Rightarrow \mathbf{x}_0$ se clasifica en π_1

de otro modo en π_2

ESTIMACION (1)

En el caso de aplicación, los parámetros $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_1 y Σ_2 son desconocidos. La función discriminante se construye con una muestra de cada población

muestra de población 1

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\Sigma_1 \Rightarrow \mathbf{S}_x$$

$$\boldsymbol{\mu}_1 \Rightarrow \bar{\mathbf{x}}$$

muestra de población 2

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mp} \end{pmatrix}$$

$$\Sigma_2 \Rightarrow \mathbf{S}_y$$

$$\boldsymbol{\mu}_2 \Rightarrow \bar{\mathbf{y}}$$

ESTIMACION (2)

En discriminante lineal se supone que: $\Sigma_1 = \Sigma_2$

La matriz de covarianza común:

$$\mathbf{S} = \frac{(n-1)\mathbf{S}_x + (m-1)\mathbf{S}_y}{n+m-2}$$

REGIONES DE CLASIFICACION

$$R_1 \;:\; D(\mathbf{x}) = (\bar{\mathbf{x}} - \bar{\mathbf{y}})'\, \mathbf{S}^{-1}\mathbf{x} \;-\; \frac{1}{2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})'\, \mathbf{S}^{-1}(\bar{\mathbf{x}} + \bar{\mathbf{y}}) \;\geq\; 0$$

$$R_2 \;:\; D(\mathbf{x}) = (\bar{\mathbf{x}} - \bar{\mathbf{y}})'\, \mathbf{S}^{-1}\mathbf{x} \;-\; \frac{1}{2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})'\, \mathbf{S}^{-1}(\bar{\mathbf{x}} + \bar{\mathbf{y}}) \;<\; 0$$

CLASIFICACION

Un nuevo sujeto representado por: $\mathbf{x}_0 = (x_1, x_2, x_3, \cdots, x_p)'$

- Será clasificado en: $\pi_1 \quad si \quad D(\mathbf{x}_0) \geq 0$
- Será clasificado en: $\pi_2 \quad si \quad D(\mathbf{x}_0) < 0$

APLICACION 1: Clasificación en dos grupos

Una empresa tiene el registro de 84 clientes. Algunos de ellos están suscritos a la revista *Wall Street Journal* (WSJ) y los otros no.

Grupos:

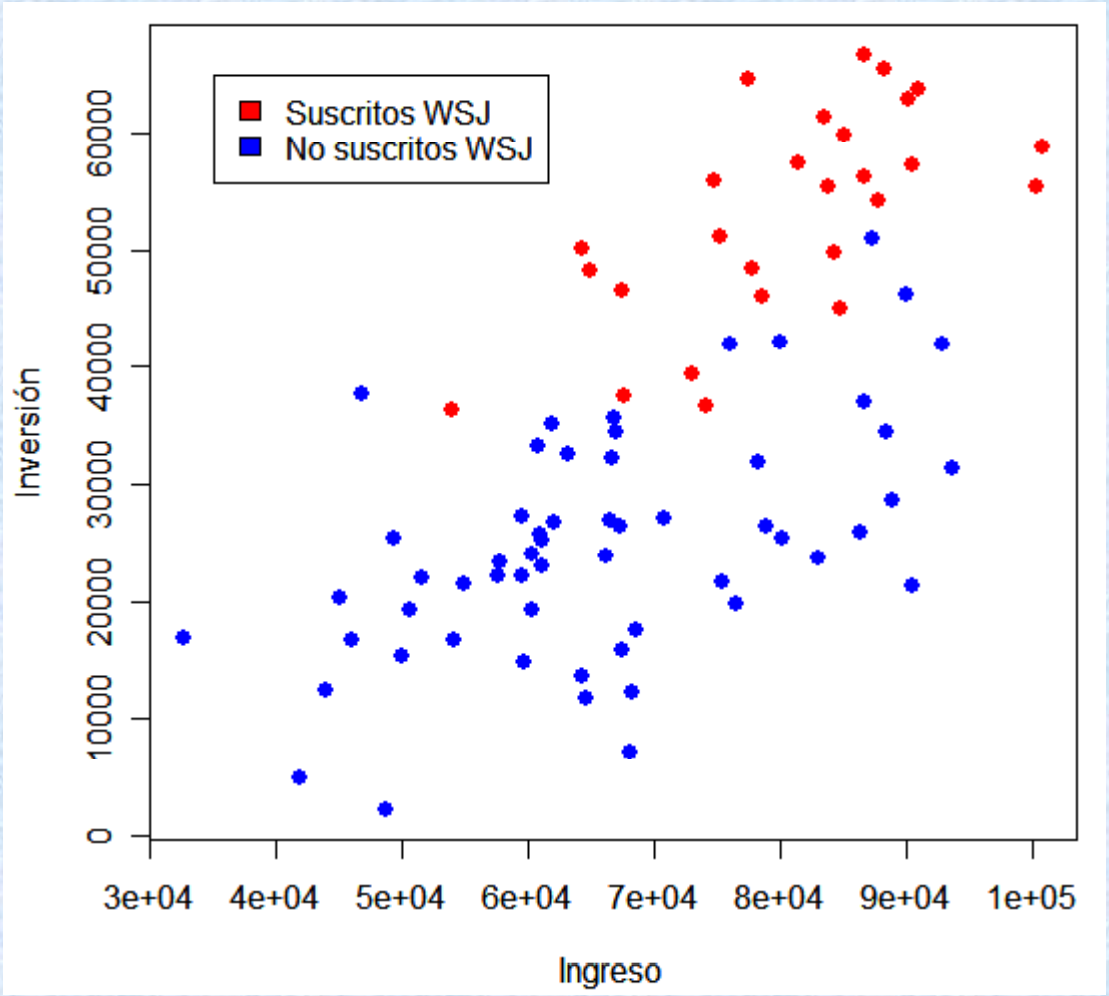
1) Clientes NO suscritos a Wall Street Journal

2) Clientes suscritos a Wall Street Journal

Variables discriminantes:

- Ingreso: ingreso anual de la persona
- Inversión: cantidad total invertido en bonos y acciones

GRAFICO DE PUNTOS: REPRESENTACION DE LOS DOS GRUPOS



ANALISIS DISCRIMINANTE (1)

PROMEDIOS

grupo	Ingreso	Inversión
N0 SUSCRITOS:NO	66042.11	24952.63
SUSCRITOS: SI	80485.19	53000.00

MATRIZ DE COVARIANZAS COMUN

	Ingreso	Inversión
Ingreso	14812033021	6123163684
Inversión	6123163684	7663422105

MATRIZ DE CLASIFICACION

	N0-suscrito	SI-suscrito	TOTAL
NO-suscritos	52	5	57
SI-suscritos	2	25	27
			84

ANALISIS DISCRIMINANTE (2)

FUNCIÓN DISCRIMINANTE (D)

$$D(\mathbf{x}) = 0.00006586 * Ingreso - 0.0003527348 * Inversión + 8.92316$$

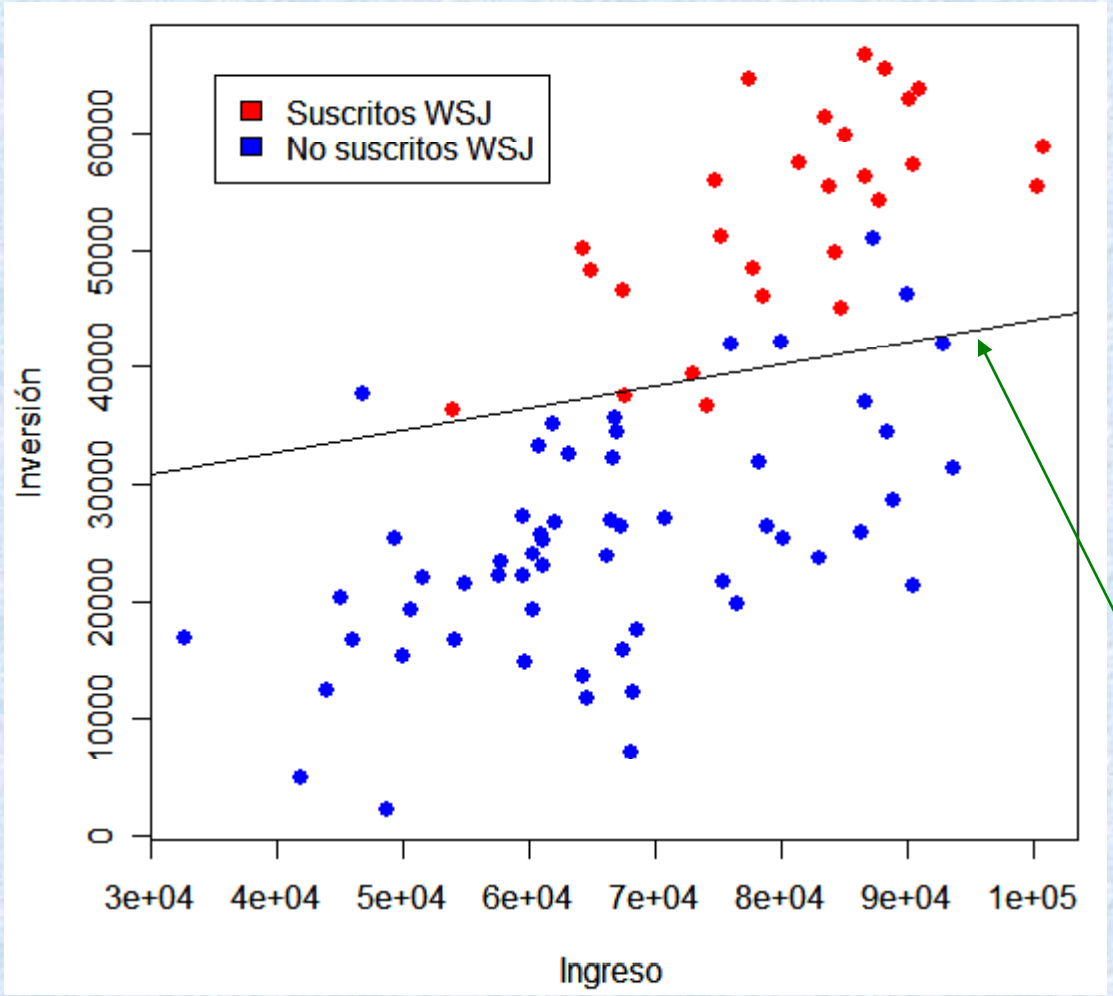
ECUACION DE LA RECTA: CLASIFICADOR

$$D(\mathbf{x}) = 0$$

$$0.00006586 * Ingreso - 0.0003527348 * Inversion + 8.92316 = 0$$

$$Inversión = 0.1867125 * Ingreso + 25297.08$$

REPRESENTACION GRAFICA DEL CLASIFICADOR



Malaclasificacion:

5 azules

2 rojos

clasificador

COMO CLASIFICAR A UN NUEVO CLIENTE?

Nuevo cliente:

Ingreso: 60000

Inversión: 10000

$$\mathbf{x}_0 = (60000,\ 10000)'$$

- Solución 1: Ubicar $\mathbf{x}_0$ en el gráfico de puntos

- Solución 2: Usar la función discriminante

$$D(\mathbf{x}_0) = 0.00006586 * \mathbf{60000} - 0.0003527348 * \mathbf{10000} + 8.92316$$

$$D(\mathbf{x}_0) = 9.347412 > 0$$

El nuevo cliente No está suscrito a WSJ

COMO CLASIFICAR A UN NUEVO CLIENTE?

- Solución 3: usando probabilidad posterior ($p_1 = p_2$)

$$P(\pi_2 \mid \mathbf{x}_0) = \frac{1}{\exp[D(\mathbf{x}_0)] + 1} = \frac{1}{\exp(9.347412) + 1} = 0.0000872$$

$$P(\pi_1 \mid \mathbf{x}_0) = \frac{1}{1 + \exp[-D(\mathbf{x}_0)]} = \frac{1}{1 + \exp(-9.347412)} = \mathbf{0.9999128}$$

Se cumple que:

$$P(\pi_1 \mid \mathbf{x}_0) > P(\pi_2 \mid \mathbf{x}_0)$$

El nuevo cliente No está suscrito a WSJ

CLASIFICACIÓN EN MÁS DE DOS GRUPOS

Trabajando con tres grupos

- Usando la función discriminante: asumiendo que $p_1 = p_2 = p_3$

clasificar $\mathbf{x}_0$ en una de las tres poblaciones

$$\pi_1 \quad si \quad D_{12}(\mathbf{x}_0) \geq 0 \quad \wedge \quad D_{13}(\mathbf{x}_0) \geq 0$$

$$\pi_2 \quad si \quad D_{12}(\mathbf{x}_0) < 0 \quad \wedge \quad D_{23}(\mathbf{x}_0) \geq 0$$

$$\pi_3 \quad si \quad D_{13}(\mathbf{x}_0) < 0 \quad \wedge \quad D_{23}(\mathbf{x}_0) < 0$$

- Usando probabilidad posterior

calcular : $P(\pi_i \mid \mathbf{x}_0) \quad , \quad i = 1, 2, 3$

clasificar $\mathbf{x}_0$ en la población donde $P(\pi_i \mid \mathbf{x}_0)$ es el valor más grande

APLICACION 2: Clasificación en tres grupos

Una compañia especializada en textos universitarios, es representante de un libro de computación con el cual ha alcanzado sus mejores ventas. La compañia tiene registrado a 119 universidades en tres grupos:

Grupos:

1) Universidades que NUNCA le compraron el libro
2) Universidades que YA NO compran el libro
3) Universidades que SIGUEN comprando el libro

Variables discriminantes:

- X1: total de alumnos en la universidad
- X2: promedio SAT
- X3: porcentaje de cursos que requieren asistencia
- X4: número de PC disponibles en la universidad
- X5: porcentaje de estudiantes con PC propia
- X6: promedio anual de estudiantes matriculados

ANALISIS DISCRIMINANTE (1)

PROMEDIOS:

GRUPO	X1	X2	X3	X4	X5	X6
1)	14799.05	1134.2750	80.59500	148.1500	54.93250	14997.500
2)	14888.46	921.7297	59.87838	101.6216	43.95135	9878.378
3)	19575.60	950.1429	59.48333	153.2857	51.98571	9680.952

MATRIZ DE COVARIANZAS COMUN

	X1	X2	X3	X4	X5	X6
X1	503920787.2	-6095139.5	-414582.4	1.40e+06	-689418.9	-229036851
X2	-6095139.5	1016284.4	28703.1	8.29e+04	58951.3	13996126
X3	-414582.4	28703.1	4432.8	1.93e+01	2743.5	1045239
X4	1403197.0	82864.9	19.3	8.30e+04	3038.1	3234841
X5	-689418.9	58951.3	2743.5	3.04e+03	11217.5	1787913
X6	-229036851.2	13996125.6	1045238.9	3.23e+06	1787912.9	1117937215

ANALISIS DISCRIMINANTE (2)

MATRIZ DE CLASIFICACION

	GRUPO1	GRUPO2	GRUPO3	TOTAL
GRUPO1	39	0	1	40
GRUPO2	0	34	3	37
GRUPO3	0	3	39	42
				119

FUNCION DISCRIMINANTE

El clasificador consta de tres funciones:

$$D_{12}(\mathbf{x}) \quad D_{13}(\mathbf{x}) \quad D_{23}(\mathbf{x})$$

Coeficientes de las funciones

Variables	$D_{12}(\mathbf{x})$	$D_{13}(\mathbf{x})$	$D_{23}(\mathbf{x})$
X1	0.0002646156	-0.0007530045	-0.0010176201
X2	0.0103142274	0.0145923633	0.0042781359
X3	0.5711362115	0.5000922141	-0.0710439974
X4	0.0583431022	-0.0060218883	-0.0643649905
X5	-0.0544522169	-0.2357332707	-0.1812810538
X6	-0.0001594750	0.0001415569	0.0003010319
Constante	-57.25612	-25.52921	31.72691

COMO CLASIFICAR A UNA NUEVA UNIVERSIDAD?

Nueva universidad:

X1	X2	X3	X4	X5	X6
17455	1068	79.3	154	46.5	17400

- Solución 1: Usar la función discriminante

$$D_{12}(\mathbf{x}_0) = 7.347388$$

$$D_{13}(\mathbf{x}_0) = 7.143177$$

$$D_{23}(\mathbf{x}_0) = -0.2042108$$

La nueva universidad pertenence al grupo 1. Nunca comprarán el libro

$$D_{12}(\mathbf{x}_0) \geq 0 \ \wedge \ D_{13}(\mathbf{x}_0) \geq 0$$

COMO CLASIFICAR A UN NUEVA UNIVERSIDAD?

- Solución 2: Usar la probabilidad posterior ($p_1 = p_2 = p_3$)

$$D_{12}(\mathbf{x}_0) = 7.347388$$

$$D_{13}(\mathbf{x}_0) = 7.143177$$

$$D_{23}(\mathbf{x}_0) = -0.2042108$$

$$P(\pi_3 \mid \mathbf{x}_0) = \frac{1}{\exp[D_{13}(\mathbf{x}_0)] + \exp[D_{23}(\mathbf{x}_0)] + 1} = 0.0007891055$$

$$P(\pi_2 \mid \mathbf{x}_0) = \frac{1}{\exp[D_{12}(\mathbf{x}_0)] + 1 + \exp[-D_{23}(\mathbf{x}_0)]} = 0.0006433502$$

$$P(\pi_1 \mid \mathbf{x}_0) = \frac{1}{1 + \exp[-D_{12}(\mathbf{x}_0)] + \exp[-D_{13}(\mathbf{x}_0)]} = \mathbf{0.9985675}$$

La nueva universidad pertenence al grupo 1. NUNCA comprarán el libro

BIBLIOGRAFIA

Albright S., Winston W., Zappe C. (2000). Managerial Statistics, Duxbury

Mardia K., Kent J., Bibby J. (1979). Multivariate Analysis, Academic Press