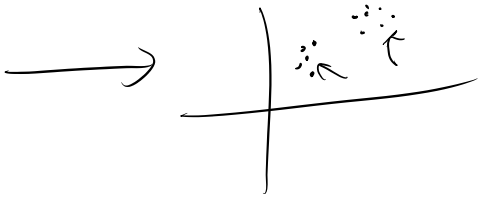
Discriminación y clasificación

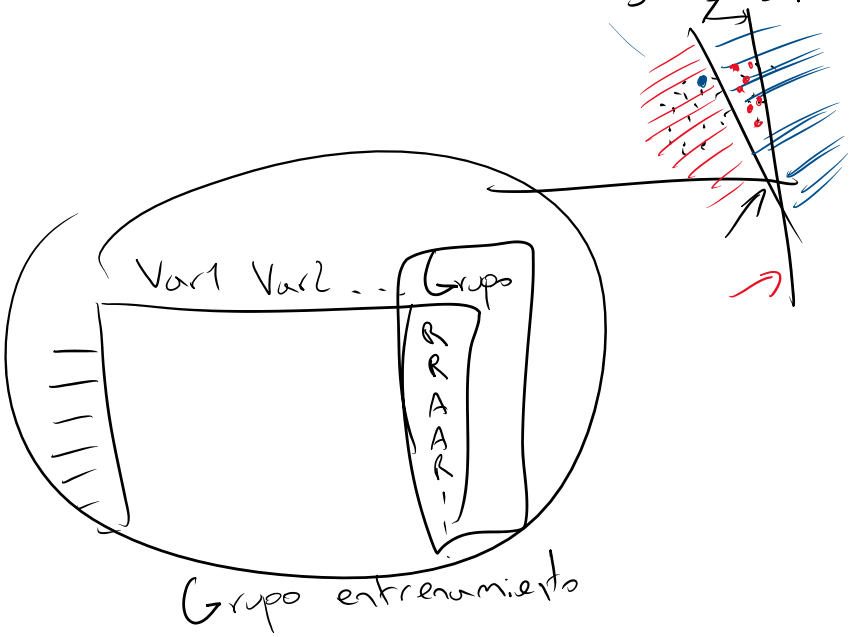
miércoles, 3 de mayo de 2023 13:55

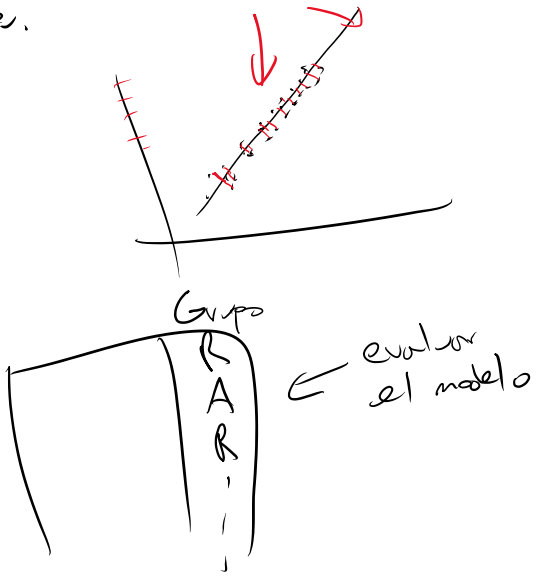
Objetivo: Separar datos (observaciones) y posicionarlos en grupos diferentes, posiblemente definidos previamente

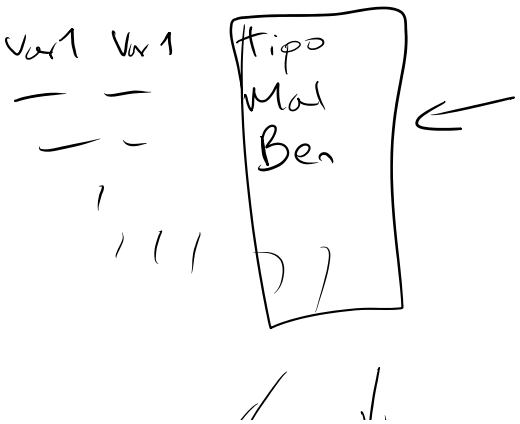
Discriminación: Normalmente es exploratorio.
Usado para investigar/ratificar diferencias en objetos representados en la muestra

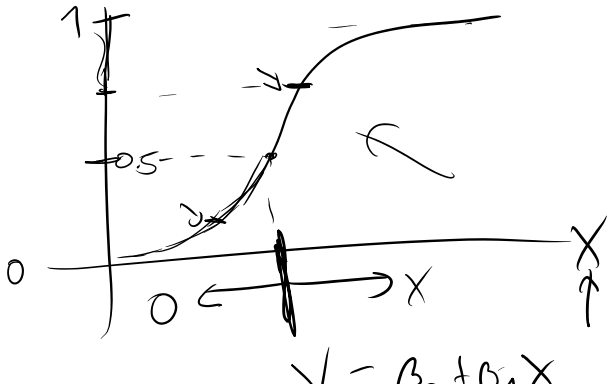


Clasificación: Mera exploratorio.
Se fijan reglas de antemano para clasificar nuevas objeto/observaciones.

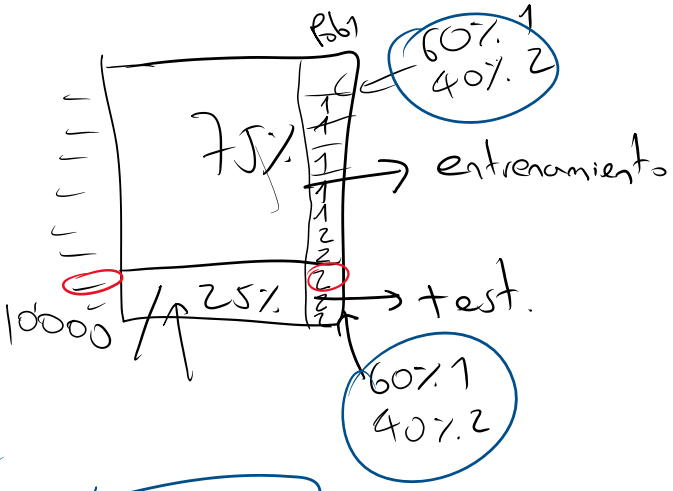


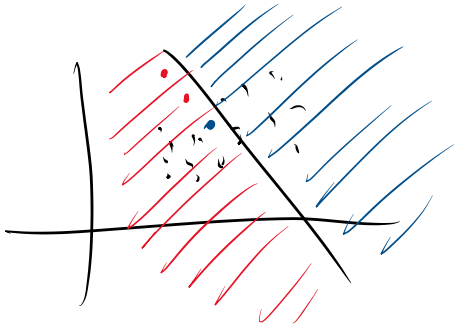






$Y = \beta_0 + \beta_1 X$





98%

Concretamente:

1) Describir gráficamente o algebraicamente diferencias entre objetos de distintas poblaciones conocidas

2) Organizar en 2 o más grupos

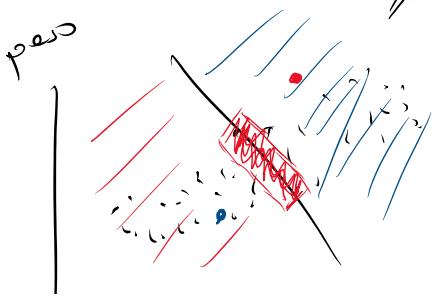
↳ derivar una regla "óptima" de clasificación para asignar nuevos objetos.

Clasificación de 2 poblaciones.

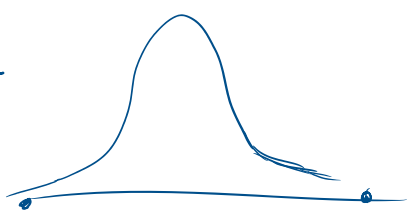
Se busca separar 2 clases de observaciones o asignar nuevas observaciones a una de las dos clases

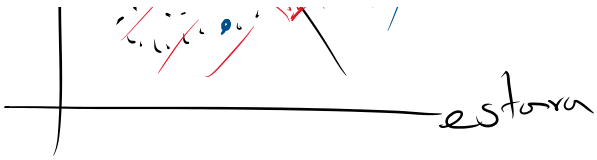
Clase Π_1 y Π_2

Los objetos se separan o se clasifican basados en las p variables aleatorias $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$

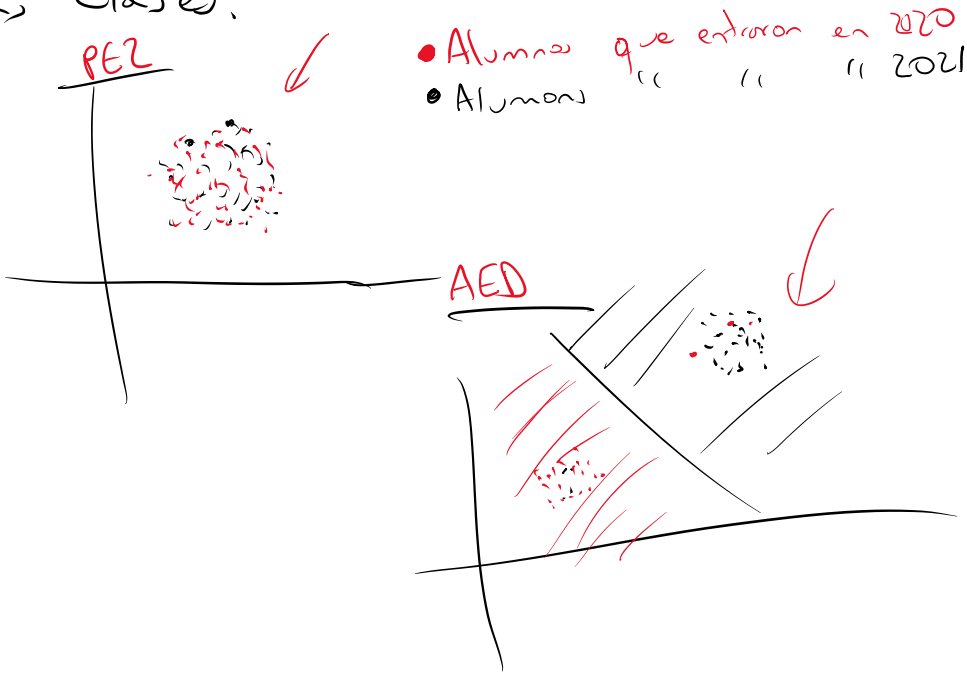


99%





Los valores obs. deberían tener alguna diferencia asociada a las clases.



La clase Π_1 se asocia a una población 1 con función de densidad (PDF) $f_1(X)$

La clase Π_2 se asocia a una población 2 con función de densidad (PDF) $f_2(X)$

Método general

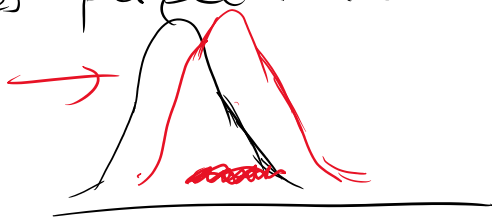
Usar reglas aprendidas de muestras de aprendizaje (o entrenamiento)

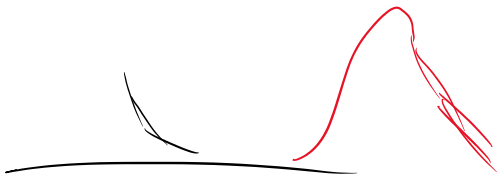
Se examinan observaciones aleatorias cuya población es conocida y se estudian sus diferencias,

Buscamos dividir el conjunto de todas las observ. en 2 regiones tq. si una nueva observ. cae en R_1 se clasifica como Π_1 y si cae en R_2 se clasifica como Π_2

En general, las clasificaciones tienen errores

En general, las clasificaciones tienen errores (la distribución entre Π_1 y Π_2 no es perfecta dados los mediciones)

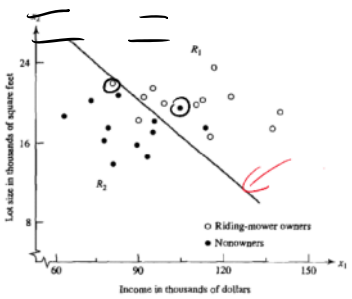




2) Buscamos minimizar la probabilidad de clasificación incorrecta.

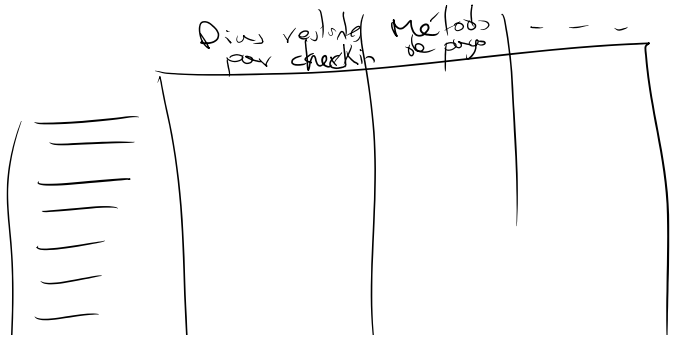
Table 11.1

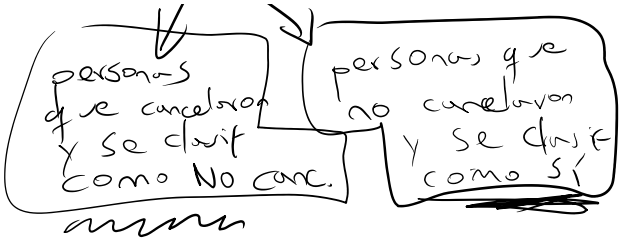
π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft²)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft²)
90.0	18.4	105.0	19.6
115.5	16.8	82.8	20.8
94.8	21.6	94.8	17.2
91.5	20.8	73.2	20.4
117.0	23.6	114.0	17.6
140.1	19.2	79.2	17.6
138.0	17.6	89.4	16.0
112.8	22.4	96.0	18.4
99.0	20.0	77.4	16.4
123.0	20.8	63.0	18.8
81.0	22.0	81.0	14.0
111.0	20.0	93.0	14.8



Obs: 1) Si es muy improbable ser Π_2, no debería clasificar la observ. como Π_2

2) Si es muy costoso clasificar un Π_1 como Π_2, pero no un Π_2 como Π_1 se debería ser cuidadoso.





Sea $f_1(\mathbb{X})$, $f_2(\mathbb{X})$ las PDF's de $\mathbb{X}$ (vect. aleat) $p\times 1$
de las poblaciones Π_1 y Π_2 respectivamente.
Un objeto $\mathbb{X}$ se clasifica como Π_1 o Π_2

Sea Ω el espacio muestral (todos los posibles valores que puede tomar $\mathbb{X}$)

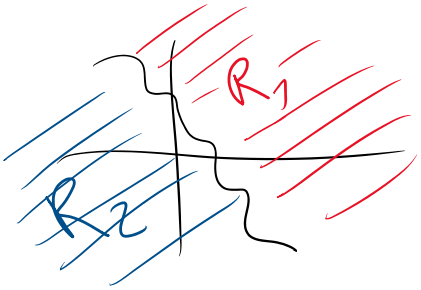
$X_1 \in [0,1]$
$X_2 \in [0,1]$

Sea $R_1 \subseteq \Omega$ región que clasifica como Π_1
$R_2 = \Omega \setminus R_1$ " " " " Π_2

$R_1 \cap R_2 = \phi$ → mutuamente excluyentes

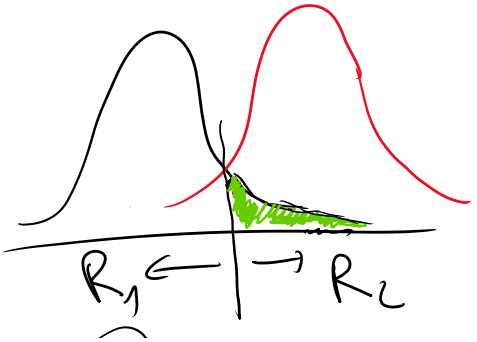
$R_1 \cup R_2 = \Omega$ → exhaustivo.

Ej: si $p=2$ $\quad \Omega = \mathbb{R}^2$

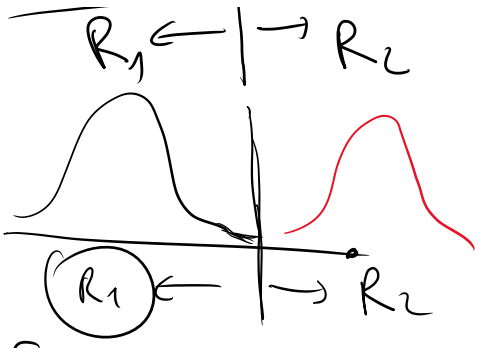


Denotamos $p(2|1)$ → prob. de clasificar objetos como Π_2 siendo de Π_1

$$p(2|1) = P(\mathbb{X} \in R_2 \mid \Pi_1) = \int_{R_2} f_1(\mathbb{X})\, dx$$



R_2



$$P(1|2) = P(X \in R_1 | \pi_2) = \int_{R_1} f_2(x)\, dx$$

Sean p_1, p_2 las probabilidades previas de π_1 y π_2, $p_1 + p_2 = 1$

1) P(obser de π_1 y clasif como π_1)

$= P$(obs $\pi_1 \cap$ clasif. como π_1)

$= P(X \in R_1 | \pi_1) \cdot P(\pi_1)$ → prob. previa de pertenece a π_1

$= P(1|1) \cdot p_1$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2) Clasificación correcta de π_2

$P(2|2) \cdot p_2$

3) P(clasif errónea en π_1)

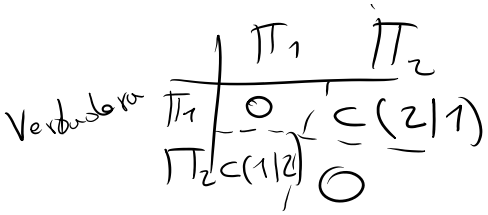
$= P(1|2) \cdot p_2$

4) P(clasif. errónea en π_2)

$P(2|1) p_1$

La matriz de costo de clasificación incorrecta

	clasificación π_1	clasificación π_2



El costo esperado o promedio de clasificación incorrecta

$$ECM = \underbrace{c(2|1)}_{2} \cdot p(2|1)\, p_1 + \underbrace{c(1|2)}_{1} \cdot p(1|2) \cdot p_2$$