RECONOCIMIENTO DE OBJETOS

- INTRODUCCIÓN

  - Clasificación: enfoque basado en la teoría de la Decisión
    - Introducción a problemas de clasificación
    - Enfoque basado en la teoría de la Decisión

PROBLEMA DE APRENDIZAJE BASADO EN DATOS: PLANTEAMIENTO GENERAL

Asumiendo un problema en el que se observa que una variable de salida Y presenta una cierta relación con un conjunto $X = (X_1, X_2, \ldots, X_p)$ de p variables de entrada, de forma general, podemos decir que:

- ***Un problema de aprendizaje basado en datos*** se refiere al conjunto de aproximaciones o técnicas que permiten ***encontrar la función f que establece la relación X-Y:***

$$Y = f(X) + \varepsilon$$

donde:

- f: alguna función fija de $(X_1, X_2, \ldots, X_p)$.
  - → Representa la información sistemática que X proporciona sobre Y.
  - → En general, esta función f es desconocida. En esta situación, se debe estimar f basándose en las observaciones entrada-salida disponibles.
- ε: término de error aleatorio, de media cero, que es independiente de X.

- **OBJETIVO:** ***predicción del valor de la variable de salida a partir de valores de las variables de entrada***

→ Dado un conjunto de variables entrada-salida X-Y, tales que $Y = f(X) + \varepsilon$, el objetivo es encontrar la función f que permita predecir Y con un error mínimo:

donde:

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}$ representa nuestra estimación sobre f
- $\hat{Y}$ representa la predicción resultante de Y

PROBLEMA DE APRENDIZAJE BASADO EN DATOS: PLANTEAMIENTO GENERAL

- **NOTACIÓN Y TERMINOLOGÍA PARA UN PROBLEMA GENERAL:**

Base de datos: compuesta por n ***observaciones*** (también llamadas ***instancias***, ***registros***, ***muestras***)

- Variables entrada-salida del problema:
  - ***Variables de entrada:*** $X = (X_1, X_2, \ldots, X_p)$

    También llamadas: ***predictores***, ***variables independientes***, ***características***, ***descriptores, atributos***

    Número de variables de entrada: p
  - ***Variable de salida:*** Y (también llamada ***respuesta***, ***variable dependiente***)

- Conjunto de datos disponibles: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, con:

  $x_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ e y_i: *valores de las variable de entrada y salida para la observación i*

  X_{ij}: *valor de la variable* X_j *para la observación i con* $i = 1, \ldots, n \quad j = 1, \ldots, p$

Matricialmente:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

CLASIFICACIÓN: se predice una categoría (no el valor de una variable numérica – ejemplo: cuál va a ser el precio de un artículo, o el número de reservas que se harán en mayo en un hotel – PROBLEMA DE REGRESIÓN).

CLASIFICACIÓN SUPERVISADA: las observaciones incluyen una variable respuesta

- Variables entrada-salida:
  - ***Variables de entrada o predictores:*** $X = (X_1, X_2, \ldots, X_p)$ (p predictores)
  - ***Variable de salida o respuesta:*** $Y = \{C_1, C_2, \ldots, C_K\} = \{1, 2, \ldots, K\}$ (K clases)

- Conjunto de datos de entrenamiento (n datos): $\longrightarrow$ $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$; $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

- Observación a clasificar:

  Observación específica: $x_0 = (X_{01}, X_{02}, \ldots, X_{0p})$

  De forma genérica, los valores de los predictores de la observación a clasificar se denotarán por :

  $X = x = (X_1, \ldots, X_p)$

PROBLEMA DE CLASIFICACÍÓN: TERMINOLOGÍA Y DEFINICIONES

- **Clases: son los posibles valores de la variable respuesta,** de la salida del modelo de predicción, categorías o grupos representativos en los que se quieren clasificar los datos.

- **Instancia, ejemplo o registro (*instance, sample, record*):** cada una de las muestras disponibles para entrenar/validar/evaluar un modelo (en los ejemplos anteriores, cada uno de los correos electrónicos; cada objeto cuadrado, circular o triangular; cada figura de madera pera/manzana; cada persona sana/enferma; cada planta).

- **Característica, atributo, propiedad o campo (*feature, attribute, property, field*):** cada instancia se describe por medio de un conjunto de atributos, descriptores o características. (En los ejemplos anteriores, cada una de las medidas que se utilizan para describir un correo electrónico; relación `perímetro^2/area`, etc…)

- **Vector de características, atributos o predictores**: al conjunto de atributos que se utilizan para entrenar el modelo (predictores) y que definen una instancia se le denomina vector de predictores. Hay que tener en cuenta, que los atributos que forman parte de este vector pueden ser el resultado de un proceso de selección de características o filtrado de todos los atributos disponibles.

- **Espacio de características (*feature space*):** espacio definido por cada uno de los atributos que componen el vector de predictores. En este espacio, cada instancia se representa mediante un punto cuyas coordenadas están definidas por los valores que tienen los atributos de dicha instancia.

- **Conjunto de datos (*dataset*):** el conjunto de datos está formado por instancias; cada instancia se compone de los valores de los predictores que conforman el vector de atributos. Además, en aprendizaje supervisado, cada instancia está etiquetada con la codificación asignada a la clase a la que pertenece.
  - **Conjunto de entrenamiento (patrones de entrenamiento):** subconjunto de datos utilizados en la fase de aprendizaje para el diseño y entrenamiento del modelo (en ocasiones este conjunto se subdivide en entrenamiento + validación).
  - **Conjunto de test:** subconjunto de datos utilizados en la evaluación del modelo entrenado.

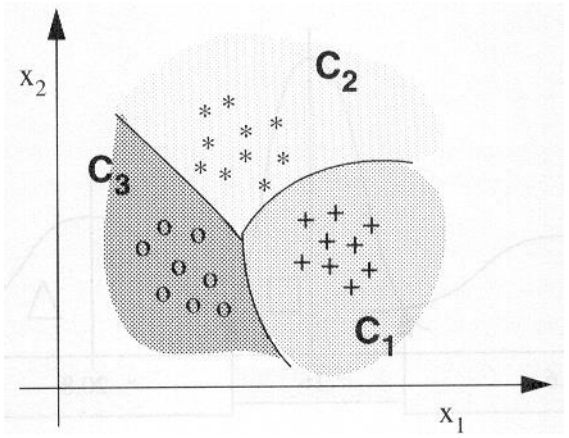
RECONOCIMIENTO DE OBJETOS

- INTRODUCCIÓN

  - Clasificación: enfoque basado en la teoría de la Decisión
    - Introducción a problemas de clasificación
    - Enfoque basado en la teoría de la Decisión

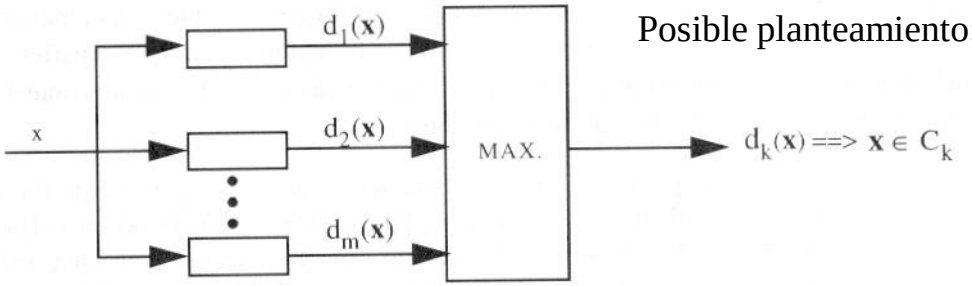
PROBLEMA DE CLASIFICACIÓN: ENFOQUE BASADO EN LA TEORÍA DE LA DECISIÓN

- **Problema de clasificación:**
  - Planteamiento matemático bien definido: Teoría de la Decisión, enfoque probabilístico-estadístico, enfoque basado en la optimización de funciones discriminantes:
    - → División del espacio de características en regiones o subespacios representativos de cada clase considerada en el problema.
    - → Implica la definición de funciones de decisión o discriminantes entre las clases del problema.



Partición del espacio de características x_1-x_2 en 3 regiones correspondientes a 3 clases

➢ **La clasificación se formula en base a unas *funciones* denominadas *de decisión* o *discriminantes* que son evaluadas para decidir la clase de una muestra «desconocida» descrita mediante su vector de atributos.**



Posible planteamiento:

- Se diseña una función de decisión para cada clase del problema.
- Estas funciones de decisión se evalúan para una muestra descrita por un vector de atributos x.
- La muestra se asigna a la clase C_k cuya función de decisión sea mayor.

RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

- Análisis discriminante
  - ✓ Clasificación basada en el Teorema de Bayes
  - ✓ Clasificación basada en distribución normal multivariante
  - ✓ Clasificador QDA: análisis discriminante cuadrático
  - ✓ Clasificador LDA: análisis discriminante lineal
  - ✓ Casos particulares LDA: clasificadores mínima distancia

- **TEOREMA DE BAYES APLICADO A UN PROBLEMA DE CLASIFICACIÓN CON K CLASES DE SALIDA:**

$$p_k(x) = Pr(Y=k \mid X=x) = \frac{Pr(X=x \mid Y=k)\, Pr(Y=k)}{\sum_{i=1}^{K}\left(Pr(X=x \mid Y=i)\, Pr(Y=i)\right)} = \frac{f_k(x)\ \pi_k}{\sum_{i=1}^{K}(f_i(x)\ \pi_i)}$$

⇒ **Probabilidad a posteriori (condicional) de pertenencia de una observación descrita por x a la clase k :**

$$p_k(x) = Pr\ (Y = k \mid X = x)$$

(Probabilidad a posteriori o probabilidad condicional: es una probabilidad que se estima después de que la evidencia sea tenida en cuenta. En nuestro caso, la evidencia son los valores de los predictores que describen a la observación que queremos clasificar $X = x$) .

⇒ **Probabilidad a posteriori de la observación $X = x$ en las muestras de la clase k:**

$$f_k(x) = Pr\ (X = x \mid Y = k)$$

(Término de verosimilitud: probabilidad de tener una descripción $X = x$ en las muestras disponibles de la clase k. Esta probabilidad será alta si hay una probabilidad alta de encontrar observaciones de la clase k con $X \approx x$).

⇒ **Probabilidad a priori (incondicinal) de pertenencia de una observación a la clase k, independientemente de su descripción X:**

$$Pr(Y=k) = \pi_k$$

(Probabilidad de pertenencia a la clase de una observación cualquiera, probabilidad estimada sin tener en cuenta los valores de sus predictores X)

- **Teorema de Bayes aplicado a un problema de clasificación con K clases de salida:**

$$p_k(x) = Pr(Y=k \mid X=x) = \frac{Pr(X=x \mid Y=k)\, Pr(Y=k)}{\sum_{i=1}^{K}\left(Pr(X=x \mid Y=i)\, Pr(Y=i)\right)} = \frac{f_k(x)\ \pi_k}{\sum_{i=1}^{K}(f_i(x)\ \pi_i)}$$

CLASIFICADOR DECISOR MÁXIMO A POSTERIORI:

$$\text{Funciones de decisión MAP: } d_k(x) = Pr(X=x \mid Y=k)\, Pr(Y=k) = f_k(x)\ \pi_k$$

CLASIFICADOR MÁXIMA VEROSIMILITUD:

$$\text{Funciones de decisión ML: } d_k(x) = Pr(X=x \mid Y=k) = f_k(x)$$

RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

- Análisis discriminante
  - ✓ Clasificación basada en el Teorema de Bayes
  - ✓ Clasificación basada en distribución normal multivariante
  - ✓ Clasificador QDA: análisis discriminante cuadrático
  - ✓ Clasificador LDA: análisis discriminante lineal
  - ✓ Casos particulares LDA: clasificadores mínima distancia

- **Teorema de Bayes aplicado para clasificar una muestra descrita por** $X = x = (X_1, \dots, X_p)$

$$p_k(x) = Pr(Y=k \mid X=x) = \frac{Pr(X=x \mid Y=k)\, Pr(Y=k)}{\sum_{i=1}^{K}\left((X=x \mid Y=i)\, Pr(Y=i)\right)} = \frac{f_k(x)\ \pi_k}{\sum_{i=1}^{K}(f_i(x)\ \pi_i)}$$

CLASIFICADOR DECISOR MÁXIMO A POSTERIORI: $d_k(x) = Pr(X=x \mid Y=k)\, Pr(Y=k) = f_k(x)\ \pi_k$

ESTIMACIÓN DE LA VEROSIMILITUD O PROBABILIDAD A POSTERIORI DE x EN LA CLASE k, $f_k(x)$:

→ **CLASIFICADOR NAIVE BAYES:** asume independencia estadística entre los atributos:

$$x = x_0 = (X_{01}, \dots, X_{0p})$$

$$f_k(x_0) = Pr(X=x_0 \mid Y=k) = [Pr(X_1 = X_{01} \mid Y=k)] * [Pr(X_2 = X_{02} \mid Y=k)] * \cdots * [Pr(X_p = X_{0p} \mid Y=k)] = \prod_{i=1}^{p} Pr(X_i = X_{0i} \mid Y=k)$$

→ **Atributos numéricos de naturaleza cuantitativa:** estimar $Pr(X_i = X_{0i} \mid Y=k)$ asumiendo que los datos de X_i en la clase k provienen de una determinada función densidad de probabilidad.

Si distribución normal o gausiana – Clasificador Naive Bayes Gaussiano:

$$\text{Para la clase } k\text{-ésima: } X_i \sim N(\mu_{ik}, \sigma_{ik}) \Rightarrow Pr(X_i = X_{0i} \mid Y=k) \sim \frac{1}{\sqrt{2\pi}\ \sigma_{ik}} exp\left(-\frac{(X_{0i} - \mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

- **Teorema de Bayes aplicado para clasificar una muestra descrita por** $X = x = (X_1, \dots, X_p)$

$$p_k(x) = Pr(Y=k \mid X=x) = \frac{Pr(X=x \mid Y=k)\, Pr(Y=k)}{\sum_{i=1}^{K}\left((X=x \mid Y=i)\, Pr(Y=i)\right)} = \frac{f_k(x)\ \pi_k}{\sum_{i=1}^{K}(f_i(x)\ \pi_i)}$$

CLASIFICADOR DECISOR MÁXIMO A POSTERIORI: $d_k(x) = Pr(X=x \mid Y=k)\, Pr(Y=k) = f_k(x)\ \pi_k$

ESTIMACIÓN DE LA VEROSIMILITUD O PROBABILIDAD A POSTERIORI DE x EN LA CLASE k, $f_k(x)$:

→ **ANÁLISIS DISCRIMINANTE: Clasificación basada en distribución normal multivariante**

- Vector de predictores continuos: $X = (X_1, X_2, \dots, X_p)$
- Asume que las observaciones provienen de una ***distribución gaussiana multivariante*** (*normal multivariante*) con vector de medias y matriz de covarianzas específicos para cada clase, μ_k y Σ_k, respectivamente.

Para la clase *k-ésima*: $X \sim N(\mu_k, \Sigma_k) \rightarrow f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_k)^T(\Sigma_k)^{-1}(x-\mu_k)\right]$

$x = [X_1, X_2, \dots, X_p]^T$; $\mu_k = [\mu_1, \mu_2, \dots, \mu_p]_k^T$ (vectores columna *px1*) ; Σ_k: *matriz de covarianzas* (*pxp*)

RECORDANDO: ***ANÁLISIS DISCRIMINANTE - Clasificación basada en distribución normal multivariante***

Instancia a clasificar: $x = (X_1, \ldots, X_p)$

1. **Teorema de Bayes: Clasificador Decisor Máximo A Posteriori:**

$$d_k(x) = Pr(X=x \mid Y=k)\, Pr(Y=k) = f_k(x)\ \pi_k$$

2. **Estimación de la verosimilitud o probabilidad a posteriori de** x **en la clase** k, $f_k(x)$, *asumiendo que los datos de entrenamiento X siguen* ***una distribución normal multivariante (p-dimensional) para cada clase del problema:***

$$X \sim N(\mu_k, \Sigma_k) \rightarrow f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_k)^T(\Sigma_k)^{-1}(x-\mu_k)\right]$$

3. **A partir del conjunto de datos de entrenamiento, diseño de una función de decisión para cada clase del problema:**

$$d_k(x) = f_k(x)\ \pi_k \Rightarrow d_k(x) = \log[f_k(x)\ \pi_k] = \log[f_k(x)] + \log[\pi_k] \text{ con } f_k(x) = N(\mu_k, \Sigma_k)$$

donde $x = (X_1, \ldots, X_p)$ *representa los valores de los predictores de cualquier muestra a clasificar*

4. **Criterio de clasificación:** evaluar cada función de decisión para la observación a clasificar dada por $X = x$ y asociarla a la clase cuya función de decisión es máxima.

$$Y(x) = i \quad si \quad d_i(x) > d_j(x) \quad \forall\ j \neq i$$

RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

- Análisis discriminante
  - ✓ Clasificación basada en el Teorema de Bayes
  - ✓ Clasificación basada en distribución normal multivariante
  - ✓ Clasificador QDA: análisis discriminante cuadrático
  - ✓ Clasificador LDA: análisis discriminante lineal
  - ✓ Casos particulares LDA: clasificadores mínima distancia

K clases ; Instancia a clasificar: $x = (X_1, \ldots, X_p)$ *; Datos de entrenamiento:* $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$; $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

CLASIFICADOR QDA: ANÁLISIS DISCRIMINANTE CUADRÁTICO

$$d_k(x) = -\frac{1}{2}(x-\mu_k)^T(\Sigma_k)^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log[\pi_k]$$

$$Y(x) = i \quad si \quad d_i(x) > d_j(x) \quad \forall \; j \neq i$$

→ Divide el espacio de características en tantas regiones como clases tenga el problema mediante ***fronteras de decisión cuadráticas.***

→ **Si clases equiprobables** $(\pi_1 = \pi_2 = \cdots = \pi_K)$

$$d_k(x) = -(x-\mu_k)^T(\Sigma_k)^{-1}(x-\mu_k) - \log|\Sigma_k|$$

CLASIFICADOR QDA: ANÁLISIS DISCRIMINANTE CUADRÁTICO

$$d_k(x) = -\frac{1}{2}(x-\mu_k)^T(\Sigma_k)^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log[\pi_k]$$

Para diseñar las funciones de decisión de cada clase, se requiere calcular a partir del conjunto de entrenamiento:

1. **Vector de medias μ_k y matriz de covarianzas Σ_k de la clase k** (calculados sobre las n_k observaciones de la clase k disponibles en el número total n de observaciones de entrenamiento):

$$x = [X_1, X_2, \dots, X_p]^T \;;\; \mu_k = [\mu_1^k, \mu_2^k, \dots, \mu_p^k]^T \;\; con\; \mu_i^k = E[X_i^k] = \frac{1}{n_k}\sum_{z:\, y_z = k} X_{zi}$$

$$\Sigma_k = \begin{bmatrix} \sigma_{11}^k & \sigma_{12}^k & \cdots & \sigma_{1p}^k \\ \sigma_{21}^k & \sigma_{22}^k & \cdots & \sigma_{2p}^k \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1}^k & \sigma_{p2}^k & \cdots & \sigma_{pp}^k \end{bmatrix} \; con \;\; \sigma_{ij}^k = \sigma_{ji}^k = E[(X_i^k - \mu_i^k)(X_j^k - \mu_j^k)] = \frac{1}{n_k - 1}\sum_{z:\, y_z = k}(X_{zi} - \mu_i^k)(X_{zj} - \mu_j^k)$$

2. **Probabilidad de priori que tiene una muestra de pertenecer a la clase de la clase k:**

   - Si no se tiene conocimiento a priori que pueda ser utilizado para estimar esta probabilidad, se suele calcular a partir de la proporción de las muestras de entrenamiento de pertenecer a la clase en cuestión:

$$\pi_k = \frac{n_k}{n}$$

RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

- Análisis discriminante
  - ✓ Clasificación basada en el Teorema de Bayes
  - ✓ Clasificación basada en distribución normal multivariante
  - ✓ Clasificador QDA: análisis discriminante cuadrático
  - ✓ Clasificador LDA: análisis discriminante lineal
  - ✓ Casos particulares LDA: clasificadores mínima distancia

K clases ; Instancia a clasificar: $x = (X_1, \ \dots \ , X_p)$ *; Datos de entrenamiento:* $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$; $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

CLASIFICADOR LDA – ANÁLISIS DISCRIMINANTE LINEAL

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma$$

$$x = [X_1, X_2, \dots, X_p]^T$$

$$d_k(x) = -\frac{1}{2}(x - \mu_k)^T (\Sigma)^{-1}(x - \mu_k) + \log[\pi_k]$$

Si conjunto de entrenamiento balanceado en las clases (clases equiprobables) →

$$d_k(x) = -(x - \mu_k)^T (\Sigma)^{-1}(x - \mu_k)$$

LDA requiere calcular a partir del conjunto de datos entrenamiento:

$$\Rightarrow \ \mu_k = [\mu_1^k, \mu_2^k, \dots, \mu_p^k]^T \ con \ \mu_j^k = E[X_j^k] = \frac{1}{n_k} \sum_{i:\, y_i = k} X_{ij}$$

$$\Rightarrow \ \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \ con \ \ \sigma_{ij} = \sigma_{ji} = \frac{1}{n - K} \sum_{k=1}^{K} \ \sum_{z:\, y_z = k} (X_{zi} - \mu_i^k)(X_{zj} - \mu_j^k)$$

$$\Rightarrow \ \pi_k = \frac{n_k}{n}$$

Notar que Σ es una estimación de una matriz de covarianzas común a todas las clases calculada a partir de la matriz de covarianzas de cada clase, Σ_k*:*

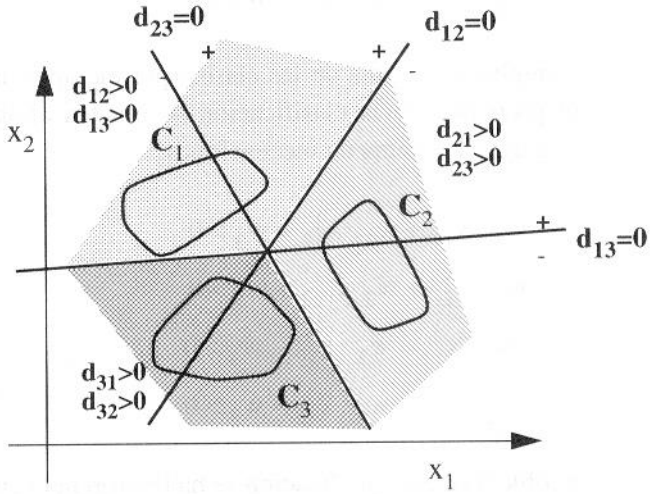
$$\Sigma = \frac{1}{n - K} \sum_{k=1}^{K} (n_k - 1)\Sigma_k$$

DOS FORMAS DE APLICACIÓN LDA: *K clases ; Instancia a clasificar:* $x=(X_1, \ldots, X_p)$

1. **A partir de las funciones de decisión cuadráticas:** diseñar K funciones de decisión (una función de decisión d_k para cada clase k del problema) de forma que una observación dada por X = x se asocia a la clase cuya función de decisión es :

$$d_k(x) = -\frac{1}{2}(x-\mu_k)^T(\Sigma)^{-1}(x-\mu_k) + \log[\pi_k] \longrightarrow Y(x) = i \;\; si \;\; d_i(x) > d_j(x) \;\; \forall \; j \neq i$$

2. **A partir de las fronteras lineales de decisión:** a partir de las funciones de decisión anteriores d_k, determinar $\binom{K}{2} = K(K-1)/2$ fronteras de decisión para separar las muestras de las clases dos a dos de la siguiente forma:



$$d_{ij}(x) = d_i(x) - d_j(x) = \beta_{ij_0} + \beta_{ij_1}X_1 + \beta_{ij_2}X_2 + \cdots + +\beta_{ij_p}X_p$$

$$Y(x) = i \;\; si \;\; d_{ij}(x) > 0 \;\; \forall \; j \neq i$$

Ejemplo – Criterio de clasificación para el caso de las 3 clases de la figura:

$$X \in C_1 \; si \; d_{12} > 0 \; y \; d_{13} > 0 \; ; \; X \in C_2 \; si \; d_{12} < 0 \; y \; d_{23} > 0 \; ; \; X \in C_3 \; si \; d_{13} < 0 \; y \; d_{23} < 0$$

FRONTERAS DE SEPARACIÓN ENTRE CLASES LINEALES :

- Frontera de decisión de las clases i-j: HIPERPLANO DADO POR $d_{ij}(x) = 0$

  (notar que son los puntos del espacio de predictores que están en la frontera de separación de ambas clases y cumplen que $d_i(x) = d_j(x)$ *)*

RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

- Análisis discriminante
  - ✓ Clasificación basada en el Teorema de Bayes
  - ✓ Clasificación basada en distribución normal multivariante
  - ✓ Clasificador QDA: análisis discriminante cuadrático
  - ✓ Clasificador LDA: análisis discriminante lineal
  - ✓ Casos particulares LDA: clasificadores mínima distancia

CLASIFICADOR LDA – ANÁLISIS DISCRIMINANTE LINEAL

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma \quad \rightarrow \quad d_k(x) = -\frac{1}{2}(x-\mu_k)^T(\Sigma)^{-1}(x-\mu_k) + \log[\pi_k]$$

$$x = [X_1, X_2, \ldots, X_p]^T \;;\; \mu_k = [\mu_1^k, \mu_2^k, \ldots, \mu_p^k]^T \;;\; \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)\Sigma_k$$

→ CLASIFICADOR MÍNIMA DISTANCIA DE MAHALANOBIS

- Asume clases equiprobables (conjunto de entrenamiento balanceado en las clases): $\pi_1 = \pi_2 = \cdots = \pi_K$

$$d_k(x) = -(x-\mu_k)^T(\Sigma)^{-1}(x-\mu_k) = -D_M{}^2(x, \mu_k)$$

- Este clasificador asigna una observación descrita por x a la clase cuyo vector promedio μ_i esté a distancia de Mahalanobis mínima.
- El criterio es, por tanto, medir la distancia a cada prototipo de las clases y clasificar la instancia cuyo vector prototipo esté más cerca según la Distancia de Mahalanobis (clasificador mediante el prototipo más próximo).

CLASIFICADOR LDA – ANÁLISIS DISCRIMINANTE LINEAL

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma \quad \rightarrow \quad d_k(x) = -\frac{1}{2}(x-\mu_k)^T(\Sigma)^{-1}(x-\mu_k) + \log[\pi_k]$$

$$x = [X_1, X_2, \ldots, X_p]^T \;;\; \mu_k = [\mu_1^k, \mu_2^k, \ldots, \mu_p^k]^T \;;\; \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \frac{1}{n-K}\sum_{k=1}^{K}(n_k-1)\Sigma_k$$

→ **CLASIFICADOR MÍNIMA DISTANCIA EUCLIDEA.** Asunciones:

- Clases equiprobables (conjunto de entrenamiento balanceado en las claes): $\pi_1 = \pi_2 = \cdots = \pi_K$
- Las variables de los predictores X_i son estadísticamente independientes, no están correladas: $\sigma_{ij} = 0 \quad \forall\, i \neq j$
- Las varianzas de cada variable predictora X_i son iguales: $\sigma_{ii} = {\sigma_i}^2 = \sigma^2 \quad \forall\, i = 1, 2, \ldots, p$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{22} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I$$

$$d_k(x) = -(x-\mu_k)^T(\Sigma)^{-1}(x-\mu_k) = -\frac{1}{\sigma^2}(x-\mu_k)^T(x-\mu_k) \quad \rightarrow \quad d_k(x) = -(x-\mu_k)^T(x-\mu_k) = -{D_E}^2(x, \mu_k)$$

- Este clasificador asigna una observación descrita por x a la clase cuyo vector promedio μ_i esté a distancia Euclidea mínima (clasificador mediante el prototipo más próximo según distancia Euclidea).

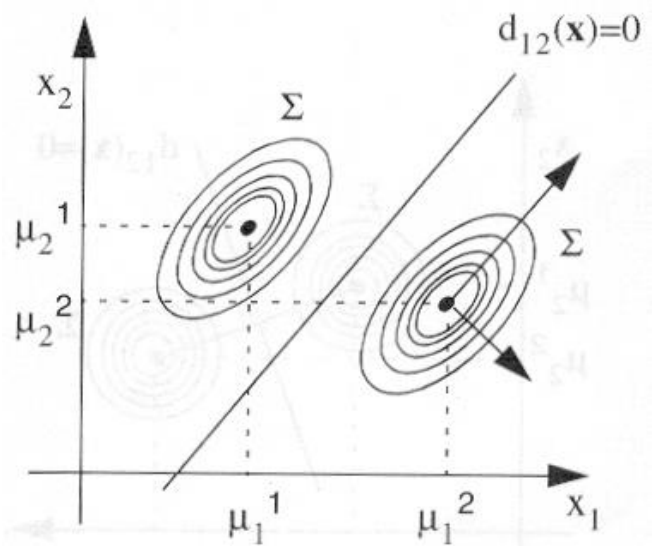
CASOS PARTICULARES LDA: Clasificadores Mínima Distancia Mahalanobis, Mínima Distancia Euclidea

Clasificador de Mínima Distancia de Mahalanobis:

⇒ El lugar geométrico de aquellos puntos equidistantes del centro (vector de medias) son hiperelipsoides.

⇒ Caso de aplicación:

- Clases equiprobables.
- Matrices de covarianzas de las clases similares.
- Las variables presentan cierta dependencia lineal o correlación entre ellas y/o tienen distinta varianza.

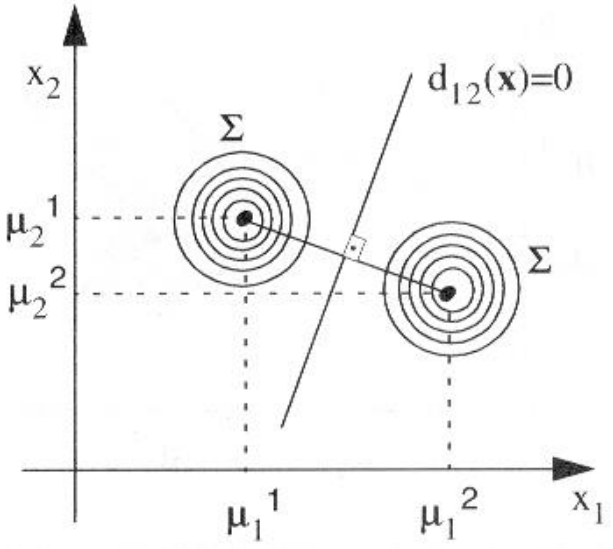


Clasificador de Mínima Distancia Euclídea

⇒ El lugar geométrico de aquellos puntos equidistantes del centro (vector de medias) son hiperesferas.

⇒ Casos de aplicación:

- Clases equiprobables.
- Matrices de covarianzas de las clases similares.
- Las variables son independientes (no están correladas) y presentan una varianza similar.



RECONOCIMIENTO DE OBJETOS

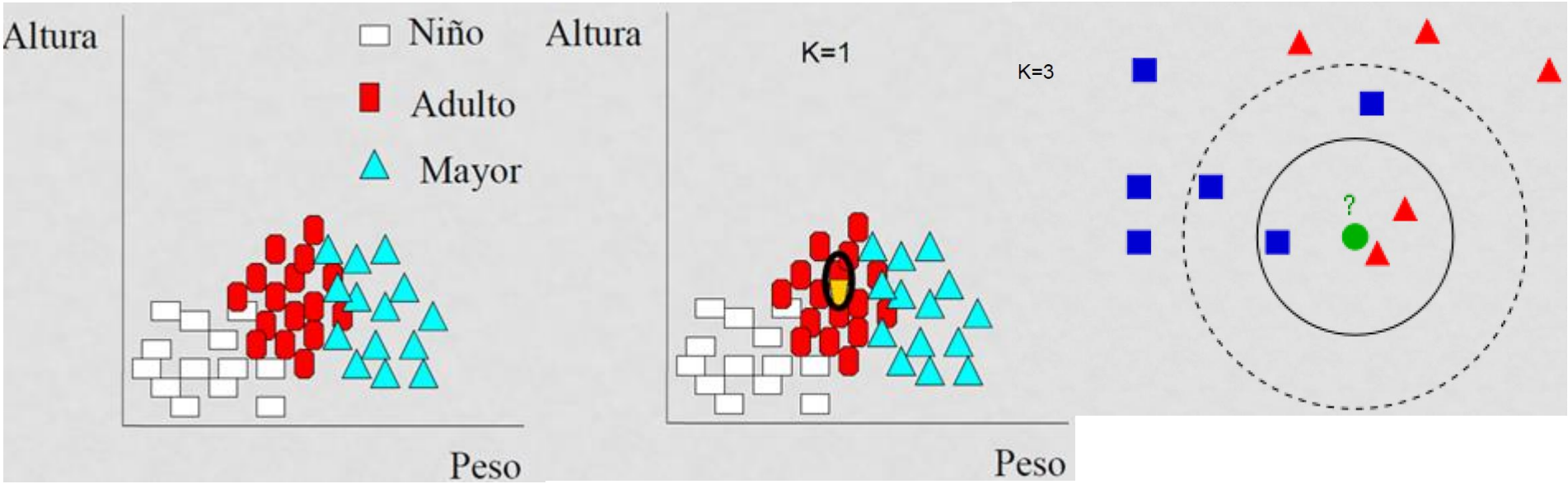
- TECNICAS BÁSICAS DE CLASIFICACIÓN

- K-vecinos más cercanos
  - Clasificador K-NN
  - Selección de instancias
  - Medidas de similitud/disimilitud
  - Regresión K-NN

❑ CLASIFICADOR K-NN. Planteamiento

1. Dado un conjunto de datos de entrenamiento (observaciones o muestras descritas por los correspondientes valores de los predictores, de clase conocida) y dada una muestra de test cuya clase es desconocida, se buscan las **"K" muestras de entrenamiento más parecidas a la de test**.

2. La clase predicha para la instancia de test es la clase más numerosa de las clases a las que pertenecen las "K" muestras de entrenamiento más cercanas.



CARACTERÍSTICAS GENERALES

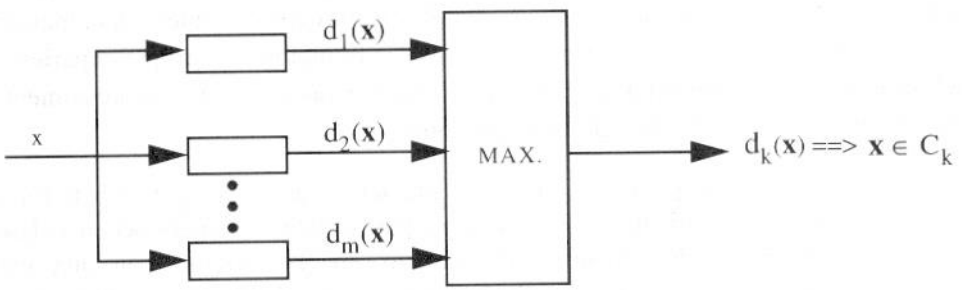
- **Algoritmo "perezoso" (*lazy*):**
  - K-NN no genera un modelo fruto del aprendizaje con datos de entrenamiento, sino ***que el aprendizaje sucede en el mismo momento en el que se prueban los datos de test.***
    - → Durante el entrenamiento, sólo "guardan" las instancias, no se construye ningún modelo.
    - → La clasificación se hace cuando llega la instancia de test.
- **Es no paramétrico:** no se hacen suposiciones sobre la distribución que siguen los datos (como, por ejemplo, hacen clasificadores basados en distribuciones normales); asume que el mejor modelo de los datos son los propios datos.
- **Es local:** la clase de un dato depende sólo de los *k* vecinos mas cercanos (no se construye un modelo global).
- **Hiperparámetro del modelo:** valor de K

Observación:

→ **Parámetros de un modelo:** son las variables que se estiman durante el proceso de entrenamiento con los conjuntos de datos (coeficientes en una regresión lineal, pesos en una red neuronal, vectores de soporte en una máquina de vector soporte).

→ **Hiperparámetros de un modelo:** son parámetros que se configuran antes del entrenamiento del modelo y no forman parte del modelo como tal; generalmente, sus valores óptimos no se conocen a priori, para establecerlos se deben utilizar reglas genéricas, valores que han funcionado en problemas similares o ajustarlos mediante prueba y error (mediante validación cruzada o, si es demasiado costoso en tiempo, utilizando un único conjunto de validación).

➢ PLANTEAMIENTO PROBLEMA DE CLASIFICACIÓN BASADO EN TEORÍA DE DECISIÓN:



- Se diseña una función de decisión para cada clase del problema.
- Estas funciones de decisión se evalúan para una muestra descrita por un vector de atributos x.
- La muestra se asigna a la clase C_k cuya función de decisión sea mayor.

➢ CLASIFICADORES BASADOS EN PROBABILIDAD :

- Asigna una observación dada por `x` a la clase más probable.
- Función de decisión asociada a la clase `j` (la variable de respuesta `Y` tiene el valor `j`):

  $d_j(x) = P(Y=j \mid X=x)$ – probabilidad que una muestra descrita por `x` (`X=x`) sea de la clase `j` (`Y=j`)

➢ CLASIFICADOR K-NN:

Dado un entero positivo `K` y una observación de test $X=x_0$:

1. El clasificador calcula N_0: conjunto de `K` muestras del conjunto de entrenamiento que están más cerca de x_0 (cercanía medida en términos de distancia entre los puntos que representan las muestras en el espacio de características).
2. El clasificador estima la probabilidad condicional de una clase como la fracción de puntos de N_0 que son de la clase en cuestión. ⇒ Probabilidad que una muestra dada por x_0 sea de la clase `j`:

$$d_j(x_0) = P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i, j) \qquad \text{con} \quad I(m,n) = \begin{cases} 1 & si\ m = n \\ 0 & si\ m \neq n \end{cases}$$

❑ Ejemplo de estimación de probabilidades: Clasificador K-vecinos más próximos (K-NN, *K-Nearest Neighbors*)

❖ Clasificador K-NN:

→ También calcula la probabilidad condicional para cada clase del problema dada una determinada observación y clasifica dicha observación a la clase con mayor probabilidad estimada.

→ Dado un entero positivo `K` y una observación de test $X=x_0$:

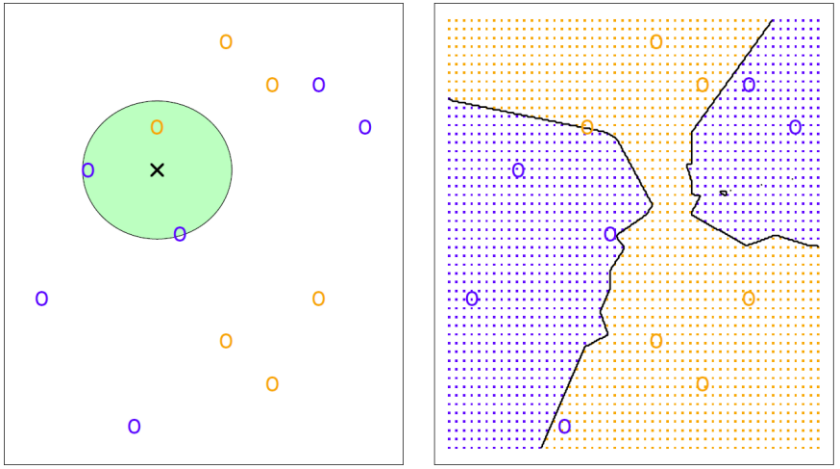
1. El clasificador calcula N_0: conjunto de `K` muestras del conjunto de entrenamiento que están más cerca de x_0 (cercanía medida en términos de distancia entre los puntos que representan las muestras en el espacio de características).
2. El clasificador estima la probabilidad condicional de una clase como la fracción de puntos de N_0 que son de la clase en cuestión
⇒ Probabilidad que una muestra dada por x_0 sea de la clase `j`:

$$P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j)$$

con $I(m,n) = \begin{cases} 1 & si\ m = n \\ 0 & si\ m \neq n \end{cases}$

Ejemplo: clasificación binaria utilizando un K-NN con K=3. Se dispone de un conjunto de entrenamiento formado por 6 observaciones para cada clase (círculos naranjas y azules).

- Izquierda: clasificación de una muestra de test (cruz negra). Se identifican las tres muestras del conjunto de entrenamiento más cercanas a la de test. Se clasifica esta muestra como de la clase azul (clase más numerosa).
- Derecha: frontera de decisión KNN y partición del espacio de características según un clasificador 3-NN.



❑ Ejemplo para Clasificación basada en Teorema de Bayes:

- ❖ Implica el diseño de una función de decisión para cada clase del problema de acuerdo a su probabilidad condicionada: $Pr(Y=j|\ X=x)$ (probabilidad que una instancia dada por x ($X=x$) sea de la clase j ($Y=j$)

- ❖ **Ejemplo:** supongamos que tenemos 200 observaciones descritas por dos atributos X_1 y X_2 pertenecientes a dos categorías (Clasificación binaria: $Y = \{1, 2\}$, clases 1 y 2, 100 observaciones por clase)

  - ➢ **Diseño del clasificador:** con las 100 observaciones disponibles de cada clase, diseñamos una función de decisión para cada del problema:

$$d_1 = Pr(Y=1|\ X=x)\ ;\quad d_2 = Pr(Y=2|\ X=x)$$

Observación: este es un ejemplo de datos simulados, se conoce la función distribución de probabilidad con la que han sido generados los datos de una determinada clase (esto es, en este caso, conocemos las funciones reales d_1 y d_2 (en la práctica, se deben estimar para clasificar una observación a la clase con mayor probabilidad estimada)

- ➢ **Aplicación del clasificador.** Criterio de clasificación:

  → Una observación dada por $X=x_0$ se asocia a la clase 1 si $d_1 > d_2$, esto es, si

$$Pr(Y=1|\ X=x_0) > Pr(Y=2|\ X=x_0)$$. En caso contrario, la muestra se asigna a la clase 2.

  → La aplicación de este criterio provoca una partición del espacio de características bidimensional, dado por X_1 y X_2 en las dos clases del problema.

  → La frontera de separación entre las dos clases en este espacio de características vendrá dada por los puntos $x = (X_1\ ,\ X_2)$ para los que

$$Pr(Y=1|\ X=x) = Pr(Y=2|\ X=x_0)$$

❑ **Ejemplo para Clasificador de Bayes:** conjunto de datos simulados de 200 observaciones de dos clases (100 de cada clase)

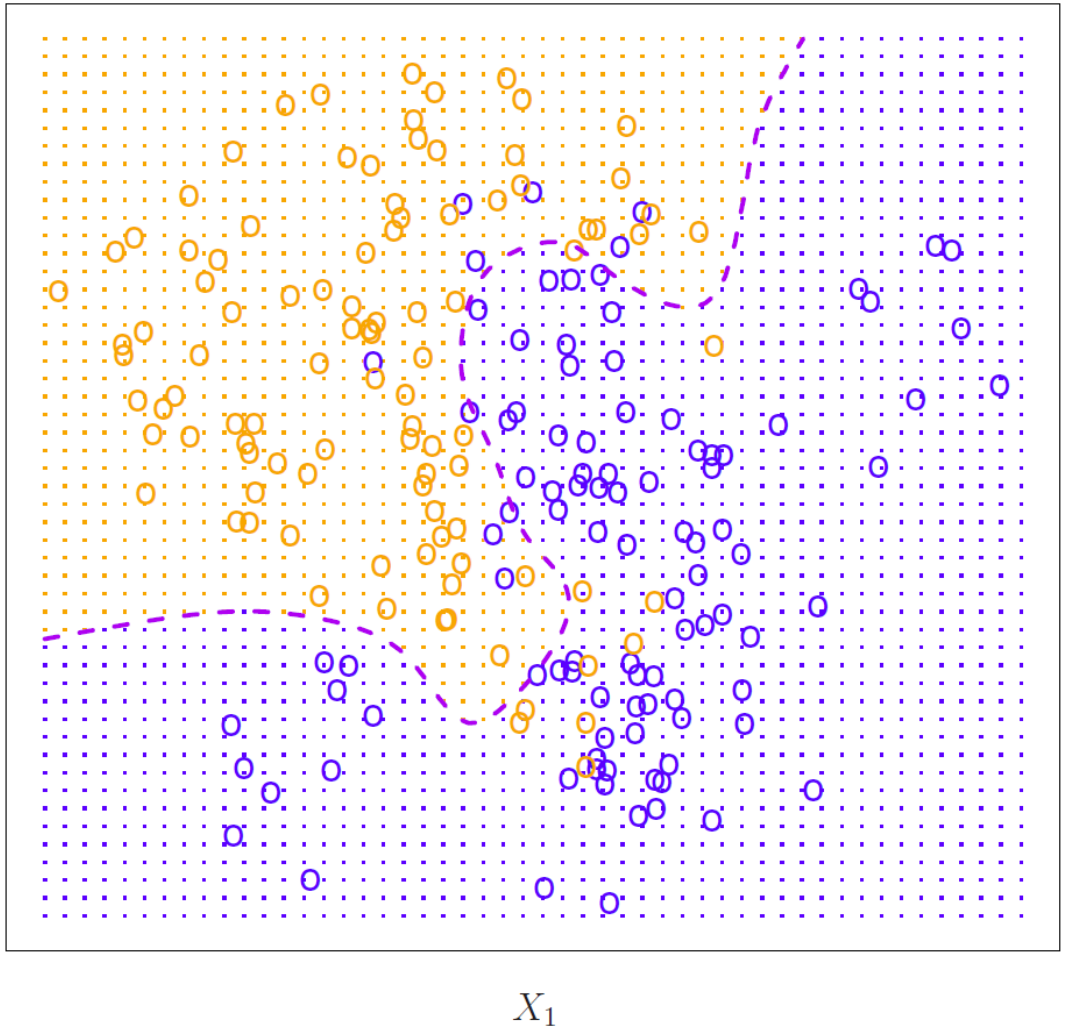
→ La figura muestra la representación del conjunto de datos disponible (círculos). En color naranja y azul, se indican la clase a la que pertenece cada observación.

→ La línea punteada púrpura representa la frontera de decisión de Bayes.

→ Los puntos de fondo naranja indican la región en la que se asignará una observación de prueba a la clase naranja.

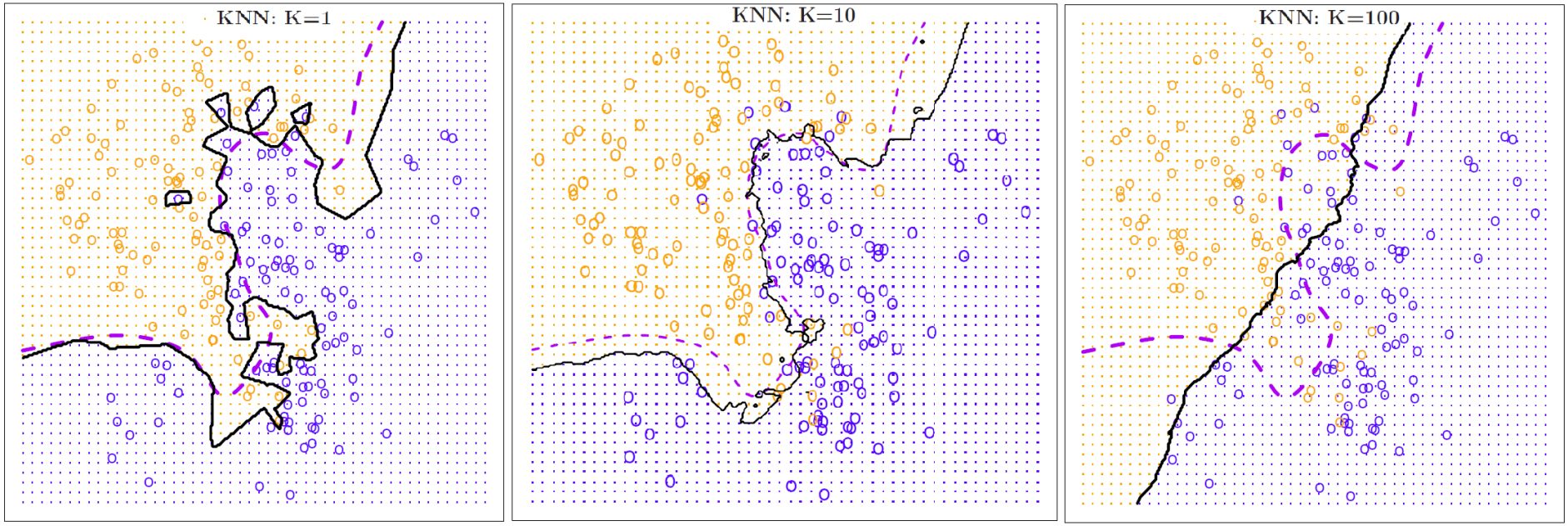
→ Loa puntos de fondo azul indican la región en la que se asignará una observación de prueba a la clase azul.

→ En estos datos simulados, la tasa de error se sitúa en 0.1304.



→ Aunque sean datos simulados y el clasificador de Bayes utilice como funciones de decisión de cada clase las funciones reales de distribución de probabilidad que generan los datos de cada clase (clasificador ideal), el error es mayor que cero porque las muestras presentan cierto solapamiento entre clases.

❑ **Ejemplo para Clasificador K-NN:** conjunto de datos simulados de 200 observaciones de dos clases



→ Curva negra continua : frontera de decisión del KNN → Curva discontinua: frontera de decisión de Bayes

❖ *A medida que K aumenta, ¿¿¿ el método se hace más o menos flexible ???*

❖ *Analiza cómo varía el comportamiento del clasificador en función de K en términos de capacidad de ajuste vs capacidad de generalización*

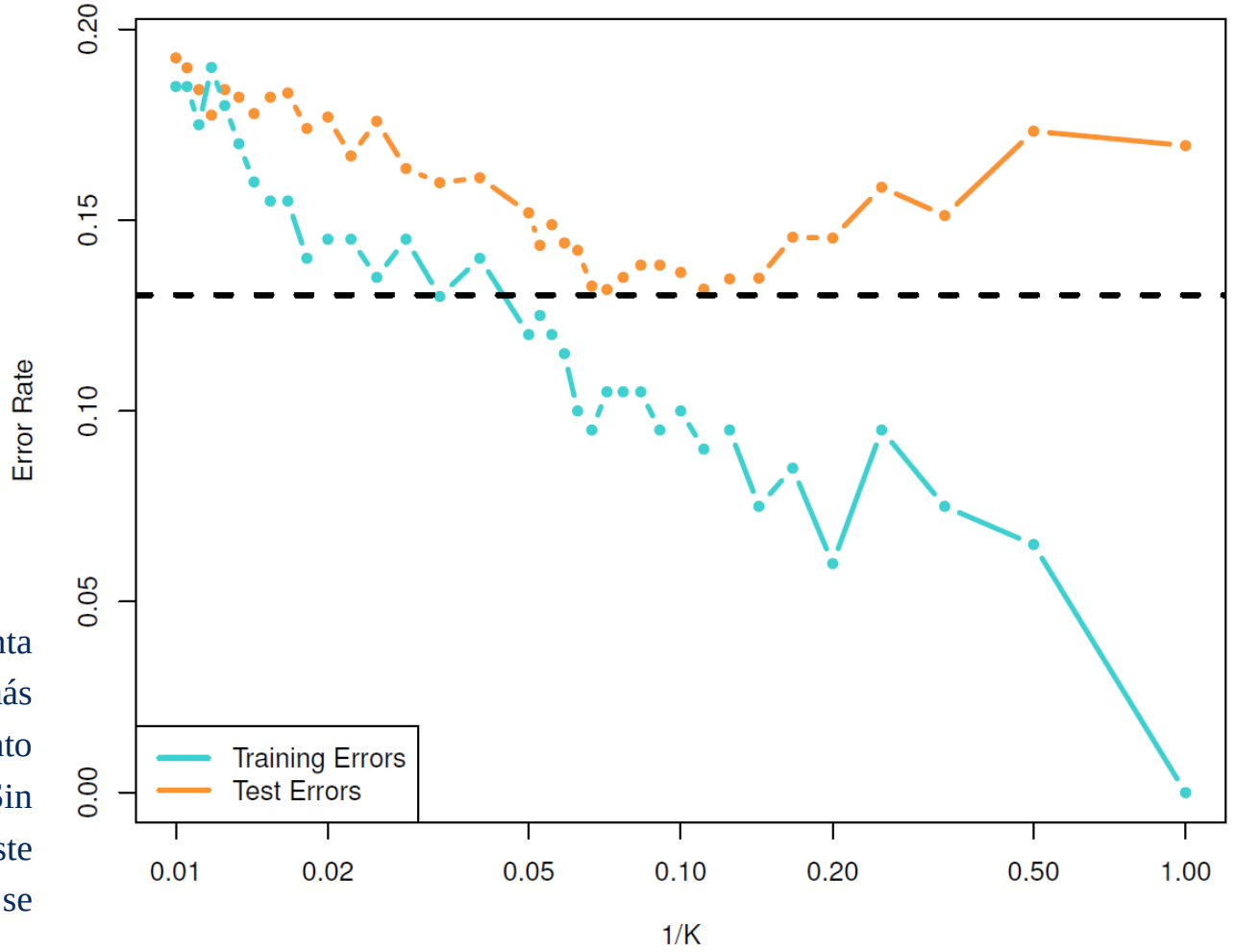
- Con K = 1, la frontera de decisión es demasiado flexible, todo lo contrario que con K = 100, que genera una frontera de decisión cercana a la lineal.
- Con K = 10, el clasificador de Bayes y el 10-NN generan fronteras de decisión muy parecidas.

❑ **Ejemplo para Clasificador K-NN:** importancia de la elección del valor de K

Evaluación de los clasificadores anteriores sobre el conjunto de entrenamiento de 200 observaciones y sobre un conjunto de test de 5000 observaciones generadas de forma similar. La línea discontinua muestra el error del Clasificador de Bayes.

→ Para K = 100 (1/K = 0.01) : altos errores en ambos conjuntos de entrenamiento y test (contorno de decisión cercano al lineal).

→ A medida que disminuye K (aumenta 1/K), el modelo es cada vez más flexible y error en el entrenamiento baja, siendo 0 para K = 1. Sin embargo el error de test para este valor de K es muy elevado → se produce sobreaprendizaje.



→ Curva de Error en Test con forma de U característica del sobreaprendizaje. El mínimo error en test se produce para K = 10; para valores menores, el error en test tiende a aumentar → los modelos son tan flexibles que sobreajustan.

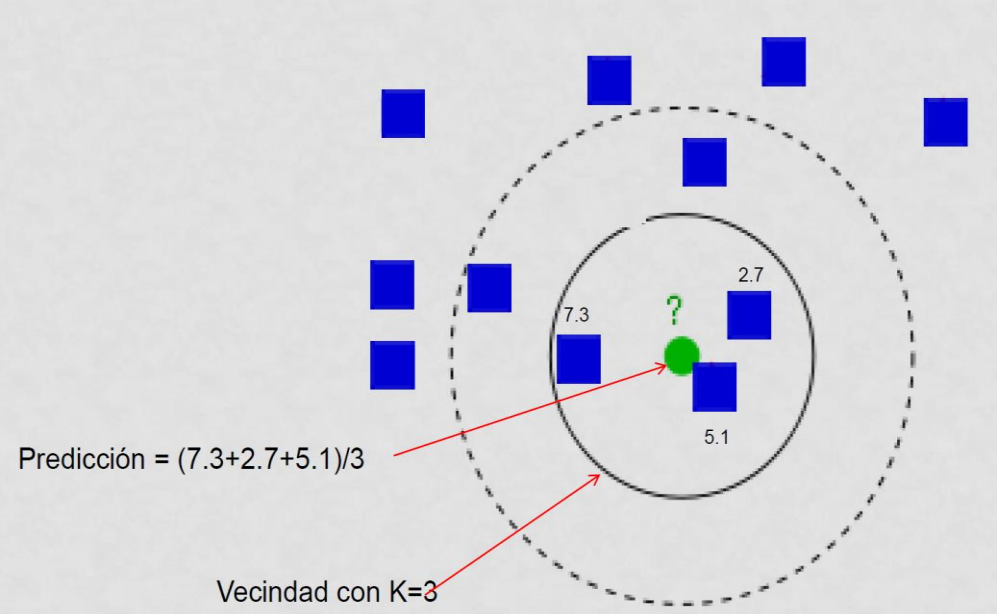
RECONOCIMIENTO DE OBJETOS

- TECNICAS BÁSICAS DE CLASIFICACIÓN

  - K-vecinos más cercanos
    - Clasificador K-NN
    - Selección de instancias
    - Medidas de similitud/disimilitud
    - Regresión K-NN

- **KNN para clasificación:** predice la clase de un instancia como la clase mayoritaria de entre los *k* vecinos mas cercanos de entre los datos de entrenamiento (la respuesta o salida del problema es de naturaleza cualitativa o discreta, la clase se refiere a uno de sus posibles valores).

- **KNN para regresión:** predice la respuesta o salida (de naturaleza numérica cuantitativa o continua) de una instancia como la media de las respuestas de los *k* vecinos mas cercanos de los datos de entrenamiento.



❑ PROBLEMAS DE REGRESIÓN EN APRENDIZAJE AUTOMÁTICO

- ***Predictores, o variables de entrada:*** $X = (X_1, X_2, \ldots, X_p)$
- ***Repuesta o variable de salida:*** Y

Problema de regresión: problemas con una respuesta *Y* cuantitativa..

Dado un conjunto de datos compuesto por n ***observaciones***: $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$

$x_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ e y_i: *valores de las variable de entrada y salida para la observación i*

X_{ij}: *valor de la variable* X_j *para la observación i con* $i = 1,\ldots, n \quad j = 1,\ldots, p$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

y asumiendo existe una función f **que relaciona las variables entrada-salida** $X-Y$**:** $Y = f(X) + \varepsilon$

- ***OBJETIVO:*** *encontrar una estimación de la función f que establece la relación* $X-Y$ *a partir de las observaciones entrada-salida disponibles.*

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}$ representa nuestra estimación sobre f,
- $\hat{Y}$ representa la predicción resultante de Y

REGRESIÓN KNN:

- Método no paramétrico, no se hace ninguna suposición explícita sobre la forma funcional de f, se busca la estimación de f que se acerque lo más posible a los puntos del conjunto de datos de entrenamiento.

Dado un entero positivo K y una observación de test $X=x_0$:

1. La regresión KNN calcula N_0: conjunto de K muestras del conjunto de entrenamiento que están más cerca de x_0.

2. Estima la salida de la observación dada por x_0 mediante el promedio de las salidas de las muestras N_0

$$\hat{y} = \hat{f}(X = x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

   - Para que las instancias más lejanas tengan menos importancia, se puede hacer una media ponderada por 1/d