Jesús García Herrero

CLASIFICADORES BAYESIANOS

En esta clase se presentan los algoritmos Análisis de Datos para abordar tareas de aprendizaje de modelos predictivos.

Se particularizan las técnicas estadísticas vistas anteriormente para resolver tareas predictivas: por un lado la regresión (lineal o no lineal) para predecir valores numéricos, y su aplicación para predecir probabilidades y clasificar instancias mediante regresión logística, y los clasificadores bayesianos como modelo de aprendizaje de parámetros de distribuciones de probabilidad para predecir probabilidades de clases.

La clasificación mediante regresión permite estimar las fronteras de decisión entre clases, presentando la equivalencia del aprendizaje de fronteras de clasificación al de las funciones de estimación de pertenencia a la clase. Se puede ver la limitación de esta técnica a problemas linealmente separables.

Los clasificadores bayesianos parten del principio de probabilidad condicionada para estimar probabilidades a posteriori de pertencia a clases de las instancia, una vez calculadas las probabilidades de los valores de los atributos (probabilidades a priori) en la fase de entrenamiento. Estos clasificadores permiten tratar con datos nominales y numéricos, en este último caso utilizando distribuciones normales para simplificar el proceso. La limitación está en el cálculo de dependencias entre los atributos, que requeriría un número de datos exponencial con la dimensión de éstos, problema habitualmente tratado con la simplificación de independencia condicional (método “naïve Bayes”). Se completa el tema revisando aspectos prácticos que surgen al aplicar técnicas de clasificación sobre datos reales: tratamiento de datos incompletos y datos insuficientes para estimar probabilidades muy pequeñas.

Clasificadores Bayesianos

Métodos probabilísticos y numéricos de clasificación

Jesús García Herrero
Universidad Carlos III de Madrid

Universidad Carlos III de Madrid

Clasificación numérica

- Modelado de datos con atributos numéricos para su aplicación a Clasificación. Generalización
- Datos representados como vectores de atributos numéricos: patrones

A_1	A_2	...	A_F
x^1_1	x^1_2	...	x^1_F
x^2_1	x^2_2	...	x^2_F
x^3_1	x^3_2	...	x^3_F
...	...	...	...
x^N_1	x^N_2	...	x^N_F

$$\rightarrow \left\{\vec{X}^1, \vec{X}^2, ..., \vec{X}^N\right\}$$

- Problemas: **dimensionalidad, sobreajuste.**

Clasificación numérica

Problema de Clasificación

- Clases: $\{C_1, ..., C_M\}$
- Muestras:E= $\{\vec{X}_1^{(1)},..., \vec{X}_{n_1}^{(1)}, \vec{X}_1^{(2)},..., \vec{X}_{n_2}^{(2)},..., \vec{X}_1^{(M)}..., \vec{X}_{n_M}^{(M)}\}$
  - Tamaño: $N = \sum_{j=1}^{M} n_j$
- Para cada clase, C_i, hay n_i patrones, cada uno con F atributos: para cada clase Ci: $\{\vec{X}_1^{(i)},..., \vec{X}_{n_i}^{(i)}\}$

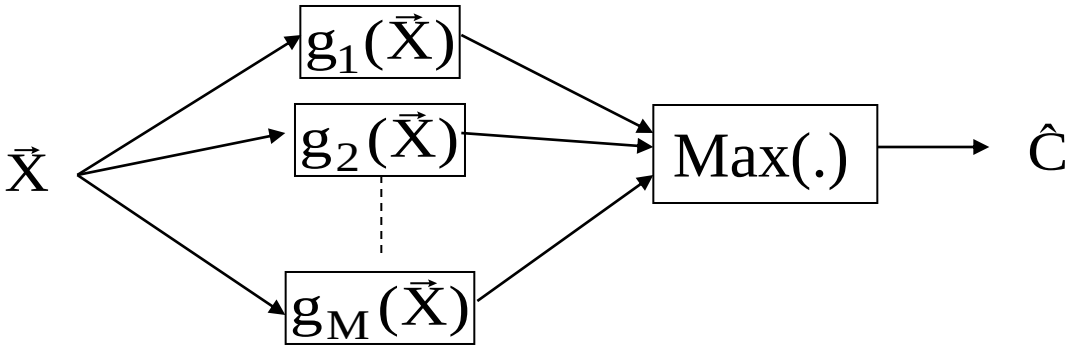
$$\vec{X}_j^{(i)} = \begin{bmatrix} x_{1j}^{(i)} \\ \vdots \\ x_{Fj}^{(i)} \end{bmatrix}; \quad j = 1,..., n_i$$

Clasificación numérica

$$g(.): \quad R^F \longrightarrow C = \{C_1,...,C_M\}$$

$$\vec{X} \longrightarrow \hat{C} = g(\vec{X})$$

- Función discriminante de cada clase: $g_i(\vec{X}), i = 1,...,M$

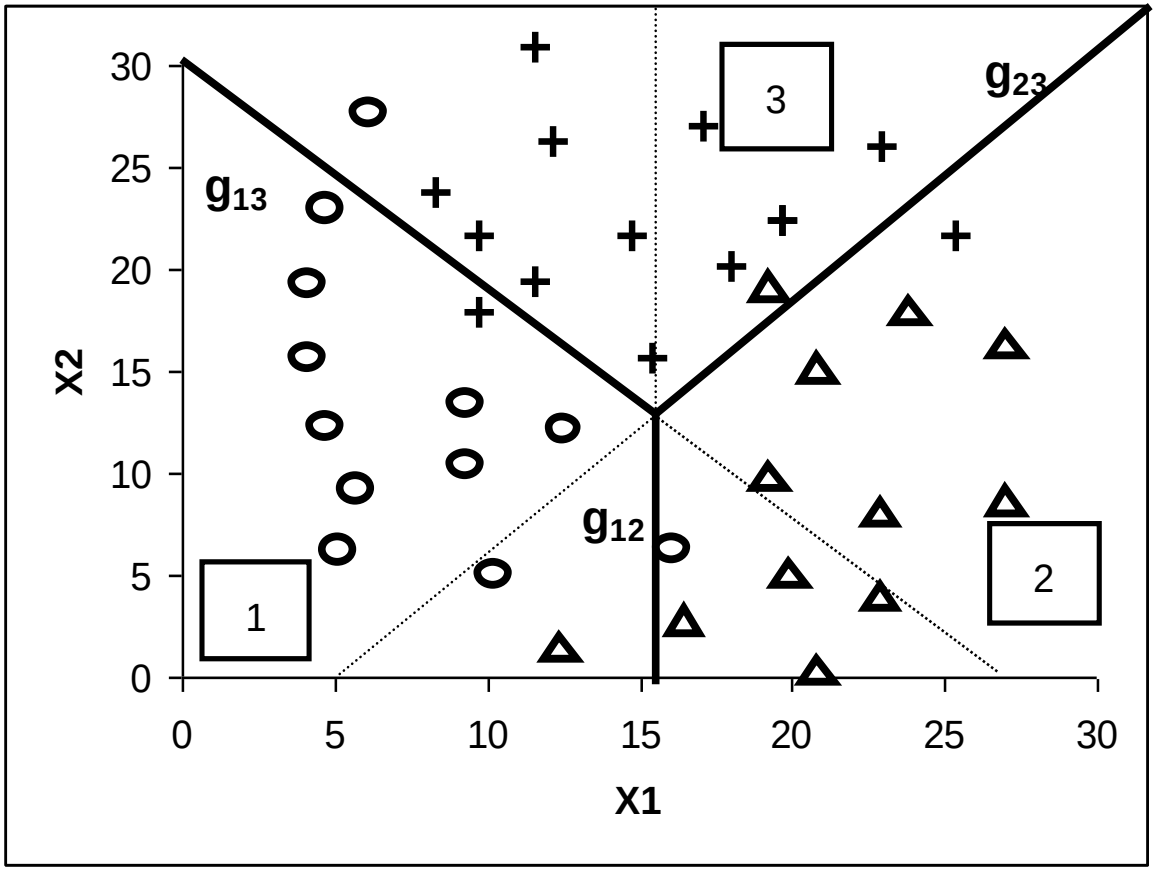


- Propiedad deseable para el diseño de $g_i(.)$: sobre el conjunto de entrenamiento E, cada patrón de la clase Ci tiene un valor máximo con el discriminante gi(.):

$$g_i(\vec{X}_j^{(i)}) = \max_{k=1,...,M} \{g_k(\vec{X}_j^{(i)})\}, \forall j = 1,...,n_i$$

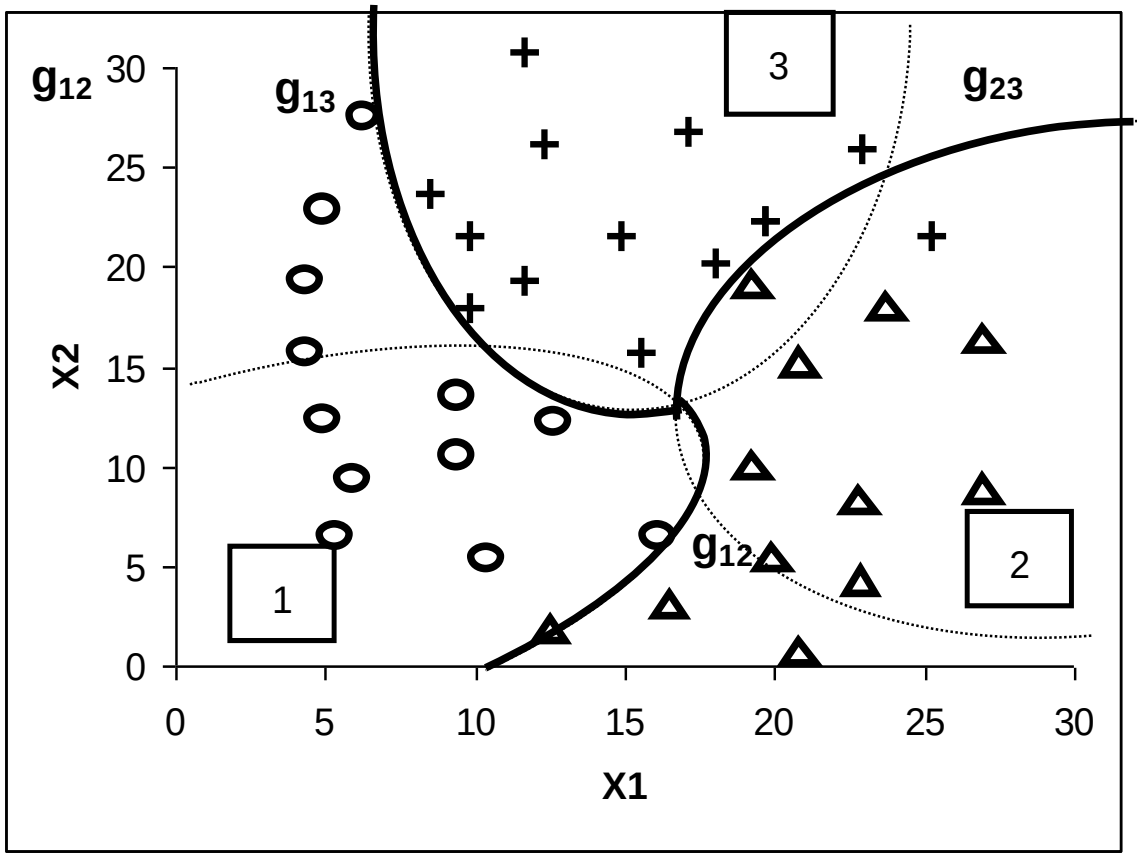
Fronteras de decisión

$$g_{ij}(\vec{X}): \quad \text{lineales}$$



Fronteras de decisión

$$g_{ij}(\vec{X}): \quad \text{cuadraticas}$$



Clasificación con Regresión Lineal

- Para cada clase se define la función de pertenencia g_i:

$$g_i(\vec{X}) = \begin{cases} 1; & \vec{X} \in C_i \\ 0; & \vec{X} \notin C_i \end{cases}$$

- Se construye una función lineal que "aproxime" g_i:

1^s en los patrones de Ci

0^s en resto

$$\vec{y}_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$H_i = \begin{bmatrix} 1 & \left(\vec{X}_1^{(i)}\right)^t \\ \vdots & \vdots \\ 1 & \left(\vec{X}_{n_i}^{(i)}\right)^t \\ 1 & \left(\vec{X}_1^{(1)}\right)^t \\ \vdots & \vdots \\ 1 & \left(\vec{X}_{n_I}^{(I)}\right)^t \end{bmatrix};$$

todos los datos

$$\vec{A}_i = [H_i{}^t H_i]^{-1} H_i{}^t \vec{y}_i$$

- Hay que "aprender" M funciones g_i

Clasificación bayesiana: aplicación de modelos estadísticos

- Clasificación con modelo de estructura probabilística conocida

  Clases: $\{C_1, ..., C_M\}$. Se conoce a priori:

  - Probabilidades de clase: $P(C_i)$
  - Distribuciones de probabilidad condicionadas (parámetros constantes)

$$F_{\vec{X}}(x_1,...,x_I \mid C_i) = P(X_1 \le x_I,...,X_I \le x_I \mid C_i) = \frac{P(X_1 \le x_1,...,X_I \le x_I, C_i)}{P(C_i)}$$

  - densidad

$$f_{\vec{X}}(x_1,...,x_I \mid C_i) = \frac{\partial F_{\vec{X}}(x_1,...,x_I \mid C_i)}{\partial x_1 ... \partial x_I}$$

Ej.: distribución normal multivariada

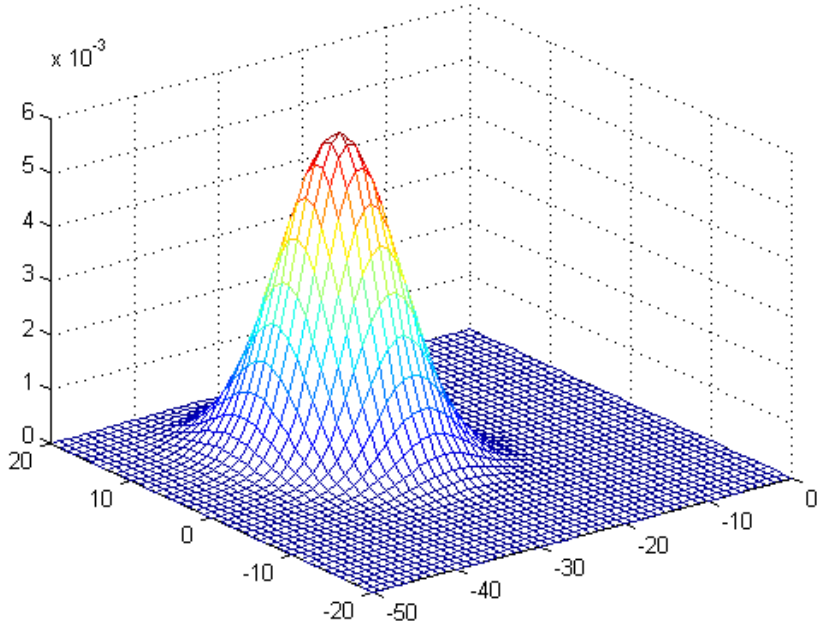
- Parámetros: vector de medias y matriz covarianzas

$$f(\vec{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|S|}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^t S^{-1}(\vec{x}-\vec{\mu})\right]$$

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}; \quad S = \begin{bmatrix} \sigma_{x1}{}^2 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_F} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{x_Fx_1} & \sigma_{x_nx_2} & \cdots & \sigma_{x_F}{}^2 \end{bmatrix}$$

- Ejemplo

$$\vec{\mu} = \begin{bmatrix} -30 \\ 5 \end{bmatrix}; \quad S = \begin{bmatrix} 21 & -6 \\ -6 & 21 \end{bmatrix}$$



Teorema de Bayes aplicado a clasificación

$$P(C_i \mid \vec{X}) = \frac{f(\vec{X} \mid C_i)p(C_i)}{f(\vec{X})}$$

- Probabilidad a posteriori: es la probabilidad de que el ejemplo tenga clase Ci: $P(C_i \mid \vec{X})$
- Probabilidad a priori: $P(C_i)$ es la probabilidad total de cada clase
- Verosimilitud: $f(\vec{X} \mid C_i)$: es la distribución de C_i aplicada a $\vec{X}$
- Densidad total: $f(\vec{X}) = f(\vec{X} \mid C_1)P(C_1) + \ldots + f(\vec{X} \mid C_M)P(C_M)$

Criterio de clasificación MAP:

$$\text{Clase}(\vec{X}) = \underset{i}{\text{máximo}} \left\{P(C_i \mid \vec{X})\right\} = \underset{i}{\text{máximo}} \left\{f(\vec{X} \mid C_i)p(C_i)\right\}$$

- función discriminante de C_i: proporcional a su prob a posteriori:

$$g_i(\vec{X}) = f(\vec{X} \mid C_i)p(C_i)$$

- la clase es la de aquella que maximiza el discriminante

Clasificación bayesiana y distrib normal

- Distribuciones condicionales gaussianas. Para cada clase C_i hay una función discriminante de parámetros μ_{ij}, σ_{ij}, j=1...I

$$g_i(\vec{x}) = \log(P(C_i) f(\vec{x} \mid C_i)) = \log \frac{P(C_i)}{(2\pi)^{n/2} \sigma_{1i}\sigma_{2i}...\sigma_{Fi}} - \frac{1}{2}(\vec{x} - \vec{\mu})^t S^{-1} (\vec{x} - \vec{\mu})$$

$$\text{simplificaion}: K - \frac{1}{2}\sum_{i=1}^{F} (x_j - \mu_{ij})^2 / {\sigma_{ij}}^2$$

- Parámetros de distribución condicionada a cada clase
- Regiones de decisión:
  - Funciones cuadráticas (hipérbolas) dadas por diferencias:

$$g_{ij}(\vec{x}) = g_i(\vec{x}) - g_j(\vec{x})$$

  - Si son iguales, y diagonales: regiones lineales (caso particular)

Ejemplo con distribución normal

- C1:

$$P_1 = 0.3; \quad C_1 = \begin{bmatrix} -30 & 5 \end{bmatrix}^t; \quad R_1 = \begin{bmatrix} 21 & -6 \\ -6 & 21 \end{bmatrix}$$

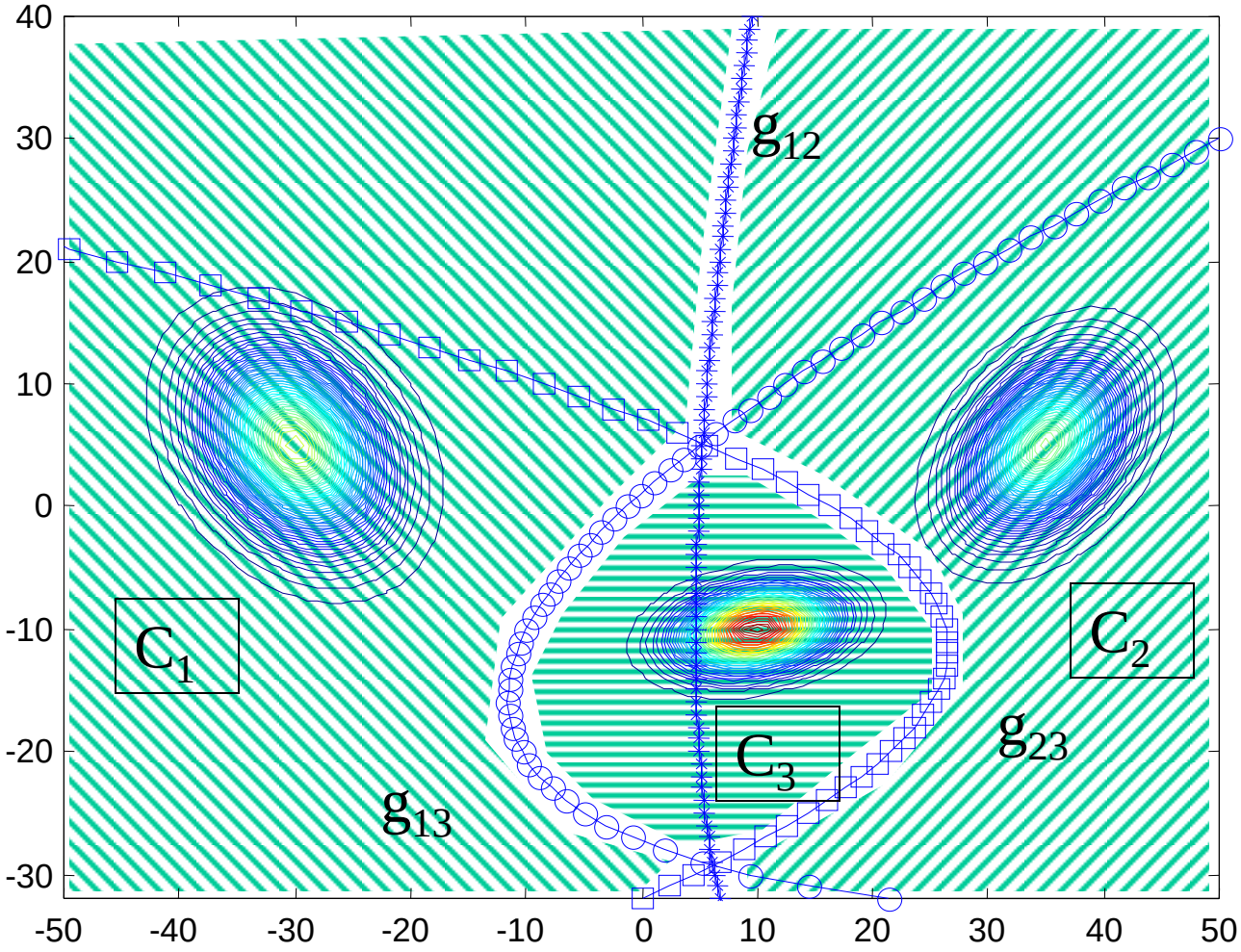
- C2:

$$P_2 = 0.2; \quad C_2 = \begin{bmatrix} 35 & 5 \end{bmatrix}^t; \quad R_2 = \begin{bmatrix} 16 & 6 \\ 6 & 16 \end{bmatrix}$$

- C3:

$$P_3 = 0.5; \quad C_3 = \begin{bmatrix} 10 & -10 \end{bmatrix}^t; \quad R_3 = \begin{bmatrix} 16 & 2 \\ 2 & 4 \end{bmatrix}$$

Ejemplo


40
30
20
10
0
-10
-20
-30
-50
-40
-30
-20
-10
0
10
20
30
40
50
g_{12}
C_1
C_2
C_3
g_{23}
g_{13}

Resumen clasificador bayesiano numérico

- Algoritmo:
  - Estimar parámetros de cada clase C_i (entrenamiento)

$$C_i : \{\vec{X}_1^{(i)}, \ldots, \vec{X}_{n_i}^{(i)}\} \longrightarrow \bar{\mu}_i, \quad C_i$$

$$\hat{\bar{\mu}}_i = \frac{1}{n} \sum_{j=1}^{n_i} \vec{x}_j^{(i)}$$

$$C_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\vec{x}_i - \mu_i)(\vec{x}_i - \mu_i)^t$$

  - Estimar probabilidad de cada clase

$$\hat{P}(C_i) = \frac{n_i}{N}; \qquad n = \sum_{i=1}^{M} n_i$$

  - Obtener regiones de decisión: $g_{ij}(.)$

Clasificación Bayesiana con Atributos Nominales

Atributos nominales con valores discretos

- $A_i=\{V_1,...,V_{ni}\}$: atributo con n_i valores posibles
- Pasamos de densidades a probabilidades: probabilidad a priori: $p(A_i=V_j|C_k)$?
- Estimación “contando” el número de casos:

$$p(A_i = V_j \mid C_k) = \frac{\text{nº de ejemplos de clase } C_k \text{ con } A_i = V_j}{\text{nº de ejemplos de clase } C_k}$$

Clasificación Bayesiana con Atributos Nominales

- Simplificación: independencia de atributos (“Naive Bayes”): la probabilidad conjunta de varios atributos se pone como producto

$$X_i = (A_1 = V_1, A_2 = V_2, ..., A_I = V_I)$$
$$p(X_i \mid C_k) = p(A_1 = V_1 \mid C_k) * p(A_2 = V_2 \mid C_k * ... * p(A_I = V_I \mid C_k)$$

- Clasificación:

$$p(C_k \mid X_i) = \frac{p(X_i \mid C_k) * p(C_k)}{p(X_i)} =$$
$$\frac{p(A_1 = V_1 \mid C_k) * p(A_2 = V_2 \mid C_k) * ... * p(A_F = V_F \mid C_k) * p(C_k)}{p(X_i)}$$

Ejemplo con atributos nominales

SALARIO	CLIENTE	EDAD	HIJOS	CRÉDITO
Poco	Sí	Joven	Uno	**NO**
Mucho	Si	Joven	Uno	**SI**
Mucho	Si	Joven	Uno	**SI**
Poco	Si	Joven	Uno	**NO**
Mucho	Si	Joven	Dos	**SI**
Poco	Si	Joven	Dos	**NO**
Mucho	Si	Adulto	Dos	**SI**
Mucho	Si	Adulto	Dos	**SI**
Poco	No	Adulto	Dos	**NO**
Mucho	Si	Adulto	Dos	**SI**
Medio	No	Adulto	Tres	**NO**
Mucho	Si	Adulto	Dos	**SI**
Medio	Si	Adulto	Dos	**SI**
Medio	No	Adulto	Tres	**NO**
Medio	No	Adulto	Dos	**SI**
Mucho	No	Mayor	Tres	**NO**
Poco	No	Mayor	Tres	**SI**
Poco	No	Mayor	Tres	**SI**
Mucho	No	Mayor	Tres	**NO**
Mucho	No	Mayor	Tres	**SI**

$p(SI) = 12/20$

$p(NO) = 8/20$

Salario \ Crédito	No	Sí
Poco	4/8	2/12
Mucho	2/8	8/12
Medio	2/8	2/12

Cliente \ Crédito	No	Sí
Sí	3/8	8/12
No	5/8	4/12

Edad \ Crédito	No	Sí
Joven	3/8	3/12
Adulto	3/8	6/12
Mayor	2/8	3/12

Hijos \ Crédito	No	Sí
Uno	2/8	2/12
Dos	2/8	7/12
Tres	4/8	3/12

Ejemplo con atributos nominales

- Ej.: **(salario=poco, cliente=si, edad=adulto, hijos=tres)**

$$p(SI \mid X_i) =$$
$$p(s = poco \mid SI) * p(c = si \mid SI) * p(e = adulto \mid SI) * p(h = tres \mid SI) * p(SI) / p(X_i) =$$
$$2/12 * 8/12 * 6/12 * 3/12 * 12 / 20 / p(Xi) = 0.0083 / p(Xi)$$

$$p(NO \mid X_i) =$$
$$p(s = poco \mid NO) * p(c = si \mid NO) * p(e = adulto \mid NO) * p(h = tres \mid NO) * p(NO) / p(X_i) =$$
$$4/8 * 3/8 * 3/8 * 4/8 * 8 / 20 / p(Xi) = 0.0141 / p(Xi)$$

Atributos sin valores

- Si el ejemplo a clasificar no tiene un atributo, simplemente se omite.
  - Ej.: (**salario=poco, cliente=si, edad=?, hijos=3**)

$$p(SI \mid X_i) =$$
$$p(s = poco \mid SI) * p(c = si \mid SI) * p(h = tres \mid SI) * p(SI) / p(X_i) =$$
$$2/12 * 8/12 * 3/12 * 12/20 / p(Xi) = 0.0167 / p(Xi)$$
$$p(NO \mid X_i) =$$
$$p(s = poco \mid NO) * p(c = si \mid NO) * p(h = tres \mid NO) * p(NO) / p(X_i) =$$
$$4/8 * 3/8 * 4/8 * 8/20 / p(Xi) = 0.0375 / p(Xi)$$

- Si hay faltas en la muestra de entrenamiento, no cuentan en la estimación de probabilidades de ese atributo

Faltas en atributo EDAD

SALARIO	CLIENTE	EDAD	HIJOS	CRÉDITO
Poco	Sí	Joven	Uno	**NO**
Mucho	Si	Joven	Uno	**SI**
Mucho	Si	Joven	Uno	**SI**
Poco	Si	**?**	Uno	**NO**
Mucho	Si	**?**	Dos	**SI**
Poco	Si	**?**	Dos	**NO**
Mucho	Si	**?**	Dos	**SI**
Mucho	Si	Adulto	Dos	**SI**
Poco	No	Adulto	Dos	**NO**
Mucho	Si	Adulto	Dos	**SI**
Medio	No	Adulto	Tres	**NO**
Mucho	Si	Adulto	Dos	**SI**
Medio	Si	Adulto	Dos	**SI**
Medio	No	Adulto	Tres	**NO**
Medio	No	Adulto	Dos	**SI**
Mucho	No	Mayor	Tres	**NO**
Poco	No	Mayor	Tres	**SI**
Poco	No	Mayor	Tres	**SI**
Mucho	No	Mayor	Tres	**NO**
Mucho	No	Mayor	Tres	**SI**

p(SI) = 12/20

p(NO) = 8/20

Salario \ Crédito	No	Sí
Poco	4/8	2/12
Mucho	2/8	8/12
Medio	2/8	2/12

Cliente \ Crédito	No	Sí
Sí	3/8	8/12
No	5/8	4/12

Edad \ Crédito	No	Sí
Joven	**1/6**	**2/10**
Adulto	**3/6**	**5/10**
Mayor	**2/6**	**3/10**

Hijos \ Crédito	No	Sí
Uno	2/8	2/12
Dos	2/8	7/12
Tres	4/8	3/12

Atributos no representados. Laplace

- Problema: con muestra poco representativa, puede ocurrir que en alguna clase, un valor de atributo no aparezca: $p(A_i=V_j|C_k)=0$
  - Cualquier ejemplo X con $A_i=V_j$ generará $P(C_k|X)=0$, independientemente de los otros atributos!
- Se suele modificar la estimación de las probabilidades a priori con un factor que elimina los ceros.
  - Ej.: P(Edad|Crédito=NO)= $\left\{\text{Joven}:\frac{3}{8},\quad \text{Adulto}:\frac{3}{8},\quad \text{Mayor}:\frac{2}{8}\right\}$
  - Ley μ: $\left\{\text{Joven}:\frac{3+\mu/3}{8+\mu},\quad \text{Adulto}:\frac{3+\mu/3}{8+\mu},\quad \text{Mayor}:\frac{2+\mu/3}{8+\mu}\right\}$
  - A veces simplemente se inicializan las cuentas a 1 en vez de 0:

$$\left\{\text{Joven}:\frac{3+1}{8+3},\quad \text{Adulto}:\frac{3+1}{8+3},\quad \text{Mayor}:\frac{2+1}{8+3}\right\}$$

Atributos mixtos

- Independencia de atributos ("Naive Bayes")

$$p(X_i \mid C_k) = p(A_1 = V_1 \mid C_k) * p(A_2 = V_2 \mid C_k) * ... * p(A_F = V_F \mid C_k)$$

  - Atributos discretos: probabilidades a priori con cada clase C_k

$$p(A_i = V_j \mid C_k) = \frac{\text{nº de ejemplos de clase } C_k \text{ con } A_i = V_j}{\text{nº de ejemplos de clase } C_k}$$

  - Atributos continuos: densidades de clase Ck: normales de parámetros μ_k, σ_k

$$p(A_i = V_j \mid C_k) \rightarrow f_{A_i}(V_j \mid C_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp\left[-\frac{1}{2} \frac{(V_j - \mu_{ik})^2}{\sigma_{ik}^2} \right]$$

Ejemplo con atributos mixtos

SALARIO	CLIENTE	EDAD	HIJOS	CRÉDITO
525	Sí	Joven	1	**NO**
2000	Si	Joven	1	**SI**
2500	Si	Joven	1	**SI**
470	Si	Joven	1	**NO**
3000	Si	Joven	2	**SI**
510	Si	Joven	2	**NO**
2800	Si	Adulto	2	**SI**
2700	Si	Adulto	2	**SI**
550	No	Adulto	2	**NO**
2600	Si	Adulto	2	**SI**
1100	No	Adulto	3	**NO**
2300	Si	Adulto	2	**SI**
1200	Si	Adulto	2	**SI**
900	No	Adulto	3	**NO**
800	No	Adulto	2	**SI**
800	No	Mayor	3	**NO**
1300	No	Mayor	3	**SI**
1100	No	Mayor	3	**SI**
1000	No	Mayor	3	**NO**
4000	No	Mayor	3	**SI**

$p(SI) = 12/20$

$p(NO) = 8/20$

Crédito / Salario	No	Sí
Media	732	2192
Desv Estándar	249	942

Crédito / Cliente	No	Sí
Sí	3/8	8/12
No	5/8	4/12

Crédito / Edad	No	Sí
Joven	3/8	3/12
Adulto	3/8	6/12
Mayor	2/8	3/12

Crédito / Hijos	No	Sí
Media	2.25	2.08
Desv Estándar	0.89	0.67

Ejemplo con atributos mixtos

- Ej.: (salario=700, cliente=si, edad=adulto, hijos=3)

$$p(SI \mid X_i) =$$
$$f_S(s = 700 \mid SI) * p(c = si \mid SI) * p(e = adulto \mid SI) * f_H(h = 3 \mid SI) * p(SI) / p(X_i) =$$
$$\frac{1}{\sqrt{2\pi}942} \exp\left[-\frac{1}{2} \frac{(700 - 2192)^2}{942^2} \right] * 8/12 * 6/12 * \frac{1}{\sqrt{2\pi}0.67} \exp\left[-\frac{1}{2} \frac{(3 - 2.08)^2}{0.67^2} \right] * 12/20 * 1/P(X_i) =$$
$$= 5.61e-6 / p(Xi)$$

$$p(NO \mid X_i) =$$
$$f_S(s = 700 \mid NO) * p(c = si \mid NO) * p(e = adulto \mid NO) * f_H(h = 3 \mid NO) * p(NO) / p(X_i) =$$
$$\frac{1}{\sqrt{2\pi}249} \exp\left[-\frac{1}{2} \frac{(700 - 732)^2}{249^2} \right] * 3/8 * 3/8 * \frac{1}{\sqrt{2\pi}0.89} \exp\left[-\frac{1}{2} \frac{(3 - 2.25)^2}{0.89^2} \right] * 8/20 * 1/P(X_i) =$$
$$= 2.81e-5 / p(Xi)$$

Clasificación con costes

- MAP proporciona clasificación con mínima prob. de Error
  - Coste de decisión $D_i \mid \vec{X}$: prob. Error total= $1 - P(C_i \mid \vec{X})$
- Con frecuencia los costes son asimétricos, y unos errores son más graves que otros. Matriz de costes

Clasificado como →

Clase real ↓

$$\begin{pmatrix} 0 & c_{12} & c_{13} \\ c_{21} & 0 & c_{23} \\ c_{31} & c_{32} & 0 \end{pmatrix}$$

- Costes de cada decisión. Criterio de mínimo coste medio: dada una decisión, promedio los costes de cada equivocación y su coste:

$$\cos te(D_1 \mid \vec{X}) = c_{21} p(C_2 \mid \vec{X}) + c_{31} p(C_3 \mid \vec{X})$$

$$\cos te(D_2 \mid \vec{X}) = c_{12} p(C_1 \mid \vec{X}) + c_{32} p(C_3 \mid \vec{X})$$

$$\cos te(D_3 \mid \vec{X}) = c_{13} p(C_1 \mid \vec{X}) + c_{23} p(C_2 \mid \vec{X})$$

Ejemplo de clasificación con costes

- Clasificación de setas con dos atributos, (X, Y) y tres categorías: V*enenosa, Mal sabor, comestible*: {*V, MS, C*}

Clasificado como

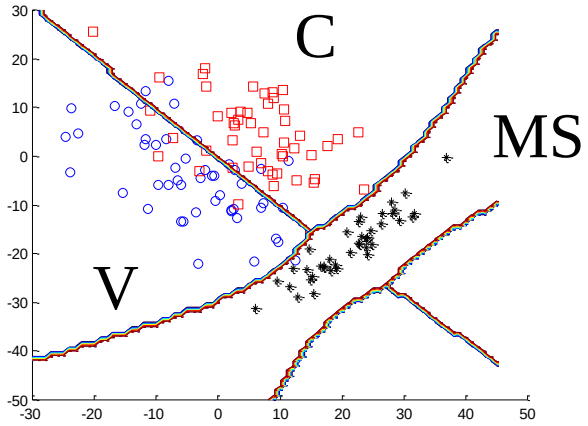
Clase real	V	MS	C
V	0	1000	1000
MS	1	0	10
C	1	1	0

$$V: \quad \mu_1 = [-5 \quad -5]^t; \quad C_1 = \begin{bmatrix} 71 & -50 \\ -50 & 71 \end{bmatrix}$$

$$C: \quad \mu_2 = [5 \quad 5]^t; \quad C_2 = \begin{bmatrix} 71 & -40 \\ -40 & 71 \end{bmatrix}$$

$$MS: \quad \mu_3 = [20 \quad -20]^t; \quad C_3 = \begin{bmatrix} 51 & 45 \\ 45 & 51 \end{bmatrix}$$

Mínimo error



Mínimo coste