Clasificadores generativos vs discriminantes

José Anibal ARIAS

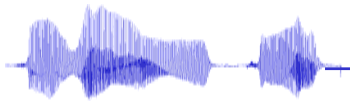
Seminario de investigación

Universidad Tecnológica de la Mixteca

15-Abril-2010

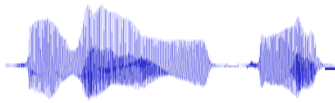
Temario

- Introducción
- Datos multivariables
- Modelos generativos
- Modelos discriminantes
- Ejemplo práctico
- Análisis teórico



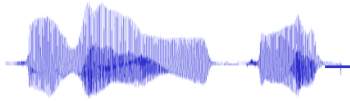
Introducción

- Existen diferentes razones -ademas de una especie de sabiduría popular- para utilizar clasificadores discriminantes en lugar de generativos
- V. Vapnik « debemos resolver el problema de clasificación directamente en lugar de tratar de resolver problemas intermedios »
- Clasificadores generativos: GMM, Redes Bayesianas, HMM, ...
- Clasificadores discriminantes: k-nn, SVM, AdaBoost, Conditional Random Fields, …



Temario

- Introducción
- Datos multivariables
- Modelos generativos
- Modelos discriminantes
- Ejemplo práctico
- Análisis teórico



Datos multivariables

- En muchas aplicaciones, los eventos observados generan vectores (entradas, características, atributos) de varias dimensiones:

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_d^1 \\ x_1^2 & x_2^2 & \dots & x_d^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^N & x_2^N & \dots & x_d^N \end{bmatrix}$$

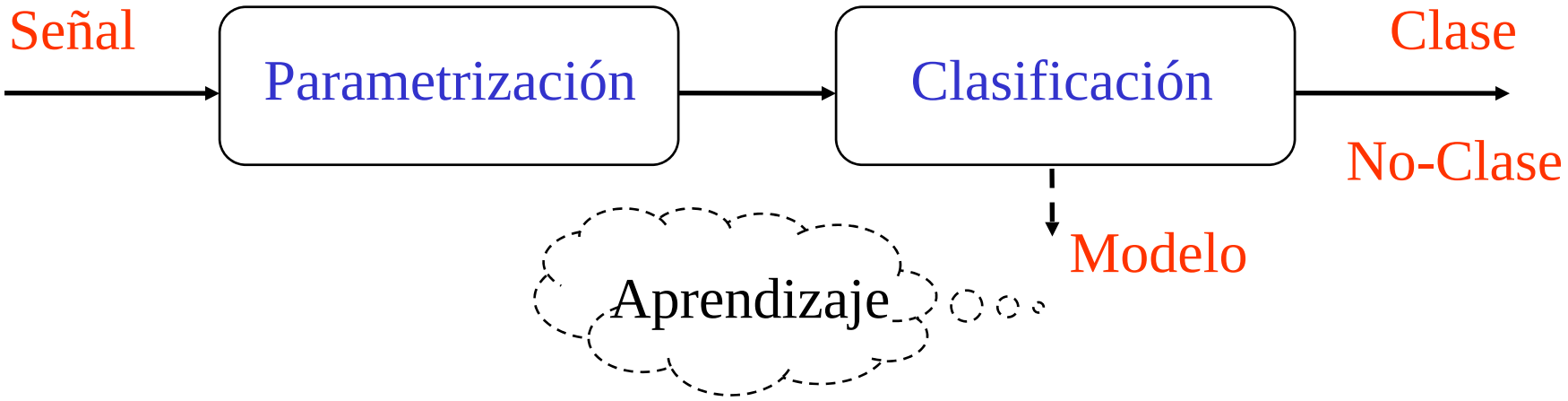
Datos multivariables

- En los problemas de aprendizaje supervisado, una parte de los datos está « etiquetada », y a cada x le corresponde una etiqueta C_k, k = 1,…K.

$$(x^1, C_k^1), \ldots, (x^m, C_k^m) \in \Re^d \times \aleph$$

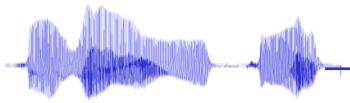
Clasificación

- Se evalúa la información contenida en el vector de parámetros para tomar una desición sobre la clase o región del espacio al que pertenece



Temario

- Introducción
- Datos multivariables
- Modelos generativos
- Modelos discriminantes
- Ejemplo práctico
- Análisis teórico



Modelos generativos

- Objetivo del aprendizaje automático: encontrar la relación entre los conjuntos X (datos) y C (etiquetas)
- El enfoque generativo trata de encontrar un modelo paramétrico de la distribución conjunta p(X,C), utilizarla para definir la distribución condicional p(C|X) y estar en condiciones de hacer predicciones de la etiqueta C para los valores de X
- Se denomina generativo porque una vez que se ha encontrado la distribución p(X,C), ésta se puede utilizar para generar muestras sintéticas de X según su etiqueta C

Probabilidad básica

$$P(X,\ Y) = P(Y \mid X) \cdot P(X)$$

$$Bayes \ :$$

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}$$

- Utilizamos el conjunto de aprendizaje para calcular P(X|Y) y P(Y)

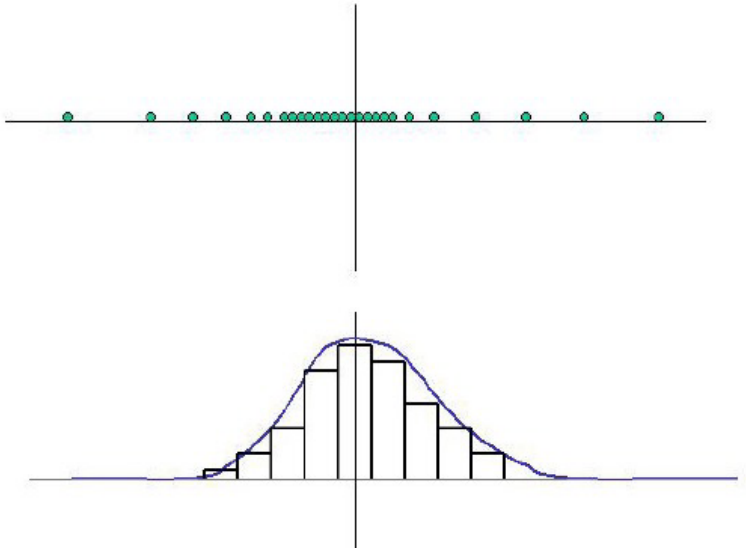
Ley Gaussiana - propiedades

- Teorema del límite central: la distribución de la suma de variables aleatorias converge en una ley Gaussiana cuando la cantidad de variables es muy grande
- Dada una distribución conjunta gaussiana, la distribución condicional de un conjunto dado otro es también gaussiana

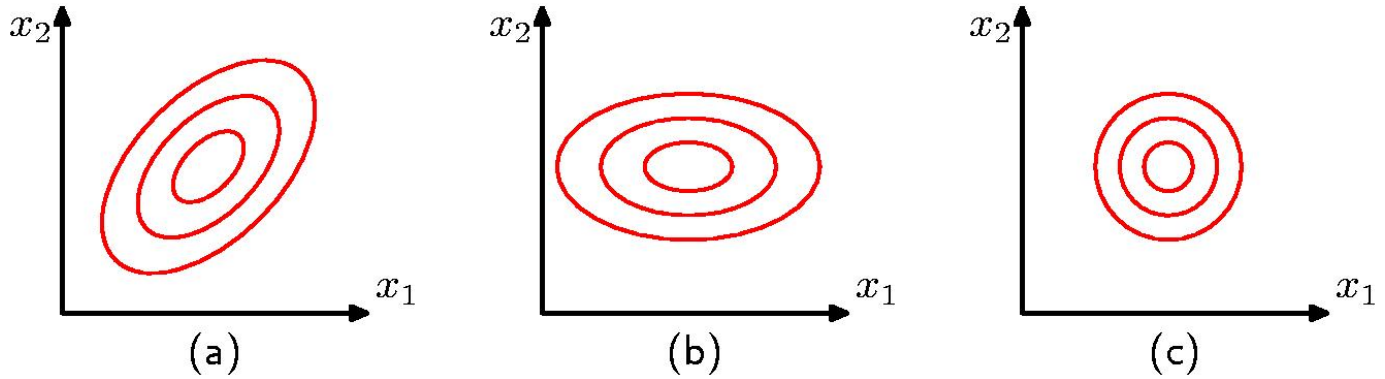
Ley Gaussiana - caso monovariable

- Distribución gaussiana

$$\mathcal{N}(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(O-\mu)^2}{2\sigma^2}}$$



- En dos dimensiones



Ley Gaussiana - caso multivariable

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{O}-\mu)^T \Sigma^{-1}(\mathbf{O}-\mu)}$$

$$-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{O}-\mu)^T\Sigma^{-1}(\mathbf{O}-\mu)$$

- Si la matriz de covarianza es diagonal

$$-\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\ln\sigma_i - \frac{1}{2}\sum_{i=1}^{n}(O_i-\mu_i)^2/\sigma_i^2$$

Estimación de los parámetros

- Consideramos que la verosimilitud de una cadena de observaciones es igual al producto de las verosimilitudes individuales

$$\mathcal{L}(O_1^N|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(O_i-\mu)^2}{2\sigma^2}}$$

$$L(O_1^N|\mu,\sigma) = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\frac{(O_i-\mu)^2}{\sigma^2}$$

- ML: se deriva L con respecto a la media y a la covarianza y se iguala a cero

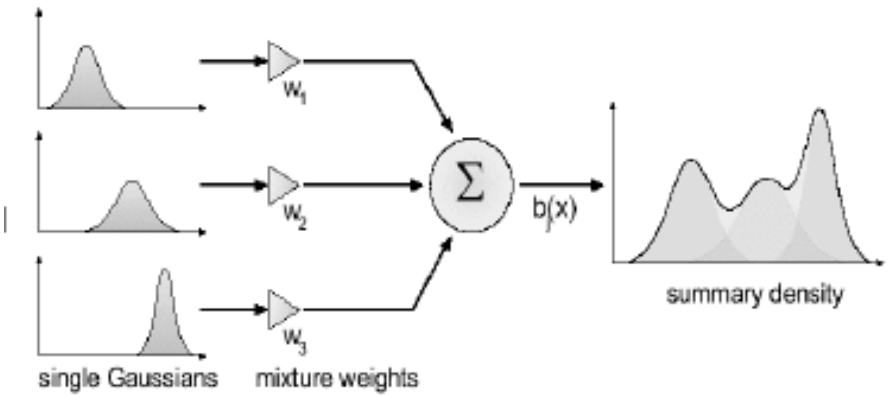
$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{O}_i \qquad \mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{O}_i-\mu)^T(\mathbf{O}_i-\mu)$$

GMM

- Mezcla de leyes normales

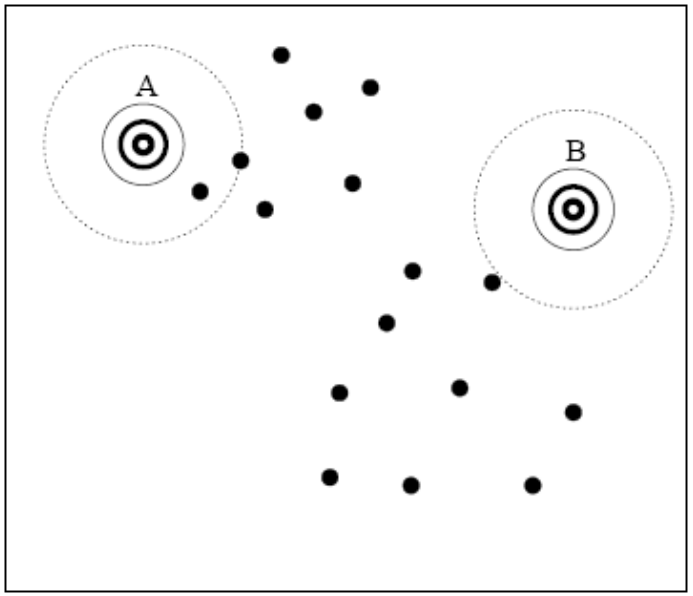
$$f(x)=\sum_{k=1}^{K} p_k N(x,\mu_k,\Sigma_k)$$

$$1\geq p_k \geq 0 \quad et \quad \sum_{k=1}^{K} p_k = 1$$

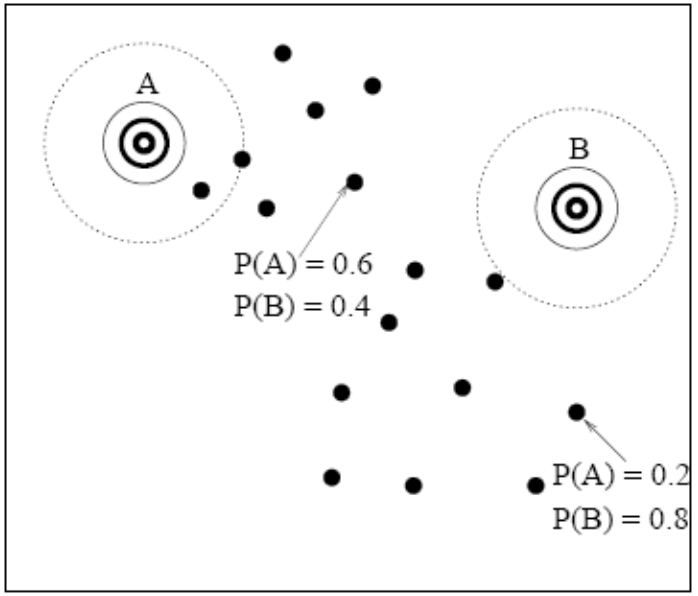


- Si supieramos de antemano que componente genero qué dato, la solución sería asociar cada dato a la ley correspondiente

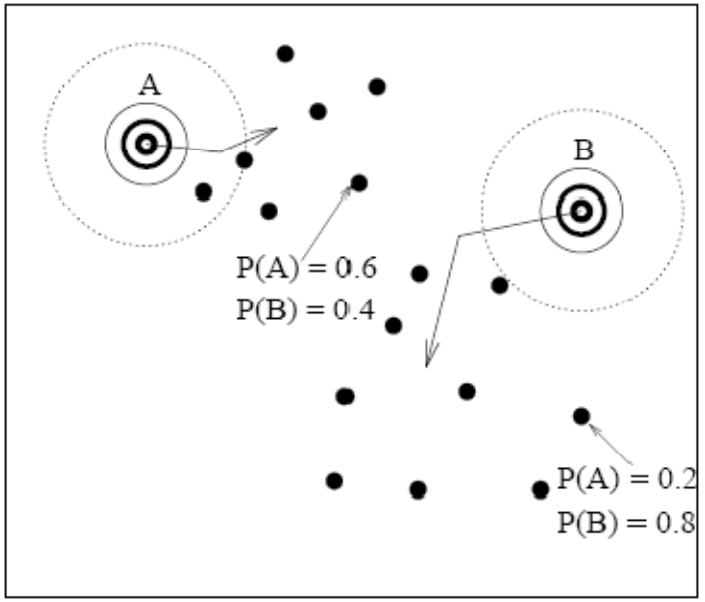
¿Que Gaussiana generó qué dato?



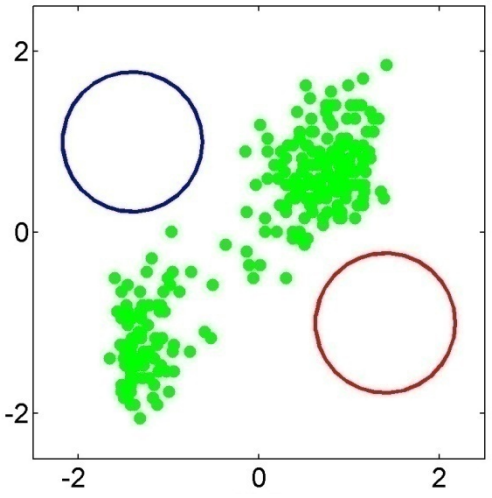
E

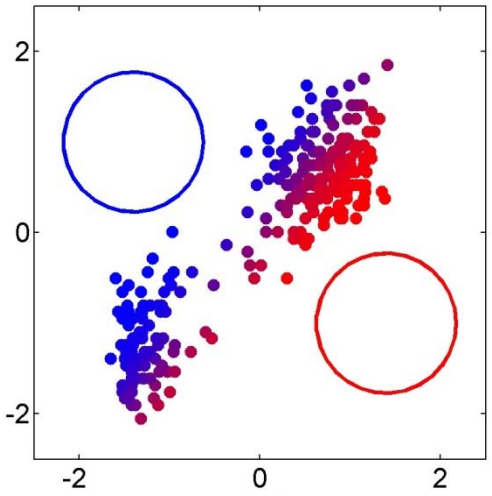


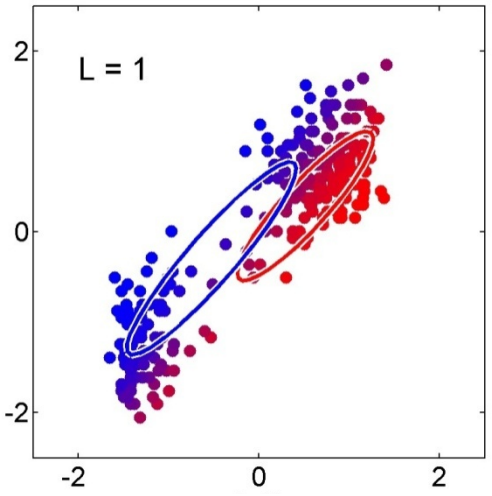
M

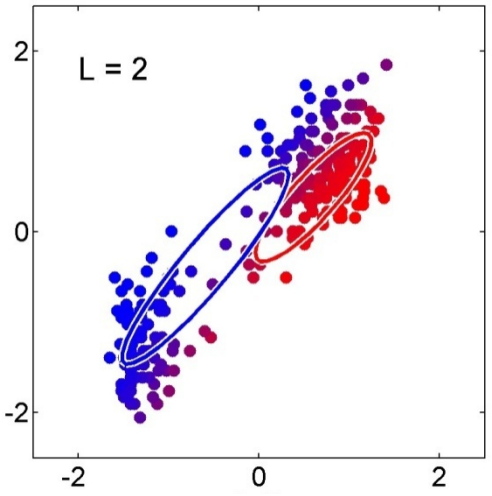


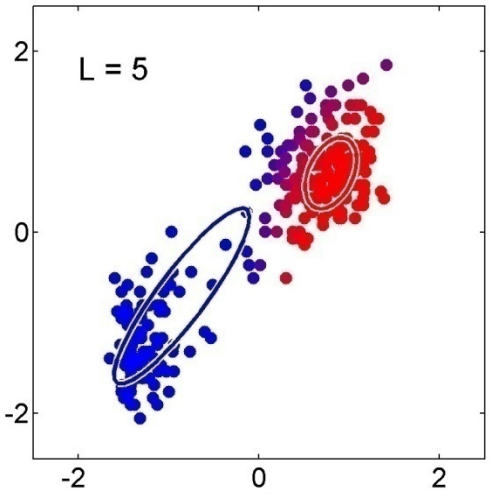
EM en acción

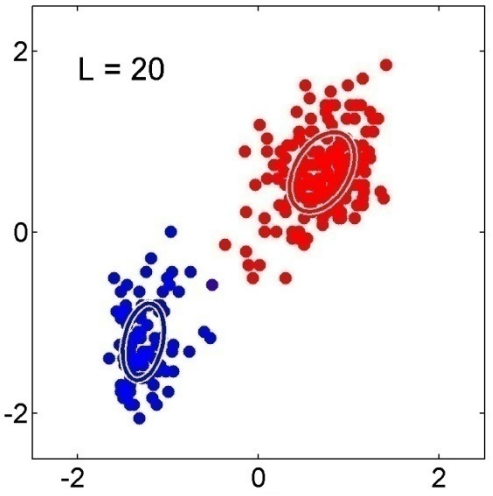






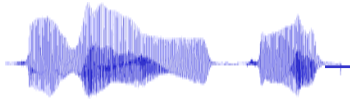






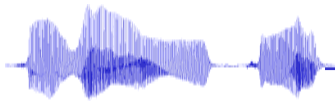
Temario

- Introducción
- Datos multivariables
- Modelos generativos
- **Modelos discriminantes**
- Ejemplo práctico
- Análisis teórico



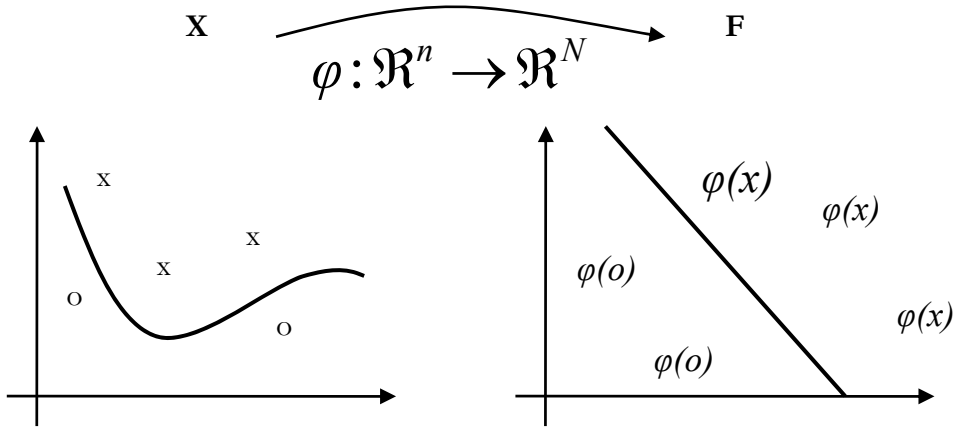
Modelos discriminantes

- Estimamos directamente la distribución condicional p(C|X) con una función $f : X \rightarrow C$
- Llamamos a este enfoque discriminante porque la distribución condicional establece una discriminación directa entre los diferentes valores de Y
- En la práctica, el poder de generalización de los métodos discriminantes es mas alto que el de los modelos generativos

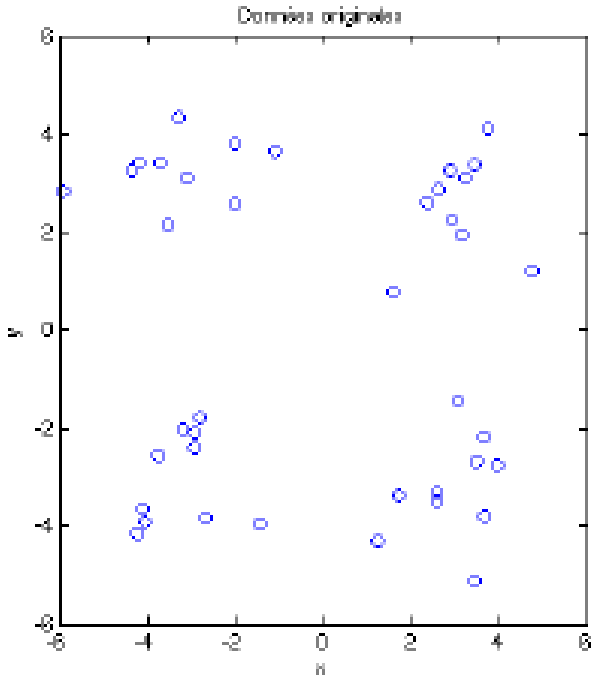


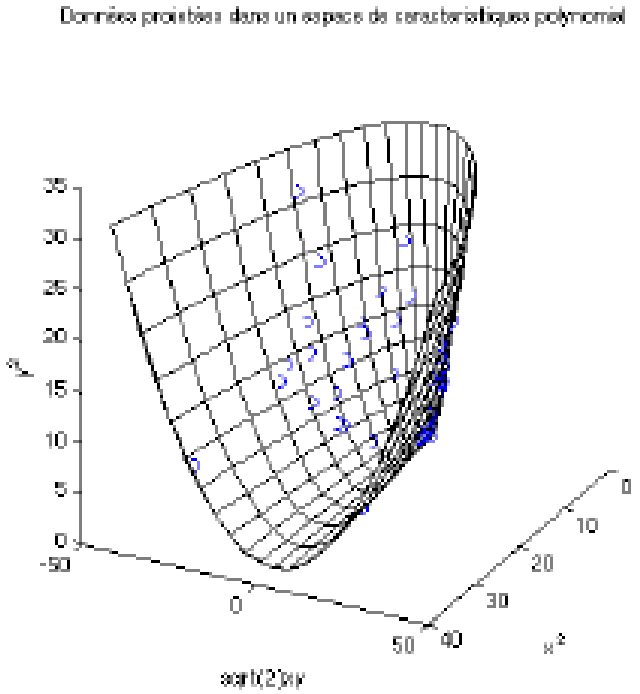
Separadores de vasto margen (SVM)

- Este método se basa en la transformación de datos hacia un espacio mas « informativo » y en la existencia de un hiperplano separador en este espacio (denominado « espacio de características »)



Transformación polinomial





SVM lineal - I

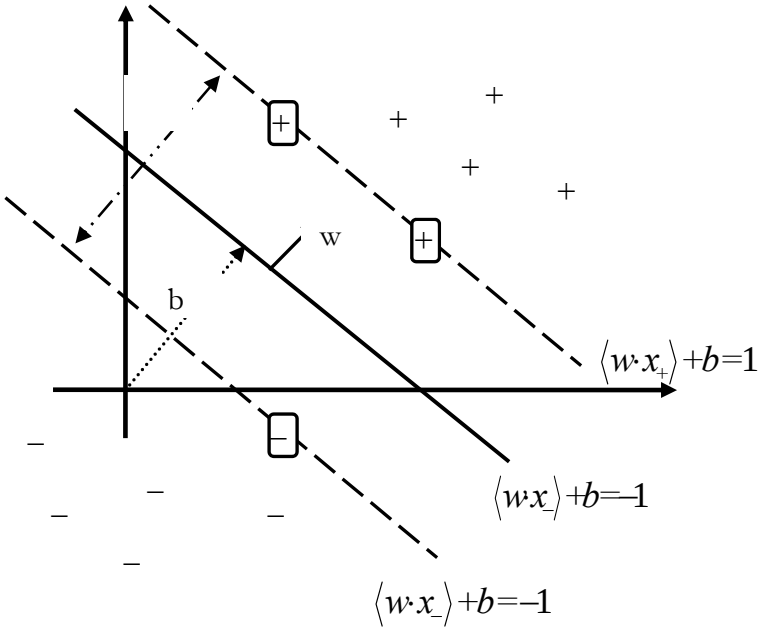
- A diferencia del perceptron, se maximiza el margen del hiperplano separador

$$\begin{cases} w \cdot x_i + b \geq +1 & si \quad y_i = +1 \\ w \cdot x_i + b \leq -1 & si \quad y_i = -1 \end{cases}$$

$$\gamma = \left(\left\langle \frac{w}{\|w\|} \cdot x_+ \right\rangle - \left\langle \frac{w}{\|w\|} \cdot x_- \right\rangle \right)$$

$$\gamma = \frac{1}{\|w\|} \left(\langle w \cdot x_+ \rangle - \langle w \cdot x_- \rangle \right)$$

$$\gamma = \frac{2}{\|w\|}$$



SVM lineal - II

- Lagrangiano L
- Minimización de L con respecto a *w* y b y maximización de W con respecto a α

$$L = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \left[y_i (\langle w \cdot x_i \rangle + b) - 1 \right]$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0$$

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

$$W = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle$$

SVM lineal - III

- La función de decisión para la clasificación de vectores desconocidos **u** está basada en los « vectores de soporte » :

$$f(u) = sign\left(\sum_{i \in sv} \alpha_i y_i \langle x_i \cdot u \rangle + b\right)$$

- A menudo los datos están afectados por el ruido y no existe una separación lineal entre las clases
- Solución: se relajan las restricciones del margen

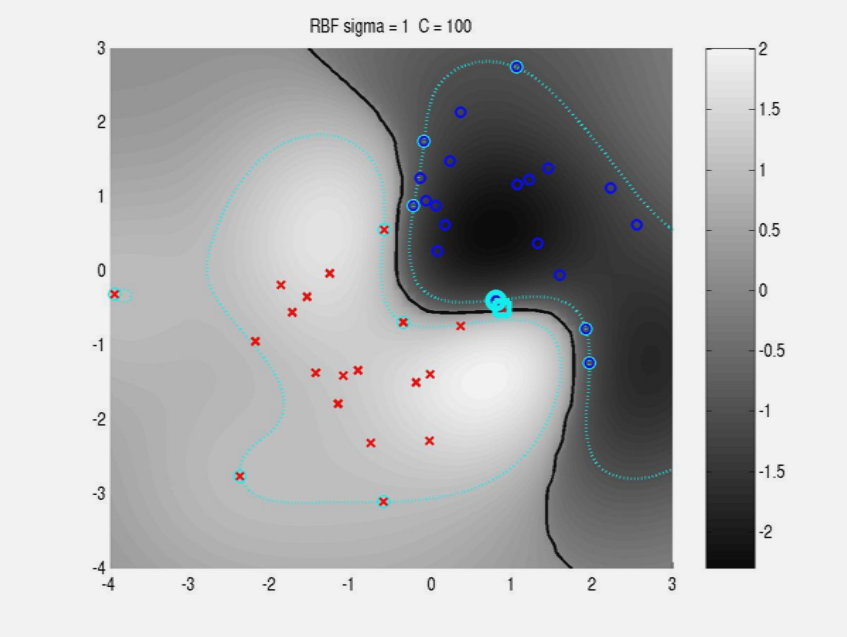
$$\begin{cases} w \cdot x_i + b \geq +1 - \zeta_i & si \quad y_i = +1 \\ w \cdot x_i + b \leq -1 + \zeta_i & si \quad y_i = -1 \end{cases}$$

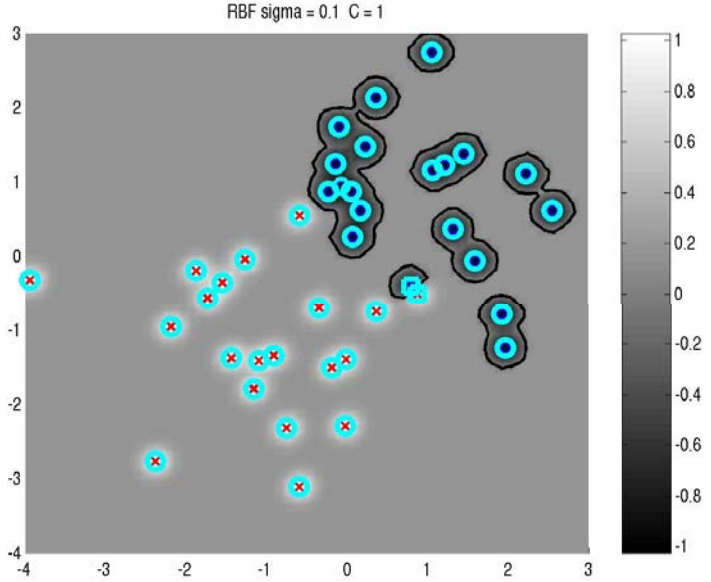
SVM non lineales

$$\Phi : X \rightarrow F$$

$$\kappa(x, z) = \langle \Phi(x) \cdot \Phi(z) \rangle$$

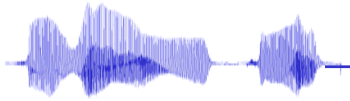
$$f(u) = sign\left(\sum_{i \in sv} \alpha_i y_i \kappa(x_i \cdot u) + b \right)$$



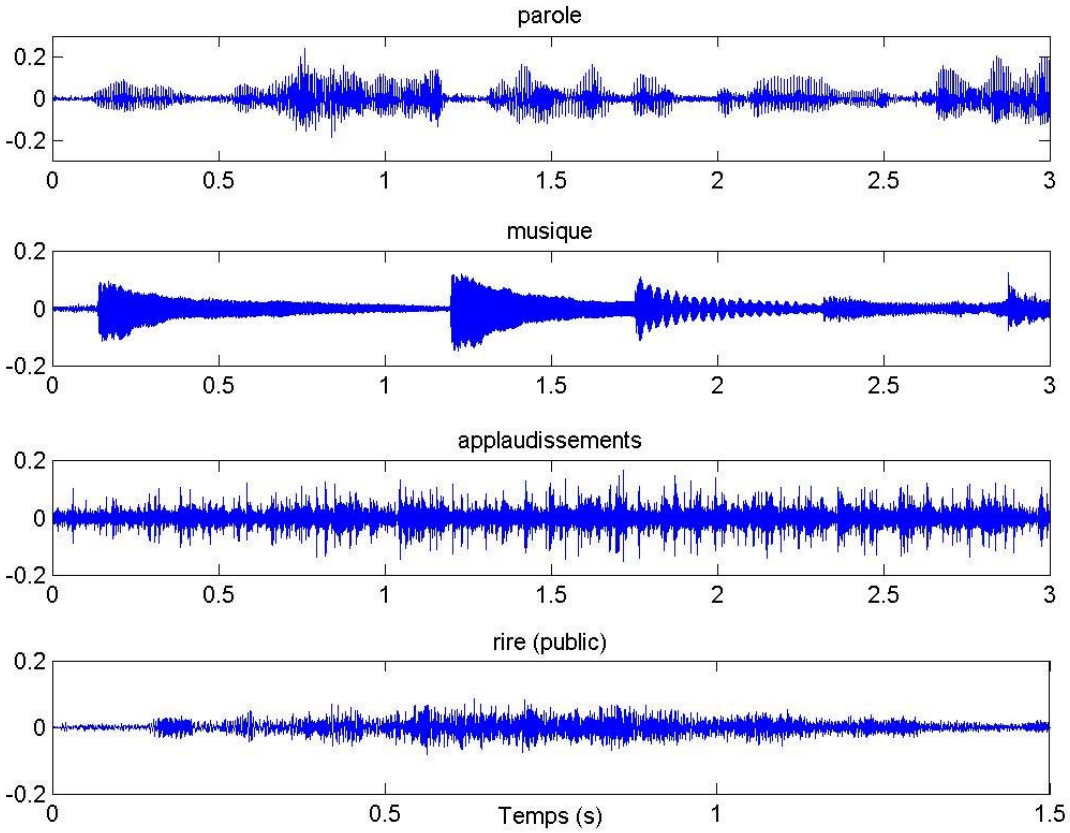


Temario

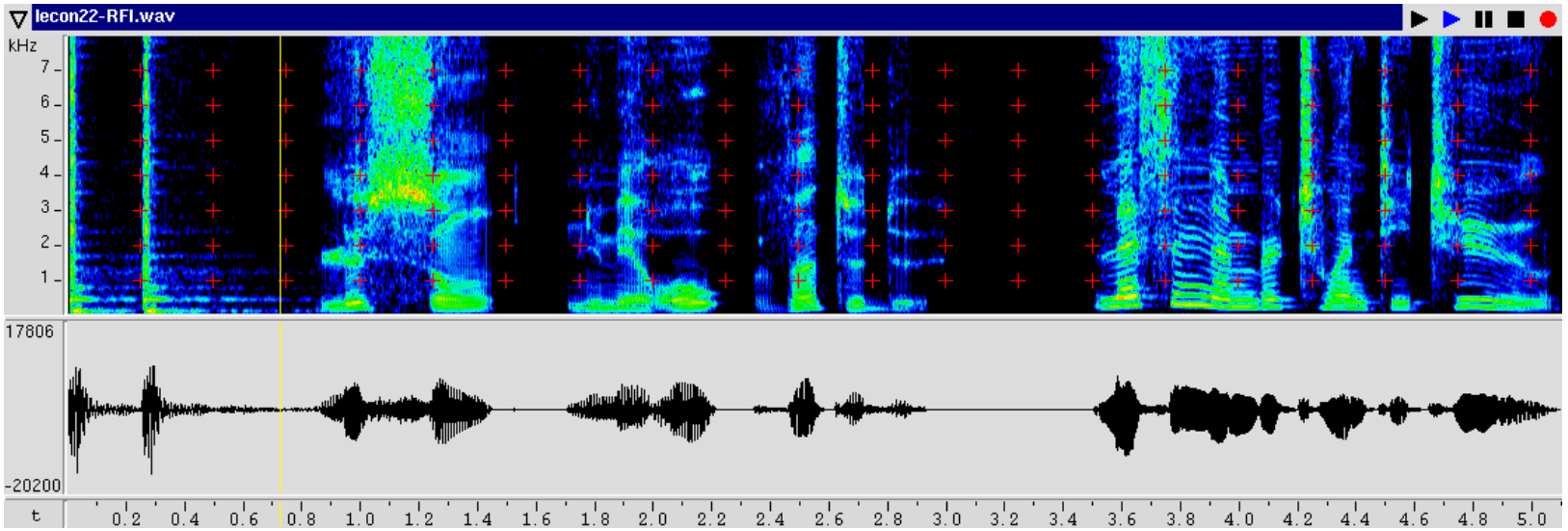
- Introducción
- Datos multivariables
- Modelos generativos
- Modelos discriminantes
- Ejemplo práctico
- Análisis teórico

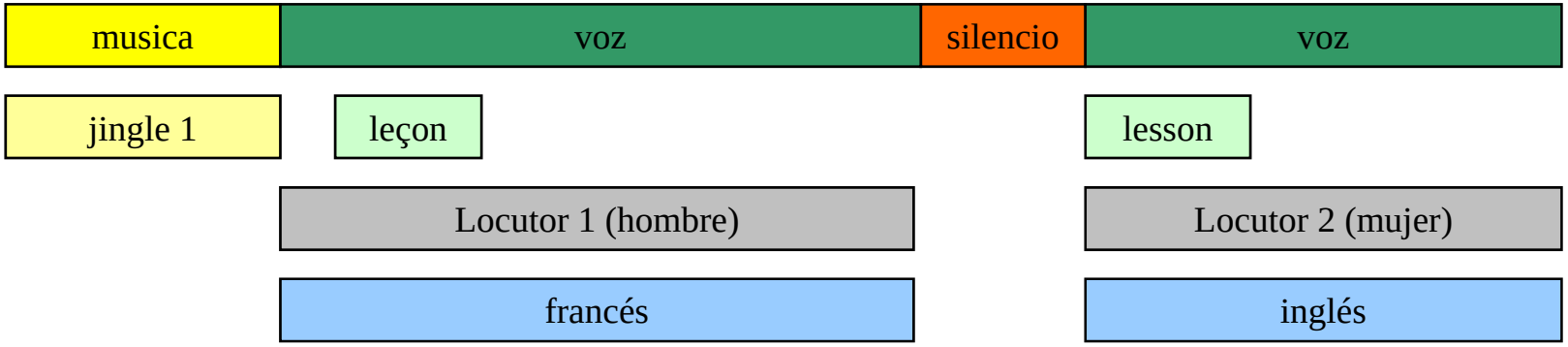


Indexado de audio



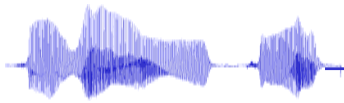
Indexado de audio





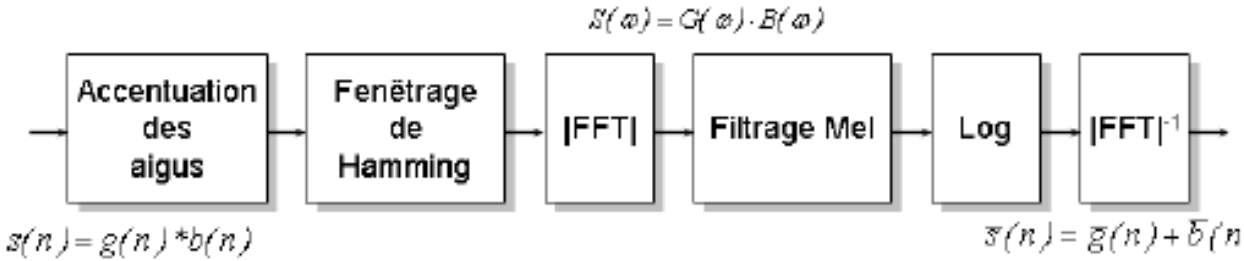
Parametrización

- El muestreo de una señal de audio produce una secuencia de valores discretos en tiempo y magnitud
  - Ej. La digitalización a 16 kHz produce 16000 muestras por segundo con valores comprendidos entre [-32768, +32767]
- Parametrización: una función asigna un valor a una serie de observaciones, extrayendo así la información pertinente y reduciendo el tiempo de cálculo



Parametrización cepstral

- Separación fuente-filtro del modelo de producción acústica

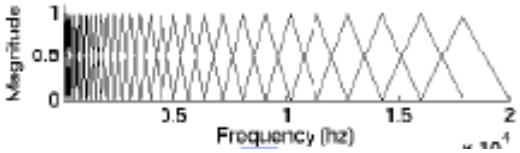


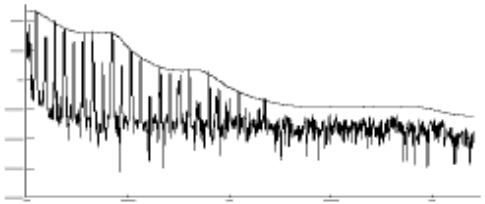
avec :

s(n), le signal acoustique

g(n), le source (cordes vocales)

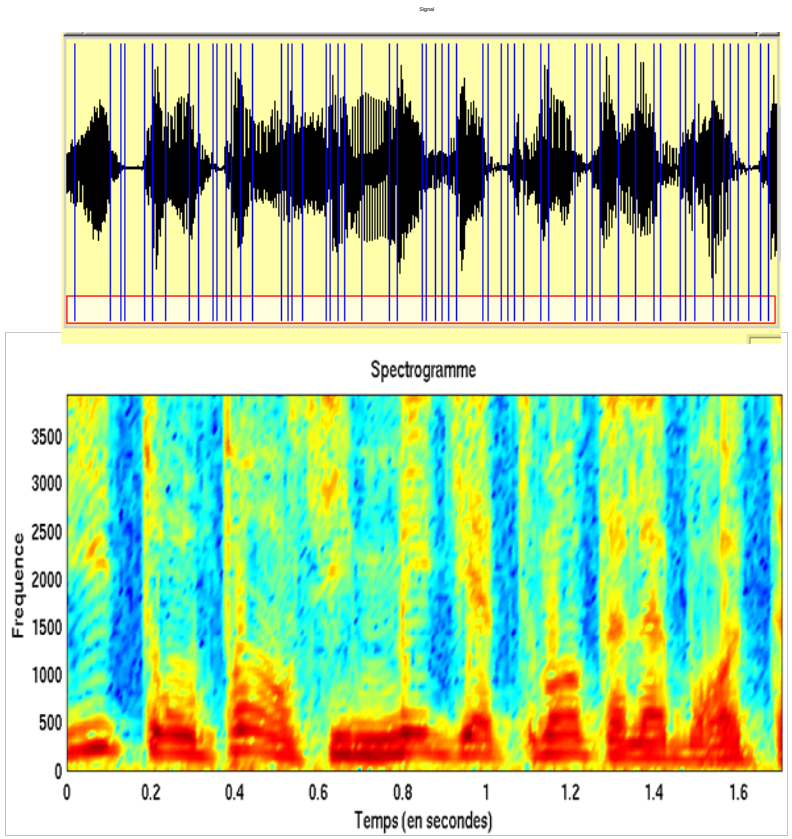
b(n), le filtre (conduit vocal)



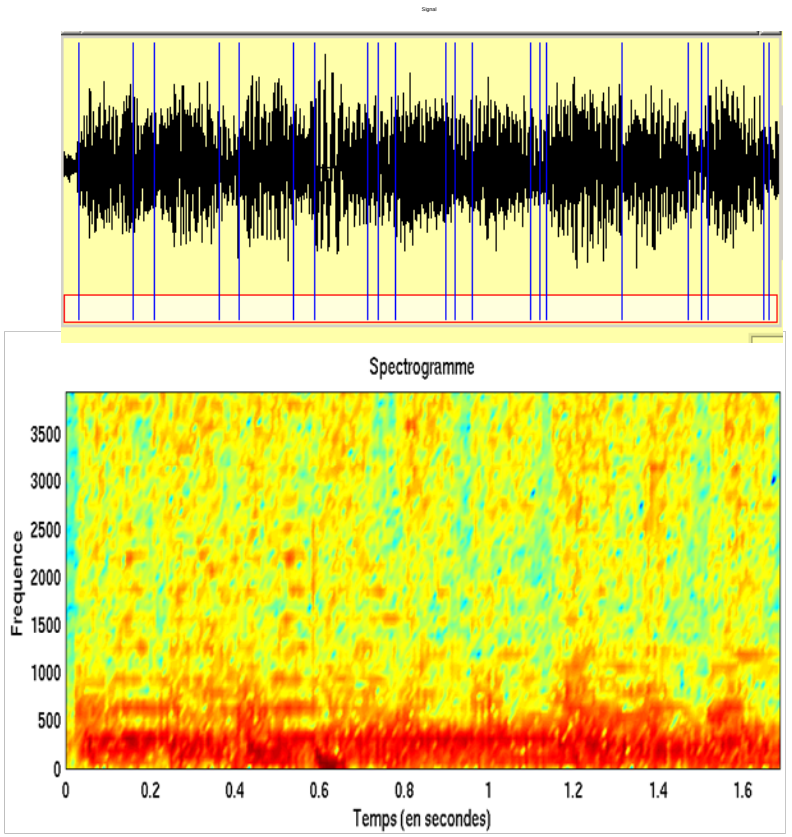


Detección de voz/música

Voz

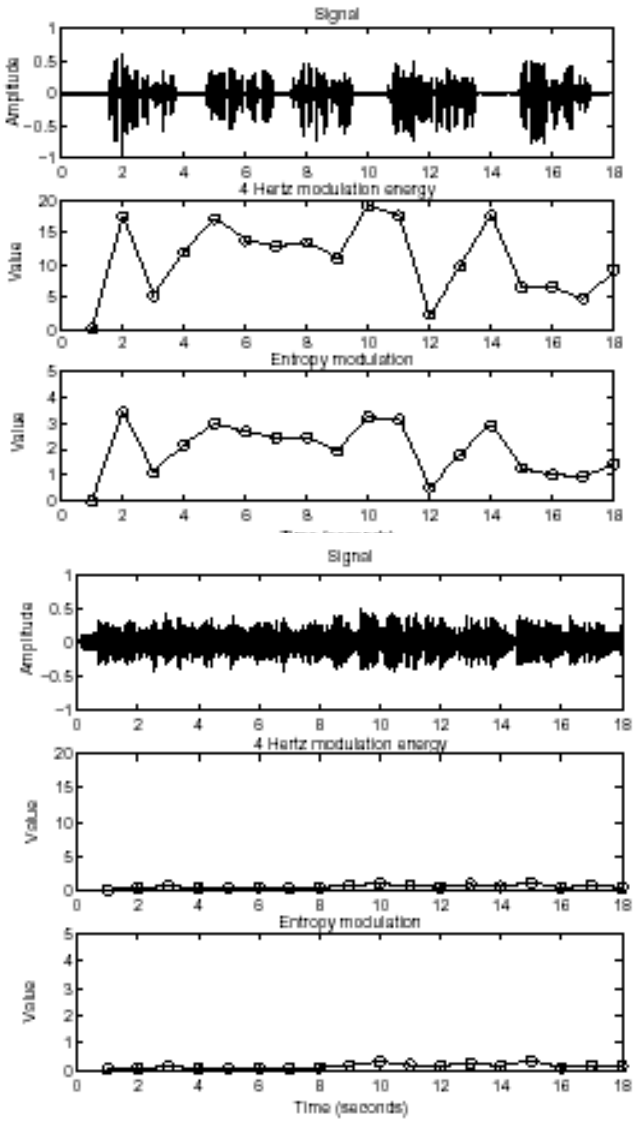


Música



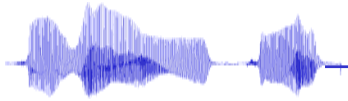
Modulación de la energía / Entropía

- Modulación de la energía a 4 Hz
  - Enventanado (16ms)
  - 40 coeficientes espectrales (Mel)
  - Filtrado (RIF pasa-banda 4hz)
  - Suma y normalización
  - Modulación (varianza durante 1s)
- Modulación de la entropía
  - Enventanado (16ms)
  - Histograma (amplitud de la señal)
  - Entropía
  - Modulación (varianza durante 1s)

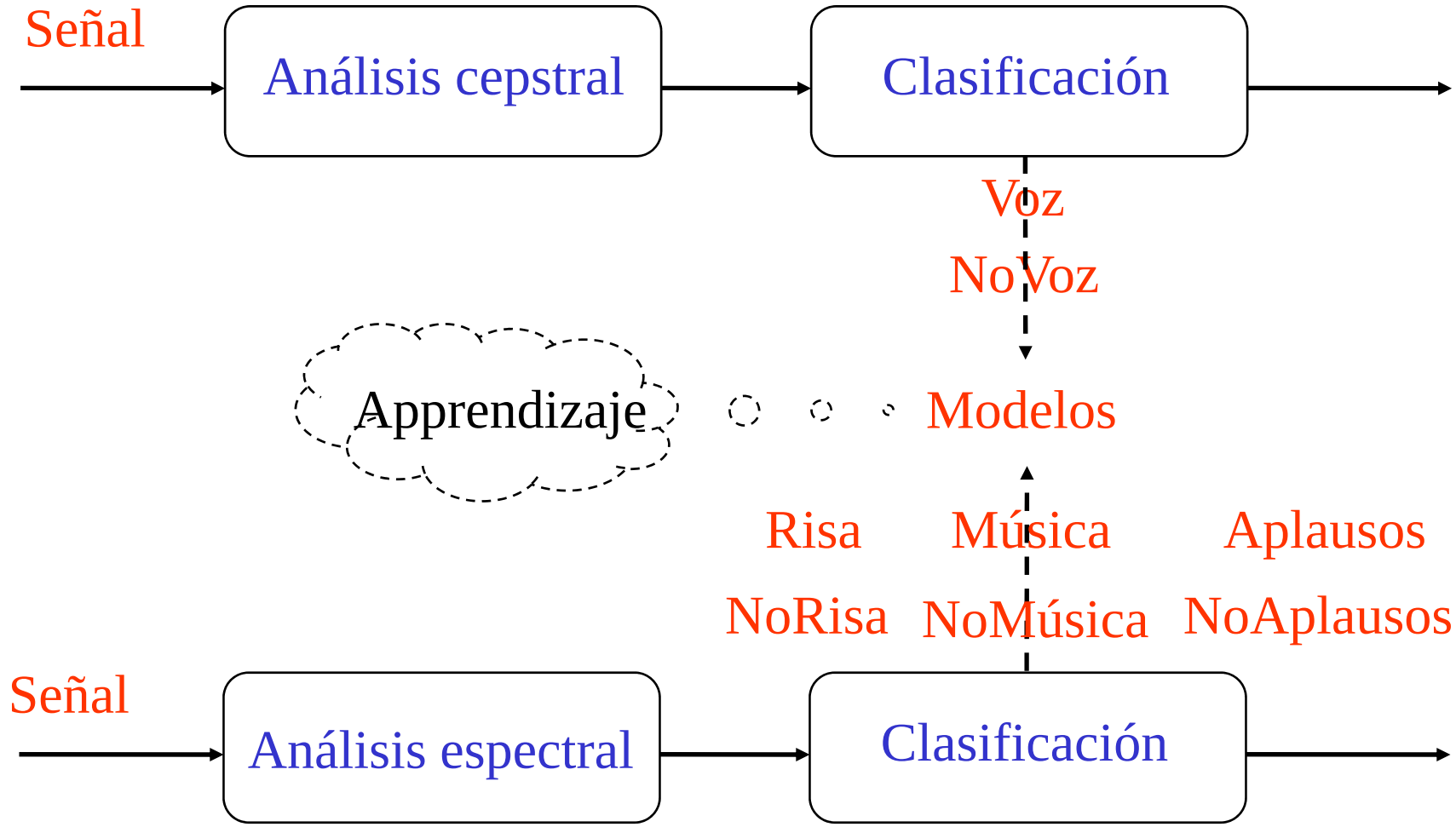


Problema práctico

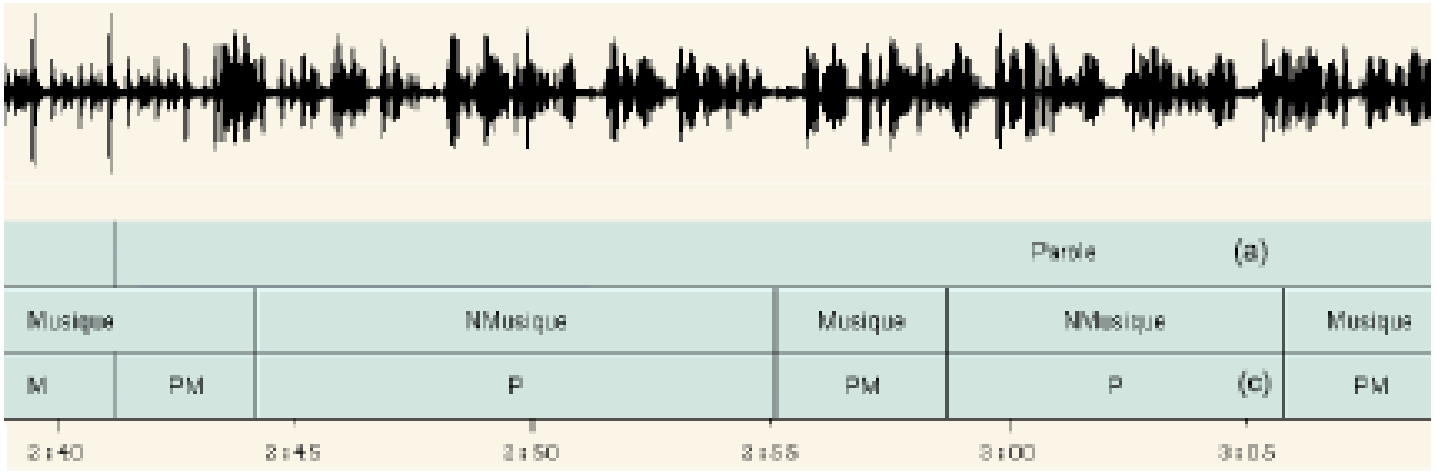
- Detección de algunos componentes básicos del audio en una emisión de « variedades »:
  - Voz
  - Música
  - Aplausos
  - Risas
- Corpus de 6 horas

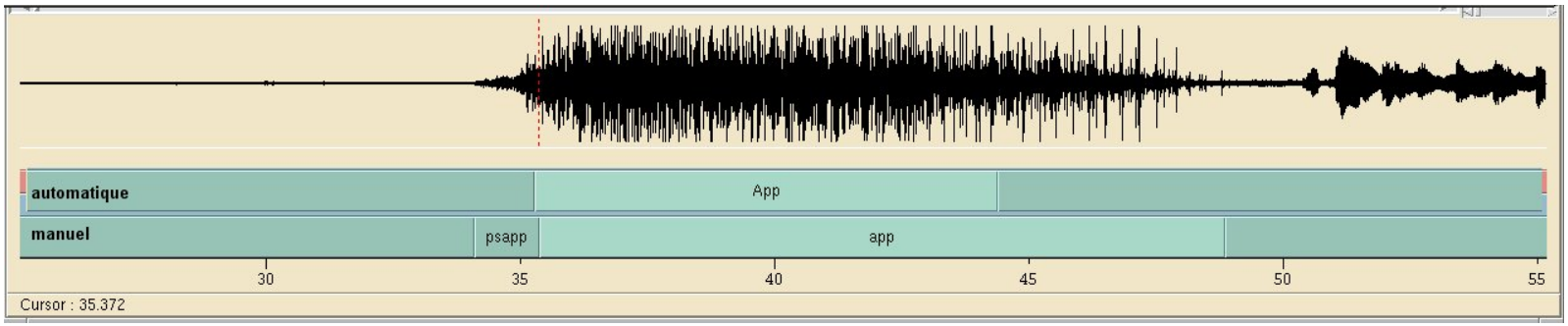


Detección de componentes: principio



Solución


Parole
(a)
Musique
NMusique
Musique
NMusique
Musique
M
PM
P
PM
P
(c)
PM
2:40
2:45
2:50
2:55
3:00
3:05


automatique
App
manuel
psapp
app
30
35
40
45
50
55
Cursor : 35.372

Solución

- Detección voz, música
  - Con GMM, se necesitaron al menos 200,000 vectores (~2 horas de audio) de cada clase para alcanzar buenos scores)
  - Cuando el número de vectores por clase es mas de 30,000, SVM se vuelve impráctico en tiempo de ejecución
    - Opción 1. Tomar un vector de cada 150
    - Opción 2. VQ
- Detección de risas y aplausos
  - En 3 horas de corpus, se obtuvieron 4 mins de aplausos y 1 min de risas

Resultados

Table 1. Classification results.

System	Music	Speech	Applause	Laughter
GMM	97.63 %	98.93 %	98.58 %	97.26 %
SVM	97.53 %	96.4 %	98.35 %	97.12 %

Table 2. Number of vectors used for training.

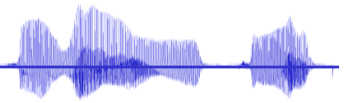
System	Music	Speech	Applause	Laughter
GMM				
class	429826	769501	7512	1714
nonclass	429826	769501	337160	342874
SVM				
class	1024	11333	7512	1714
non class	1024	5130	20000	20000

Table 3. Results for applause indexing using the same training vectors.

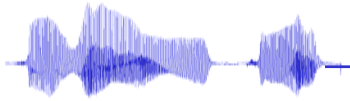
System	Class vectors	Score	CPU[1] time - training	CPU time - classif.
GMM	7500	98.58%	2' 55''	3' 25''
GMM	2500	95.04%	57''	3' 03''
SVM	7500	98.35%	6' 28''	22' 23''
SVM	2500	97.56%	2' 29''	12' 15''

[1] Processor Pentium IV 2.5GHz, 1.5 Gb of RAM

Temario



- Introducción
- Datos multivariables
- Modelos generativos
- Modelos discriminantes
- Ejemplo práctico
- Análisis teórico



Clasificación con modelos gaussianos

- Considere el caso de dos clases. La probabilidad posterior de la clase C_1 puede ser escrita:

$$p(C_1 \mid x) = \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_1)p(C_1) + p(x \mid C_2)p(C_2)}$$

$$p(C_1 \mid x) = \frac{1}{1+\exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_2)p(C_2)}$$

Clasificación con modelos gaussianos

- Si las distribuciones condicionales son gaussianas, y considerando que las matrices de covarianza son comunes:

$$p(C_1 \mid x) = \sigma(w^T x + w_0)$$

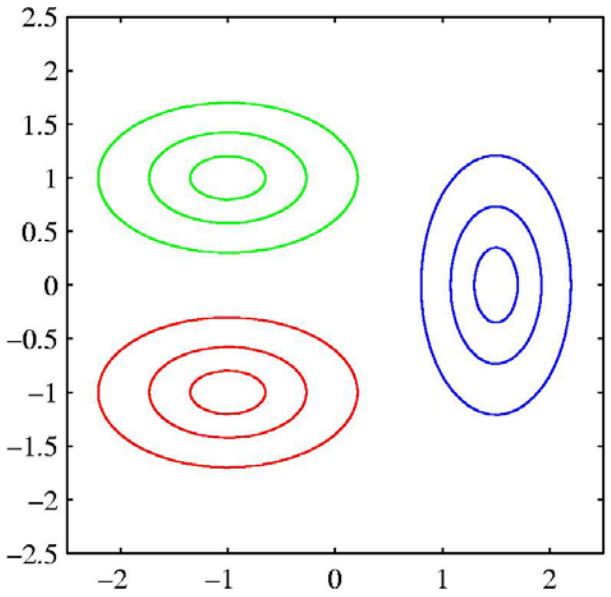
$$w = \Sigma^{-1}(\mu_1 - \mu_2)$$

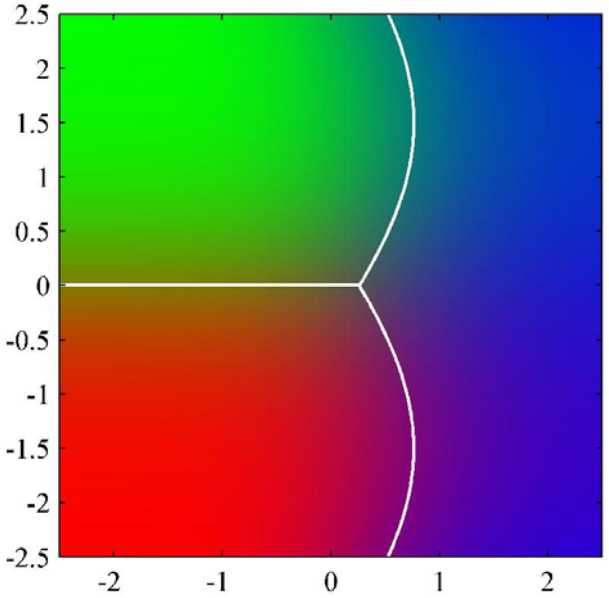
$$w_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

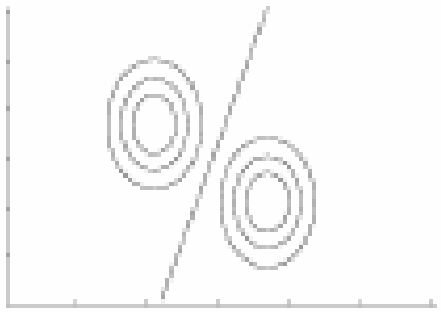
$$p(C_2 \mid x) = 1 - p(C_1 \mid x)$$

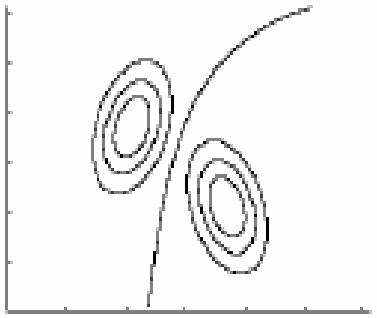
- Esto da como resultado una función lineal de x como argumento de la función logística

Clasificación con modelos gaussianos









Clasificación con regresión logística

- En la discriminación logística no se modelan distribuciones condicionales con respecto a las clases, sino su cociente. Supongamos dos clases:

$$\ln \frac{p(x \mid C_1)}{p(x \mid C_2)} = w^T x + w_0$$

$$p(C_1 \mid x) = \frac{1}{1 + \exp(-w^T x - w_0)}$$

$$p(C_2 \mid x) = 1 - p(C_1 \mid x)$$

Clasificación con regresión logística

- Con dos clases, la función discriminante es un proceso de Bernoulli:

$$C^t = 1 \; si \; x \in C_1, C^t = 0 \; si \; x \in C_2$$

$$l(w, w_0 \mid X) = \prod_t p(C_1 \mid x^t)^{C^t} \cdot (1 - p(C_1 \mid x^t))^{1-C^t}$$

$$E(w, w_0 \mid X) = -\ln l$$

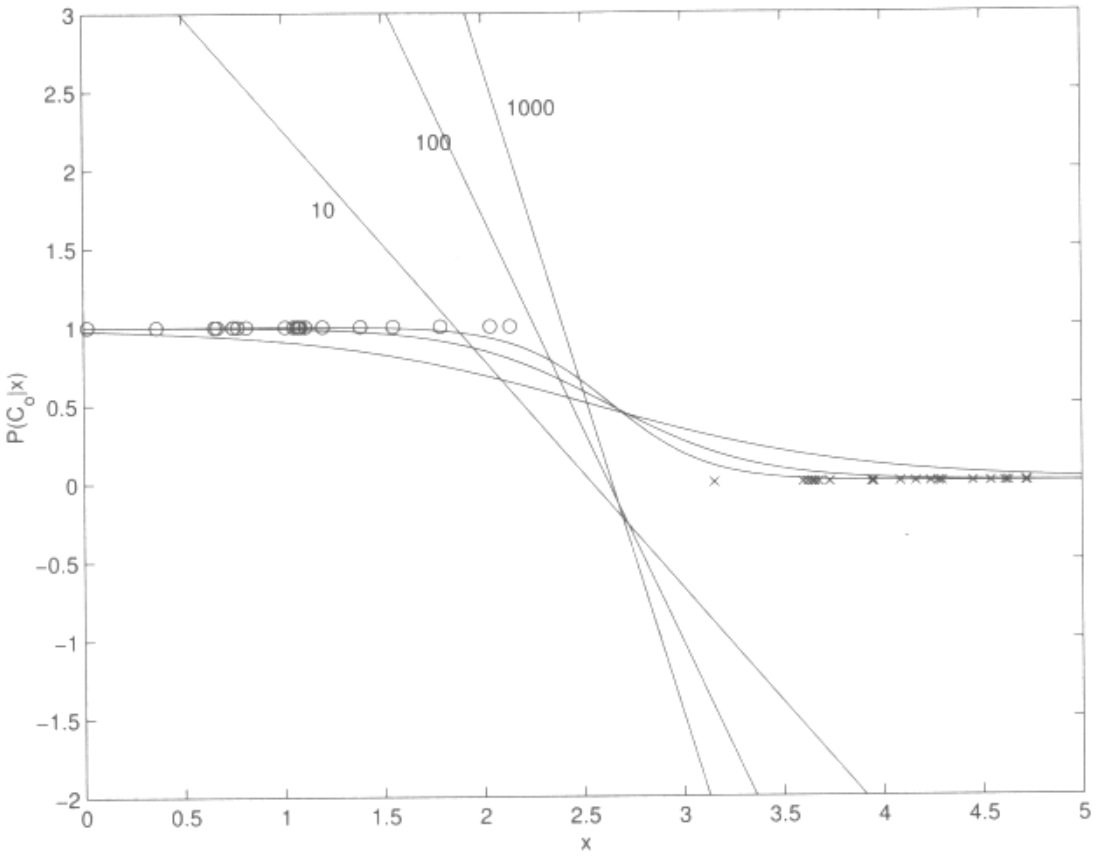
$$E(w, w_0 \mid X) = -\sum_t C^t \ln p(C_1 \mid x^t) + (1 - C^t) \ln(1 - p(C_1 \mid x^t))$$

$$w^* = \arg\min_w E(w \mid X)$$

$$\frac{\partial E}{\partial w_j} = \sum_t (C^t - p(C_1 \mid x^t)) x_j^t$$

$$\Delta w_j \leftarrow \Delta w_j + \eta \frac{\partial E}{\partial w_j}$$

Clasificación con regresión logística



En teoría…

- Asintóticamente, el error de la regresión logística es menor que el de Bayes « naive »

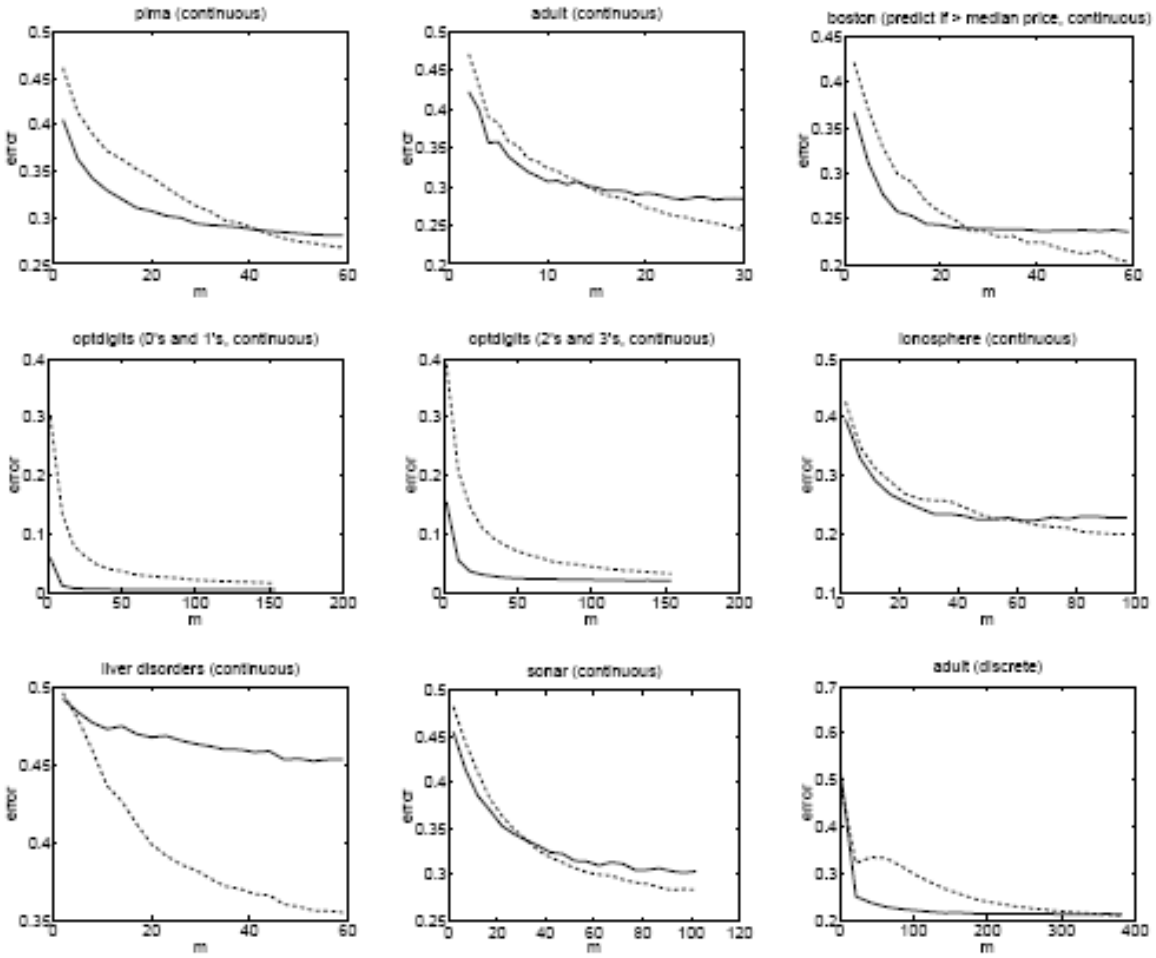
$$\varepsilon(h_{Dis,\infty}) \leq \varepsilon(h_{Gen,\infty})$$

- El error del clasificador de Bayes converge asintóticamente con un número logarítmico de ejemplos, mientras que el de regresión logística lo hace de manera lineal

$$\varepsilon(h_{Gen}) \leq \varepsilon(h_{Gen,\infty}) + G\left(O\left(\sqrt{\frac{1}{d}\log n}\right)\right)$$

$$\varepsilon(h_{Dis}) \leq \varepsilon(h_{Dis,\infty}) + O\left(\sqrt{\frac{d}{n}\log\frac{n}{d}}\right)$$

En teoría…


pima (continuous)
adult (continuous)
boston (predict if > median price, continuous)
optdigits (0's and 1's, continuous)
optdigits (2's and 3's, continuous)
ionosphere (continuous)
liver disorders (continuous)
sonar (continuous)
adult (discrete)
error
m

- Preguntas?