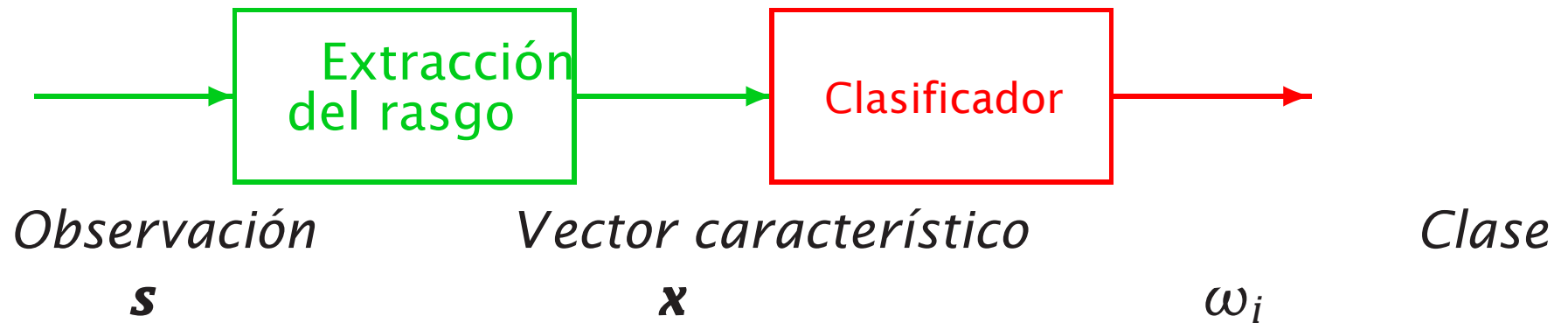


Clasificación de patrones

- Introducción
- Clasificadores paramétricos
- Clasificadores semiparamétricos
- Reducción de dimensionalidad
- Pruebas de relevancia

Clasificación de patrones

Objetivo : Clasificar objetos (o patrones) en categorías (o clases)



Tipos de problemas:

1. *Supervisados*: Las clases se conocen de antemano, y las muestras de datos de cada clase están disponibles.
2. *No supervisado*: Las clases (y /o número de clases) no son conocidas de antemano, y deben deducirse a partir de los datos.

Conceptos básicos de probabilidad

- Función de masa de probabilidad discreta (PMF): $P(\omega_i)$

$$\sum_i P(\omega_i) = 1$$

- Función de densidad de probabilidad continua (PDF): $p(x)$

$$\int p(x) dx = 1$$

- Valor esperado: $E(x)$

$$E(x) = \int xp(x) dx$$

Distancia de Kullback-Liebler

- Puede utilizarse para calcular una distancia entre distribuciones de masa de probabilidad $P(z_i)$, y $Q(z_i)$

$$D(P \parallel Q) = \sum_i P(z_i) \log \frac{P(z_i)}{Q(z_i)} \geq 0$$

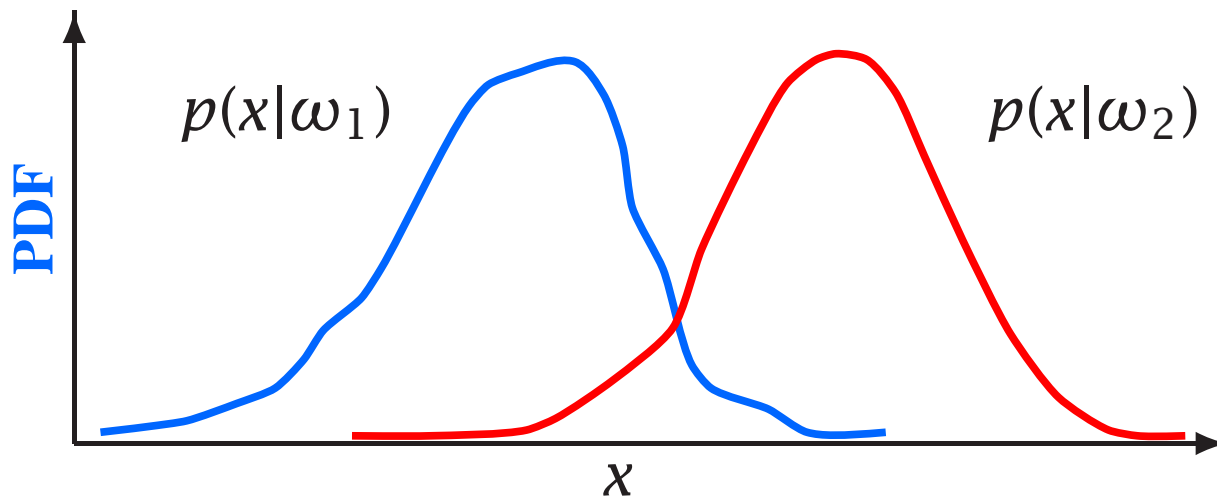
- Utiliza el logaritmo de desigualdad $\log x \leq x - 1$

$$\sum_i P(z_i) \log \frac{Q(z_i)}{P(z_i)} \leq \sum_i P(z_i) \left(\frac{Q(z_i)}{P(z_i)} - 1 \right) = \sum_i Q(z_i) - P(z_i) = 0$$

- Conocida como *entropía relativa* en teoría de la información
- La *divergencia* de $P(z_i)$ y $Q(z_i)$ es la suma asimétrica

$$D(P \parallel Q) + D(Q \parallel P)$$

Teorema de Bayes



Definir:

- $\{\omega_i\}$ un conjunto de clases M mutuamente exclusivas
- $P(\omega_i)$ una probabilidad **a priori** para la clase ω_i
- $p(\mathbf{x}|\omega_i)$ PDF para el vector característico \mathbf{x} de la clase ω_i
- $P(\omega_i|\mathbf{x})$ una probabilidad **a posteriori** de ω_i dado \mathbf{x}

De la regla de Bayes:
$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

donde
$$p(\mathbf{x}) = \sum_{i=1}^M p(\mathbf{x}|\omega_i)P(\omega_i)$$

Teoría de decisión de Bayes

- La probabilidad de cometer un error, dado \mathbf{x} es:

$$P(\text{error}|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x}) \quad \text{si se determina la clase } \omega_i$$

- Para minimizar $P(\text{error}|\mathbf{x})$ (y $P(\text{error})$):

$$\text{Elija } \omega_i \text{ si } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \quad \forall j \neq i$$

- Para un problema de clase dos, esta regla de decisión significa:

$$\text{Elegir } \omega_1 \text{ si } \frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x})} > \frac{p(\mathbf{x}|\omega_2)P(\omega_2)}{p(\mathbf{x})}; \text{ de lo contrario } \omega_2$$

- Esta regla se puede expresar como un cociente de probabilidad:

$$\text{Elegir } \omega_1 \text{ si } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}; \text{ si no elegir } \omega_2$$

Riesgo de Bayes

- Definir la función de coste λ_{ij} y el riesgo condicional $R(\omega_i|\mathbf{x})$:
 - λ_{ij} es el coste de clasificar \mathbf{x} como ω_i cuando éste es realmente ω_j
 - $R(\omega_i|\mathbf{x})$ es el riesgo de clasificar \mathbf{x} como clase ω_i

$$R(\omega_i|\mathbf{x}) = \sum_{j=1}^M \lambda_{ij} P(\omega_j|\mathbf{x})$$

- **El riesgo de Bayes** es el riesgo mínimo que se puede obtener:
Elegir ω_i si $R(\omega_i|\mathbf{x}) < R(\omega_j|\mathbf{x}) \quad \forall j \neq i$
- El riesgo de Bayes corresponde al mínimo $P(error|\mathbf{x})$ cuando:
 - Todos los errores poseen igual coste $(\lambda_{ij} = 1, \quad i \neq j)$
 - No existe un coste para ser correcto $\lambda_{ii} (\neq 0)$

$$R(\omega_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

Funciones discriminantes

- Formulación alternativa de la regla de decisión de Bayes
- Definir una función discriminante, $g_i(\mathbf{x})$, para cada clase ω_i

Elegir ω_i si $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$

- Las funciones revelan resultados de clasificación idénticos:

$$\begin{aligned} g_i(\mathbf{x}) &= P(\omega_i|\mathbf{x}) \\ &= p(\mathbf{x}|\omega_i)P(\omega_i) \\ &= \log p(\mathbf{x}|\omega_i) + \log P(\omega_i) \end{aligned}$$

- La elección de la función afecta a los costes de computación.
- Las funciones discriminantes dividen el espacio característico en **regiones de decisión**, separadas por **límites de decisión**.

Estimación de la densidad

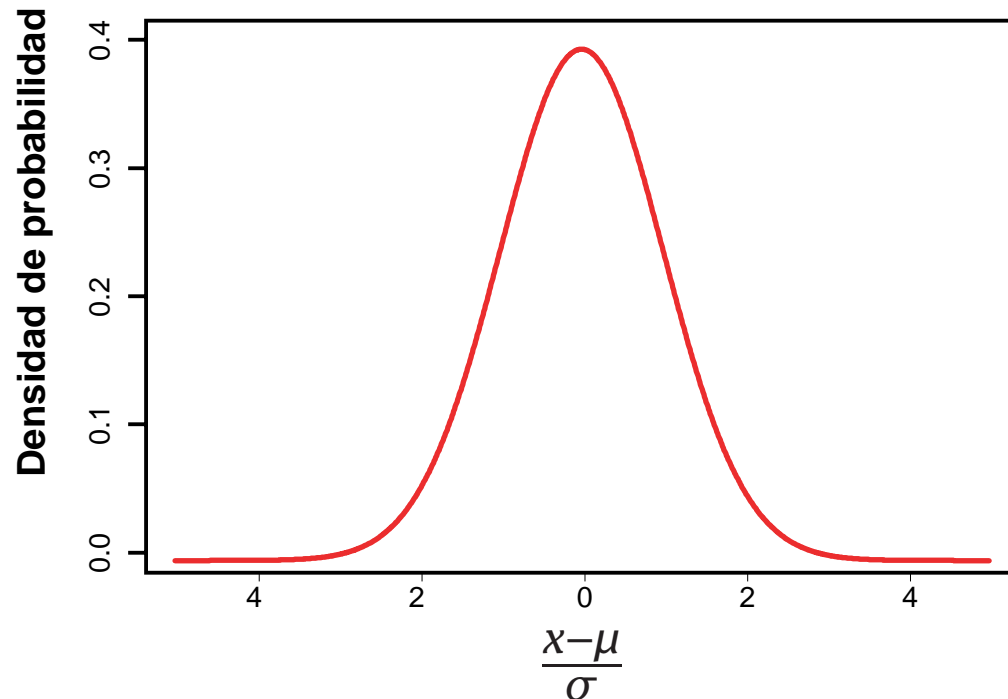
- Se utiliza para calcular la función PDF $p(\mathbf{x}|w_i)$ subyacente.
- Métodos **paramétricos**:
 - Supone una forma funcional específica para la PDF
 - Optimiza parámetros de la función PDF para ajustar datos
- Métodos **no-paramétricos**:
 - Determinan la forma de la PDF a partir de los datos
 - Aumenta el tamaño del conjunto de parámetros con la cantidad de datos
- Métodos **semi-paramétricos**:
 - Utilizan una clase general de formas funcionales para la PDF
 - Pueden cambiar el conjunto de parámetros con independencia de los datos
 - Utilizan métodos no supervisados para calcular parámetros

Clasificadores paramétricos

- Distribuciones gaussianas
- Estimación del parámetro de máxima probabilidad (ML)
- Gaussianas multivariadas
- Clasificadores gaussianos

Distribuciones gaussianas

- Los PDF gaussianos son razonables cuando un vector característico se considera una perturbación alrededor de una referencia



- Procedimientos de estimación simples para parámetros modelo
- La clasificación se reduce con frecuencia a una distancia métrica simple
- Las distribuciones gaussianas se conocen también como *Normales*

Distribuciones gaussianas: Una dimensión

- Un PDF gaussiano unidimensional se puede expresar como:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \sim N(\mu, \sigma^2)$$

- La función PDF está centrado alrededor de la media

$$\mu = E(x) = \int xp(x)dx$$

- La *propagación* de la función PDF está determinada por la varianza.

$$\sigma^2 = E((x-\mu)^2) = \int (x-\mu)^2 p(x)dx$$

Estimación del parámetro de probabilidad máxima

- La estimación del parámetro de probabilidad máxima determina un cálculo aproximado $\hat{\theta}$ para el parámetro θ al potenciar al máximo la **probabilidad** $L(\theta)$ de la observación de datos $\mathcal{X} = \{x_1, \dots, x_n\}$

$$\hat{\theta} = \arg \max_{\theta} L(\theta)$$

- Se asumen datos **independientes, distribuidos de forma idéntica**

$$L(\theta) = p(\mathcal{X}|\theta) = p(x_1, \dots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta)$$

- Las soluciones del parámetro de ML pueden obtenerse a través de la derivada

$$\frac{\partial}{\partial \theta} L(\theta) = 0$$

- Para las distribuciones gaussianas, el $\log L(\theta)$ es más fácil de resolver

Estimación de ML gaussiana: Una dimensión

- El cálculo de probabilidad máxima para μ viene dado por:

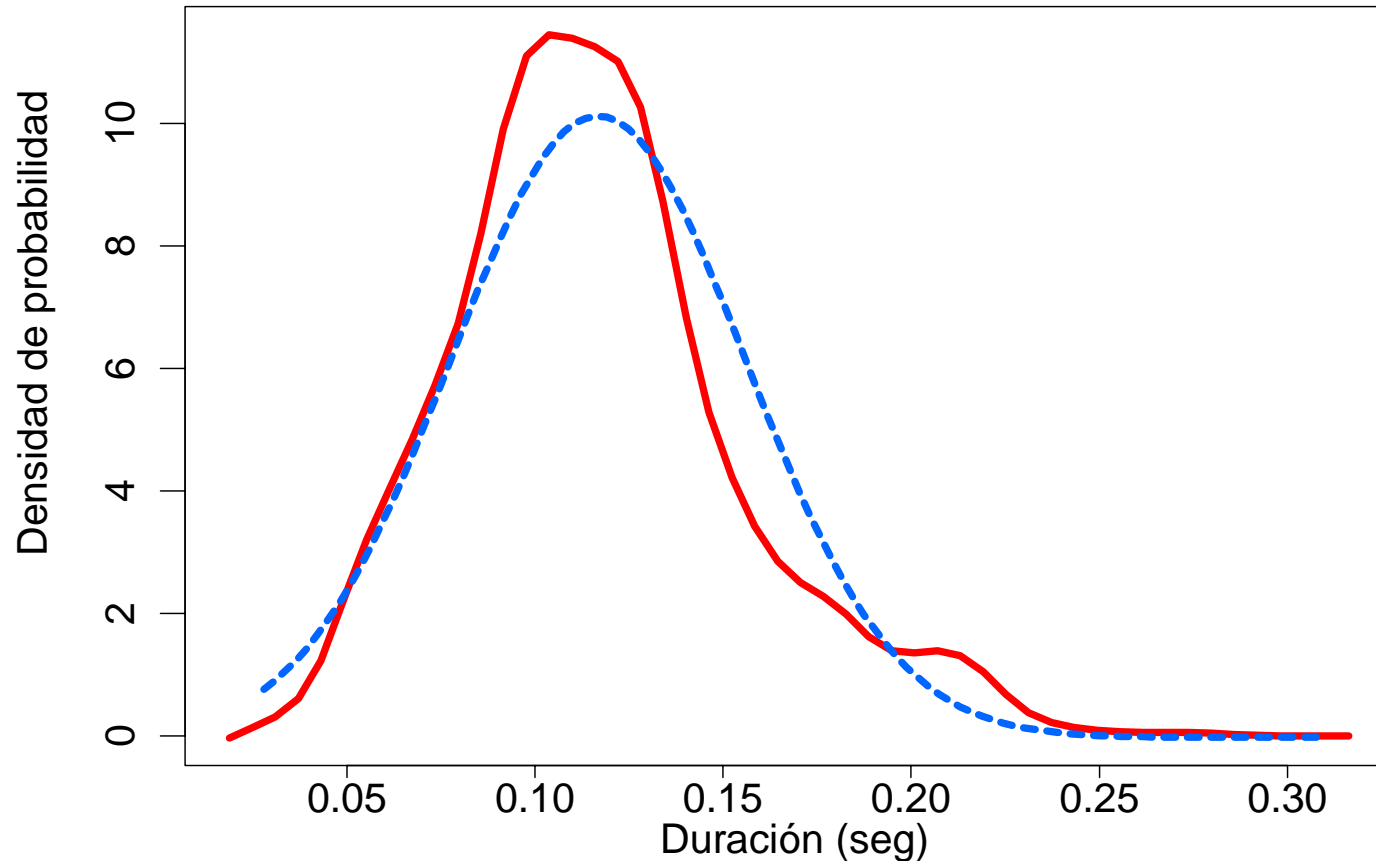
$$L(\mu) = \prod_{i=1}^n p(x_i|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$
$$\log L(\mu) = -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 - n \log \sqrt{2\pi}\sigma$$
$$\frac{\partial \log L(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_i (x_i - \mu) = 0$$
$$\hat{\mu} = \frac{1}{n} \sum_i x_i$$

- El cálculo de probabilidad máxima para σ viene dado por:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2$$

Estimación de ML gaussiana: Una dimensión

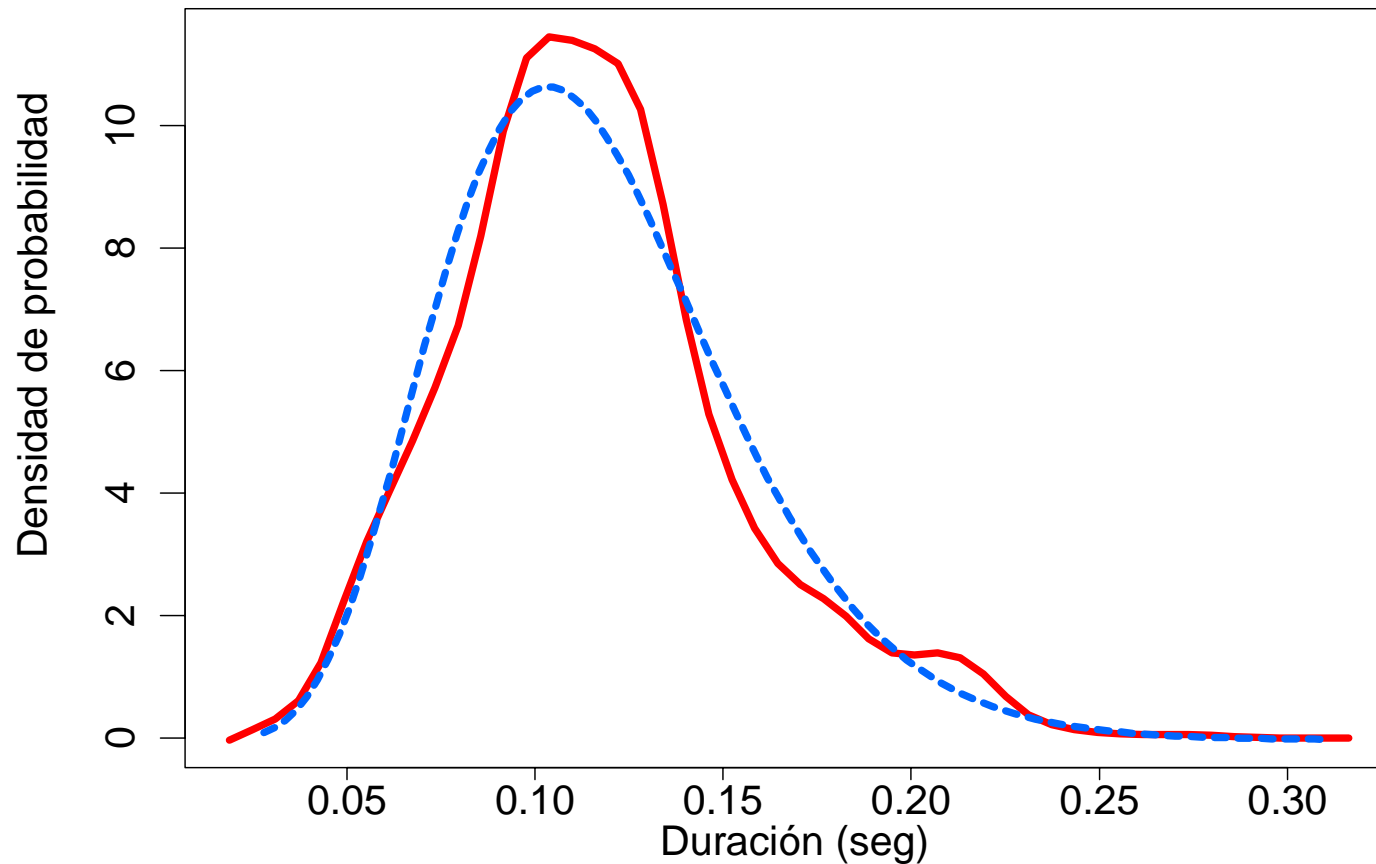
[s] Duración (1 000 enunciados, 100 hablantes)



$(\hat{\mu} \approx 120 \text{ ms}, \hat{\sigma} \approx 40 \text{ ms})$

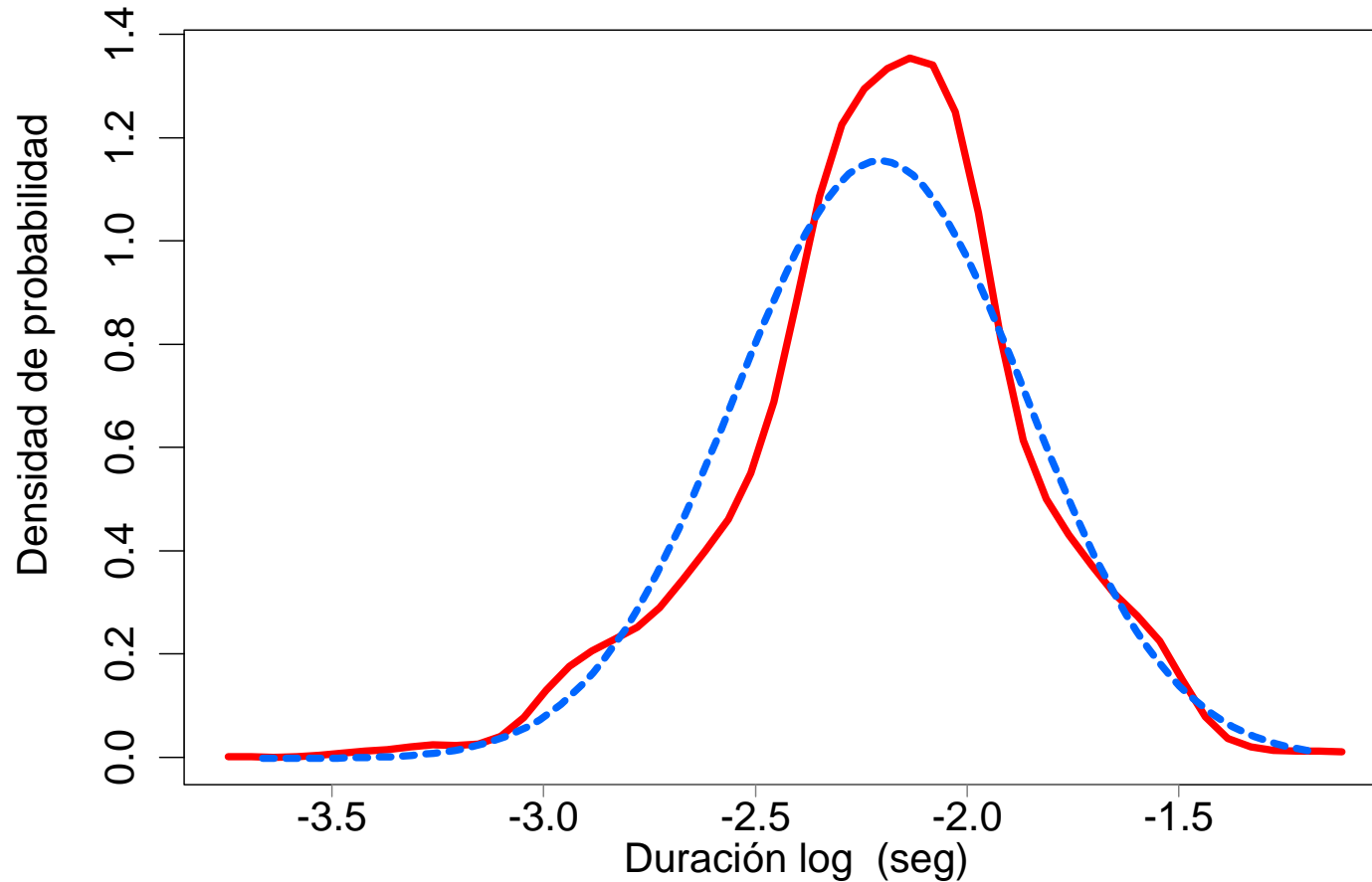
Estimación de ML: Distribuciones alternativas

[s] Duración: Distribución Gamma



Estimación de ML: Distribuciones alternativas

Duración [s] Log : Distribución normal



Distribuciones gaussianas: Dimensiones múltiples

- Un PDF gaussiano multidimensional se puede expresar como:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} \sim N(\boldsymbol{\mu}, \Sigma)$$

- d es el número de dimensiones
- $\mathbf{x} = \{x_1, \dots, x_d\}$ es el vector de entrada
- $\boldsymbol{\mu} = E(\mathbf{x}) = \{\mu_1, \dots, \mu_d\}$ es el vector de medias
- $\Sigma = E((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t)$ es la matriz de covarianza con elementos σ_{ij} , inverso Σ^{-1} , y determinante $|\Sigma|$
- $\sigma_{ij} = \sigma_{ji} = E((x_i - \mu_i)(x_j - \mu_j)) = E(x_i x_j) - \mu_i \mu_j$

Distribuciones gaussianas: Propiedades multidimensionales

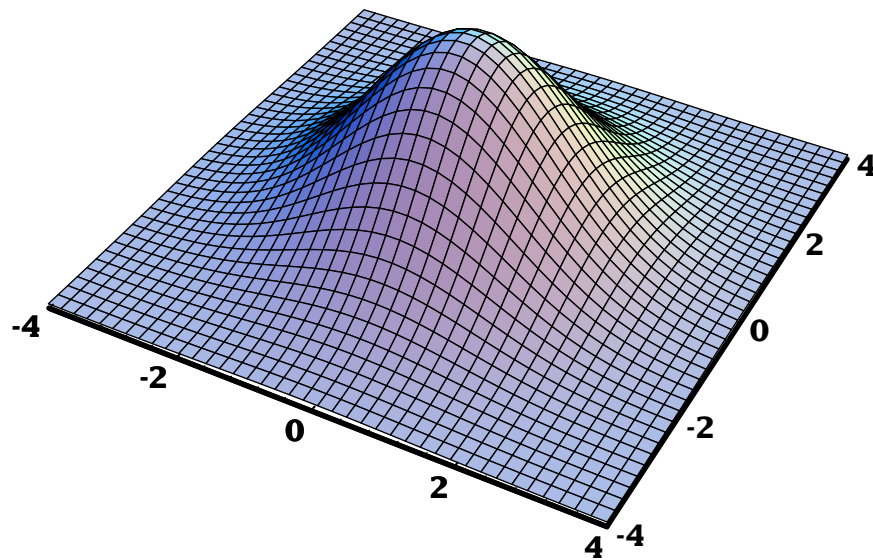
- Si las dimensiones i th y j th son estadística o linealmente independientes, entonces $E(x_i x_j) = E(x_i)E(x_j)$ y $\sigma_{ij} = 0$
- Si todas las dimensiones son estadística o linealmente independientes, $\sigma_{ij} = 0 \quad \forall i \neq j$ y Σ poseen elementos no cero sólo en la diagonal
- Si la densidad subyacente es gaussiana y Σ es una matriz diagonal, entonces las dimensiones son estadísticamente independientes y

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i) \quad p(x_i) \sim N(\mu_i, \sigma_{ii}) \quad \sigma_{ii} = \sigma_i^2$$

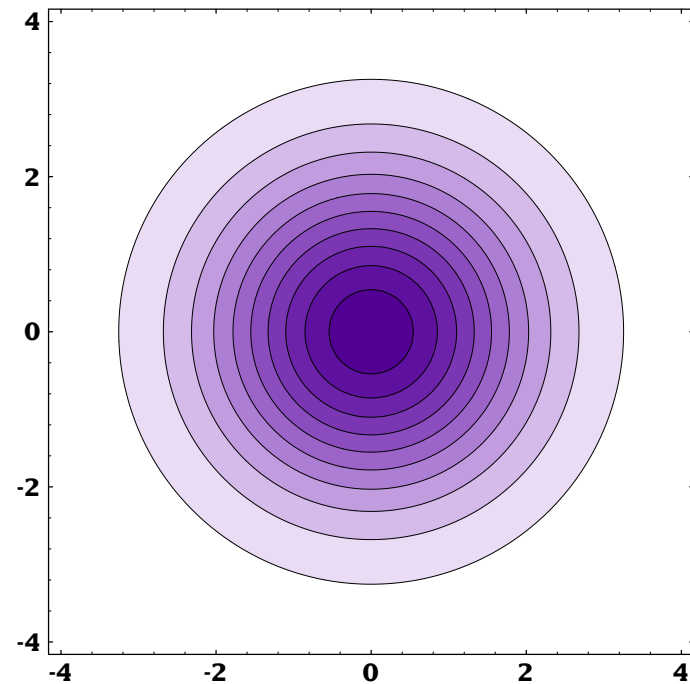
Matriz de covarianza diagonal: $\mathbf{S} = s^2\mathbf{I}$

$$\Sigma = \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix}$$

PDF 3-Dimensional



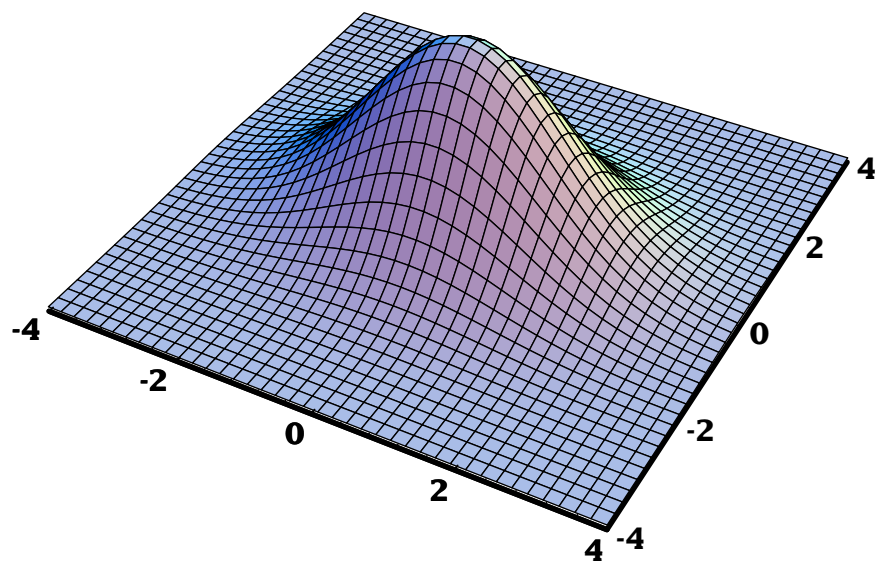
Contorno PDF



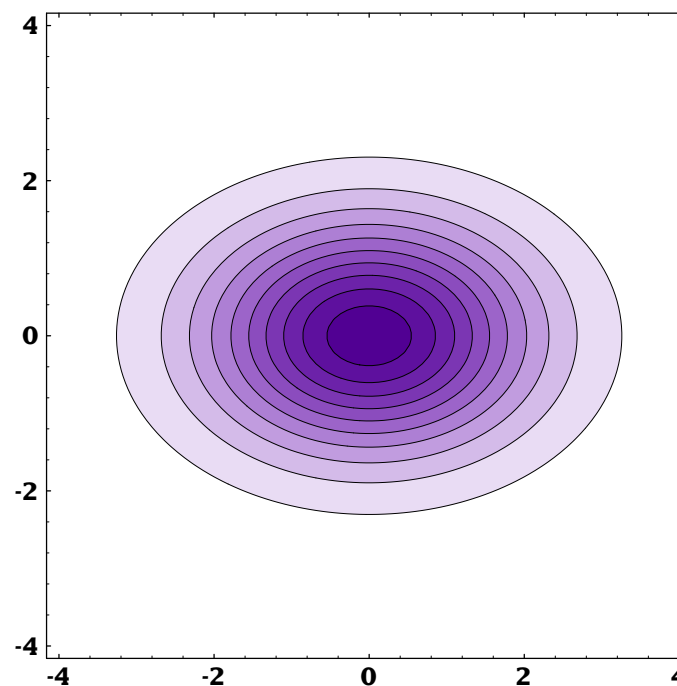
Matriz de covarianza diagonal: $s_{ij} = 0 \quad \forall i \neq j$

$$\Sigma = \begin{vmatrix} 2 & 0 \\ 0 & 1 \end{vmatrix}$$

PDF 3-Dimensional



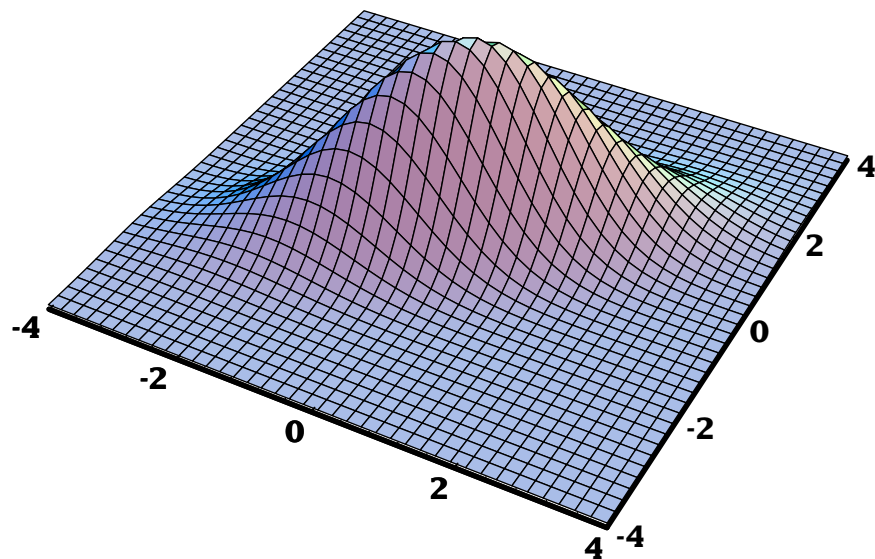
Contorno PDF



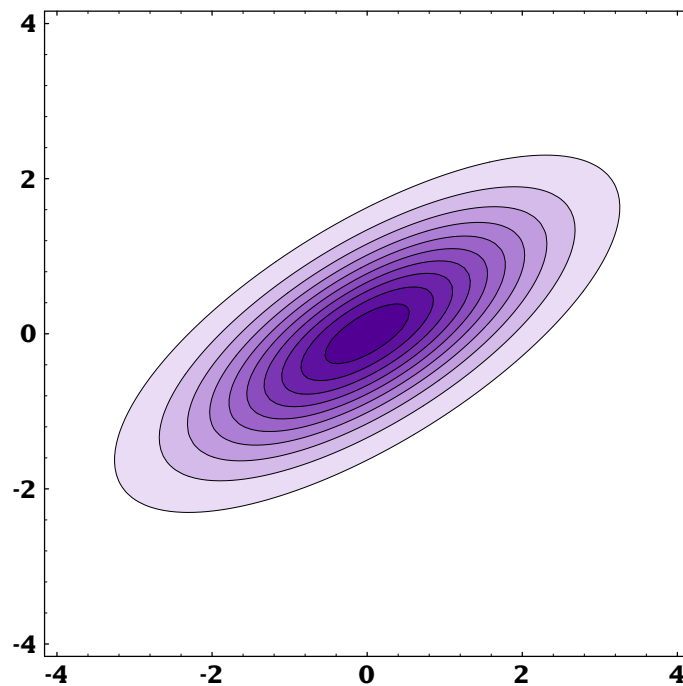
Matriz de covarianza general : $s_{ij} \neq 0$

$$\Sigma = \begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix}$$

PDF 3-Dimensional



Contorno PDF



Estimación de ML multivariada

- Las estimaciones de ML para los parámetros $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_l\}$ están determinados por la maximización conjunta $L(\boldsymbol{\theta})$ de un conjunto de datos i.i.d. $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

$$L(\boldsymbol{\theta}) = p(\mathcal{X}|\boldsymbol{\theta}) = p(\mathbf{x}_1, \dots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\theta})$$

- Para hallar $\hat{\boldsymbol{\theta}}$ resolvemos $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \mathbf{0}$, or $\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \mathbf{0}$

$$\nabla_{\boldsymbol{\theta}} = \left\{ \frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_l} \right\}$$

- Los cálculos de ML de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_i \mathbf{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t$$

Clasificador gaussiano multivariado

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Requiere un vector de medias $\boldsymbol{\mu}_i$, y una matriz de covarianza $\boldsymbol{\Sigma}_i$ para cada una de las clases $M \{\omega_1, \dots, \omega_M\}$
- Las funciones discriminantes de error mínimo son de la siguiente forma:

$$g_i(\mathbf{x}) = \log P(\omega_i|\mathbf{x}) = \log p(\mathbf{x}|\omega_i) + \log P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log P(\omega_i)$$

- La clasificación puede reducirse a una métrica de distancia simple para diversas situaciones

$$\Sigma_i = \sigma^2 \mathbf{I}$$

- Cada clase posee la misma estructura de covarianza : dimensiones estadísticamente independientes con varianza σ^2
- Las funciones discriminantes equivalentes son:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \log P(\omega_i)$$

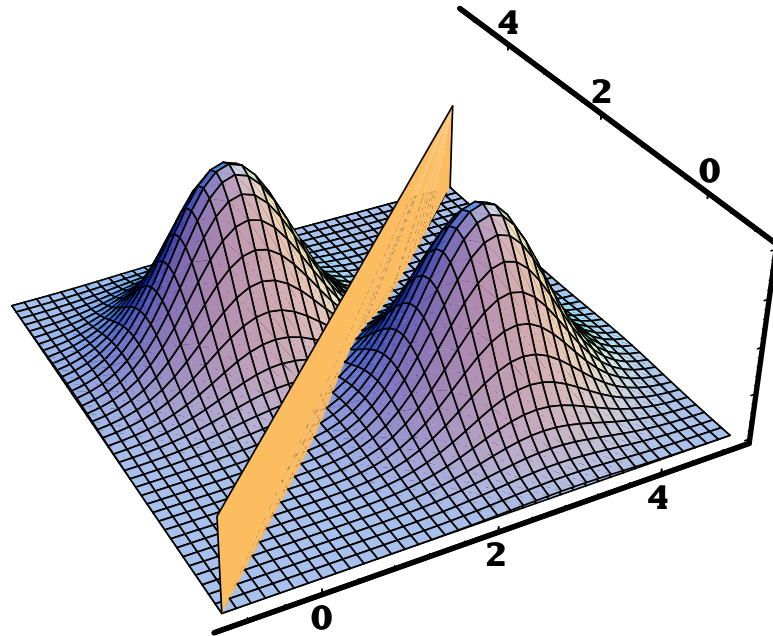
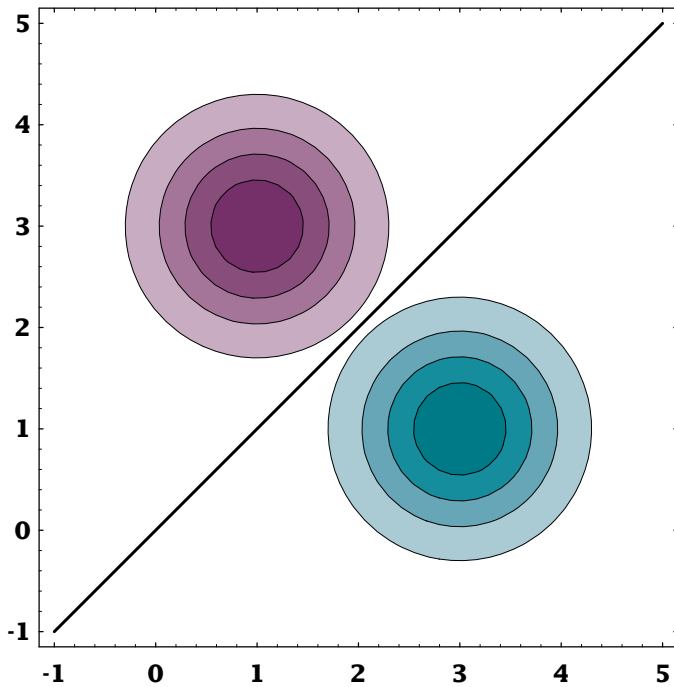
- Si cada clase es equitativamente probable, éste es un clasificador de **distancia mínima**, una forma de correlación.
- Las funciones discriminantes pueden sustituirse por la siguiente expresión **lineal**,

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + \omega_{i0}$$

donde $\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$ y $\omega_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \log P(\omega_i)$

Clasificador gaussiano: $\Sigma = \sigma^2 I$

En distribuciones con una estructura de covarianza común, las regiones de decisión son hiperplanos.



Clasificador gaussiano $\Sigma_i \neq \Sigma$

- Cada clase posee la misma estructura de covarianza Σ
- Las funciones discriminantes equivalentes son:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \log P(\omega_i)$$

- Si cada clase es igualmente probable, la regla de decisión del mínimo error es la distancia cuadrada de **Mahalanobis**.
- Las funciones discriminantes permanecen como expresiones lineares:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + \omega_{i0}$$

donde

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$$

$$\omega_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \log P(\omega_i)$$

Clasificador gaussiano: Σ_i arbitrario

- Cada clase posee una estructura de covarianza distinta Σ_i
- Las funciones discriminantes equivalentes son:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log P(\omega_i)$$

- Las funciones discriminantes son inherentemente **cuadráticas**:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + \omega_{i0}$$

donde

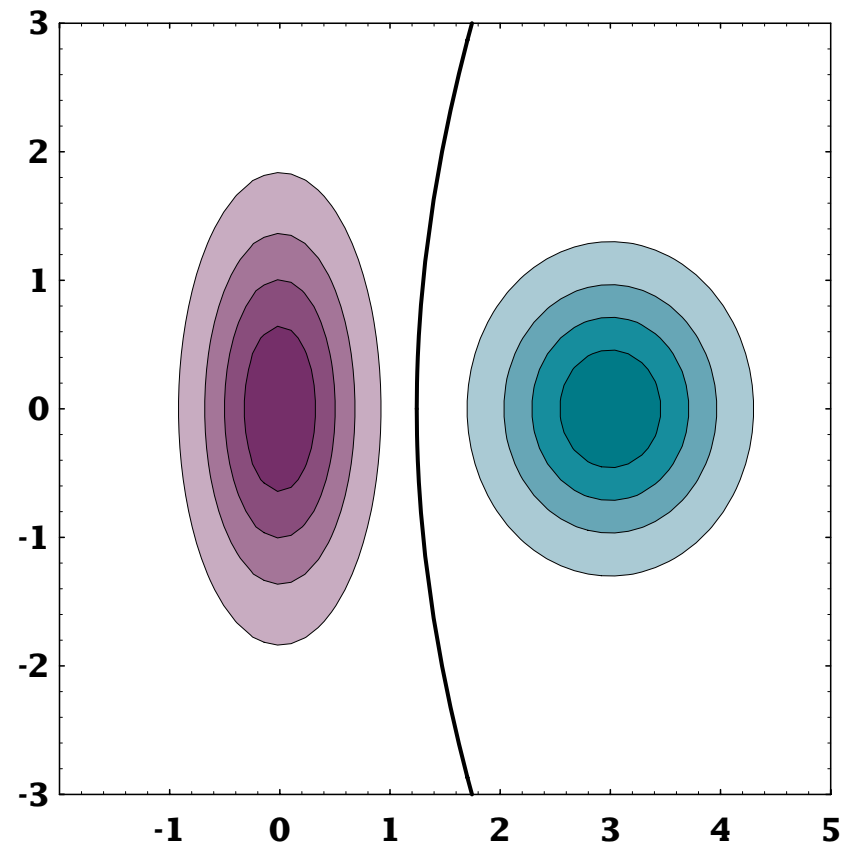
$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$$

$$\omega_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log P(\omega_i)$$

Clasificador gaussiano: Σ_i arbitrario

En distribuciones con estructuras de covarianza arbitrarias, las regiones de decisión están definidas por hiperesferas.



Clasificación de clase 3 (Atal & Rabiner, 1976)

- Distingue entre silencio, sonidos sordos y sonoros
- Utilizan 5 rasgos:
 - Cálculo de cruce por cero
 - Energía logarítmica
 - Primer coeficiente de autocorrelación normalizado
 - Primer coeficiente de predicción
 - Error de predicción normalizado
- Clasificador gaussiano multivariado, estimación de ML
- Decisión por la distancia cuadrada de Mahalanobis
- Entrenado con cuatro hablantes (2 oraciones por hablante), probado en 2 hablantes (1 oración por hablante)

Estimación del parámetro máximo a posteriori

- Los enfoques de estimación bayesiana asumen que la forma de la PDF $p(x|\theta)$ se conoce, pero el valor de θ no.
- El conocimiento de θ queda contenido en:
 - Un PDF inicial a *a priori* $p(\theta)$
 - Un conjunto de datos i.i.d. $\mathcal{X} = \{x_1, \dots, x_n\}$

- La función PDF deseada para x posee la forma de

$$p(x|\mathcal{X}) = \int p(x, \theta|\mathcal{X})d\theta = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

- El valor $\hat{\theta}$ que maximiza $p(\theta|\mathcal{X})$ se denomina el cálculo **máximo a posteriori** (MAP) de θ

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})} = \alpha \prod_{i=1}^n p(x_i|\theta)p(\theta)$$

Estimación gaussiana MAP: Una dimensión

- En una distribución gaussiana con media μ desconocida :

$$p(x|\mu) \sim N(\mu, \sigma^2) \quad p(\mu) \sim N(\mu_0, \sigma_0^2)$$

- Las estimaciones MAP de μ y x vienen dadas por:

$$p(\mu|\mathcal{X}) = \alpha \prod_{i=1}^n p(x_i|\mu)p(\mu) \sim N(\mu_n, \sigma_n^2)$$

$$p(x|\mathcal{X}) = \int p(x|\mu)p(\mu|\mathcal{X})d\mu \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

donde
$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- A medida de que n aumenta, $p(\mu|\mathcal{X})$ converge en μ , y $p(x|\mathcal{X})$ converge en el cálculo de ML $\sim N(\hat{\mu}, \sigma^2)$

MIT

Referencias

- Huang, Acero y Hon, *Spoken Language Processing*, Prentice-Hall, 2001.
- Duda, Hart y Stork, *Pattern Classification*, John Wiley & Sons, 2001.
- Atal y Rabiner, A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition, *IEEE Trans ASSP*, 24(3), 1976.