Curso: Reconocimiento de Patrones

Temas[*]:
(1) Selección y extracción de características

(2) Aprendizaje no supervisado

Alvaro Pardo
IMERL & IIE
Facultad de Ingeniería
Universidad de la República

Octubre 2003

1 Introducción

Idealmente, antes de diseñar un sistema de reconocimiento de patrones debemos estudiar el proceso de formación de estos patrones. En teoría, si entendemos las diferencias entre patrones de diferentes clases podremos representar cada una de las clases con un conjunto reducido de medidas que podrán ser usadas para el proceso de clasificación. En general, con muchas clases, este estudio no es factible. A lo sumo podremos arriesgar hipótesis respecto a cual es la información más relevante a ser incluida en la representación de los patrones. De esta manera la opción que nos queda es aplicar técnicas de selección de características.

Comúnmente la representación de los patrones contiene muchos componentes que contienen información redundante o irrelevante. En este contexto la selección de características busca:

- reducir la dimensionalidad de la representación.
- minimizar los costos extracción de las medidas
- evaluar la performance potencial del sistema de reconocimiento de patrones.
- mejorar la performance del sistema

[*]Estas notas están basadas en las notas del curso Pattern Recognition, Josef Kittler, University of Surrey, y en el libro Pattern Classification de R. Duda, P. Hart y D. Storck, Wiley Interscience 2001.

2 Selección de características

Dado un conjunto $Y = \{f_1, ..., f_n\}$ de características, y una métrica $J(.)$ sobre su relevancia, buscamos $d < n$ características $X_d = \{f_{i_1}, ..., f_{i_d}\}$ tal que:

$$J(X_d) = \max_{X \in \mathcal{X}_d} J(X)$$

con $\mathcal{X}_d = \{\psi_i : \psi_i \in Y, i = 1, ..., d, \psi_i \neq \psi_j \text{ con } i \neq j\}$

La mayoría de las funciones para selección de características satisfacen la propiedad de monotonía la cual puede ser usada en los procedimientos de optimización. La propiedad de **mononotía** se puede expresar de la siguiente manera: Dados M conjuntos encajados:

$$X_1 \subset X_2 \subset ... \subset X_M$$

la función criterio, $J(X_k)$ cumple:

$$J(X_1) \leq J(X_2) \leq ... \leq J(X_M)$$

Intuitivamente esto quiere decir que agregando una nueva observación la cantidad de información que permite la discriminación no puede decrecer.

En general las funciones J son de la forma:

$$J = \int g\left[p(\mathbf{x}|\omega_1), p(\mathbf{x}|\omega_2), P(\omega_1), P(\omega_1)\right] d\mathbf{x}$$

satisfaciendo:

- $J \geq 0$
- $J = 0$ si $p(\mathbf{x}|\omega_1)P(\omega_1) = p(\mathbf{x}|\omega_2)P(\omega_2)\ \forall \mathbf{x}$
- J tiene una máximo cuando $p(\mathbf{x}|\omega_i) = 0$ y $p(\mathbf{x}|\omega_j) \neq 0\ i \neq j$

Ejemplo: Divergencia En el caso de clases equiprobables:

$$J_D = \int \left[p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)\right] \log \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\, d\mathbf{x}$$

Para el caso de gaussianas con $\Sigma_1 = \Sigma_2 = \Sigma$:

$$J = (\mu_2 - \mu_1)^t \Sigma^{-1} (\mu_2 - \mu_1).$$

Esto se puede generalizar al caso de m clases como:

$$J = \sum_{i=1}^{m} \sum_{j>i}^{m} P(\omega_i) P(\omega_j) J^{ij}$$

2.1 Búsquedas óptimas

Son computacionalmente caras dado que son problemas combinatorios.

2.1.1 Caso particular: distribuciones normales independientes

En este caso las matrices Σ_i son diagonales. Entonces podemos usar: $J = \sum_i J(\zeta_i)$ o $J = \prod_i J(\zeta_i)$.

Por ejemplo, en el caso de la divergencia:

$$J_D = \int (p(x|\omega_1) - p(x|\omega_2)) \log \frac{p(x|\omega_1)}{p(x|\omega_2)} dx$$

con $\Sigma_i = \text{diag}(\sigma_{i1}, ..., \sigma_{in})$, y $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in})^t$ para i=1,2. Se llega a que:

$$J_D = \frac{1}{2} \sum_{i=1}^{m} \left[(\mu_{2i} - \mu_{1i})^2 \left(\frac{\sigma_{1i} + \sigma_{2i}}{\sigma_{1i}\sigma_{2i}} \right) + \frac{(\sigma_{2i} - \sigma_{1i})^2}{\sigma_{1i}\sigma_{2i}} \right]$$

Como podemos ver, el resultado final es la suma de m términos de la forma:

$$J(\zeta_i) = \frac{1}{2} (\mu_{2i} - \mu_{1i})^2 \left(\frac{\sigma_{1i} + \sigma_{2i}}{\sigma_{1i}\sigma_{2i}} \right) + \frac{(\sigma_{2i} - \sigma_{1i})^2}{\sigma_{1i}\sigma_{2i}}.$$

De esta forma, se deduce que la solución para maximizar J es ordenar los $J(\zeta_i)$ en forma decreciente y elegir los d primeros.

2.1.2 Algoritmo Branch and Bound

Veremos su funcionamiento a través de un ejemplo.

2.2 Búsquedas sub-óptimas

Cuando incluso el algoritmo Branch and Bound es computacionalmente prohibitivo, se deben usar métodos sub-óptimos (no garantizan la optimalidad de la solución).

2.2.1 Selección secuencial hacia adelante

Se comienza con un conjunto vacío y se agrega la mejor característica individual. En etapas sucesivas se va agregando la mejor característica en combinación con las ya seleccionadas.

2.2.2 Selección secuencial hacia atrás

Es similar al anterior pero se comienza con todas las características y se van eliminando de a una hasta llegar al número deseado.

2.2.3 Más l - menos r

Es una combinación de los dos algoritmos anteriores, que son aplicados alternativamente l y r veces.

3 Extracción de características

Si nos restringimos al caso de transformaciones lineales, buscamos la matriz asociada A que maximiza la métrica J.

$$\max_A J(Ay).$$

3.1 Karhunen-Loeve

Consideremos una base ortonormal $\{\mathbf{u}_i\}$. Podemos desarrollar $\mathbf{y}$ es esta base como:

$$\mathbf{y} = \sum_{i=1}^{\infty} x_i u_I$$

KL busca la base que minimiza el error cuadrático entre $\mathbf{y}$ y su versión truncada $\hat{\mathbf{y}} = \sum_{i=1}^{d} x_i \mathbf{u}_i$.

$$\begin{aligned} e = E\{(\mathbf{y}-\hat{\mathbf{y}})^t(\mathbf{y}-\mathbf{y})\} &= E\left\{\left(\sum_{i=n+1}^{\infty} x_i \mathbf{u}_i\right)\left(\sum_{j=n+1}^{\infty} x_j \mathbf{u}_j\right)\right\} \\ &= E\left\{\sum_{i=n+1}^{\infty} x_j^2\right\} \end{aligned}$$

Recordando que $x_k = \mathbf{u}_k^t y$

$$\begin{aligned} e &= E\left\{\sum_{i=n+1}^{\infty} (\mathbf{u}_i^t \mathbf{y})^t (\mathbf{u}_i^t \mathbf{y})\right\} \\ &= E\left\{\sum_{i=n+1}^{\infty} \mathbf{u}_i^t \mathbf{y}\mathbf{y}^t \mathbf{u}_i\right\} \\ &= \sum_{i=n+1}^{\infty} \mathbf{u}_i^t E\left\{\mathbf{y}\mathbf{y}^t\right\} \mathbf{u}_i \end{aligned}$$

Escribimos la matriz de autocorrelación $R = E\{\mathbf{y}\mathbf{y}^t\}$ y resolvemos el siguiente problema de optimización con restricciones para encontrar la base $\{\mathbf{u}_i\}$.

$$\min_{\{||\mathbf{u}_i||=1\ \forall i\}} \sum_{i=n+1}^{\infty} \mathbf{u}_i^t R \mathbf{u}_i \tag{1}$$

$$\frac{\partial}{\partial \mathbf{u}_j}\left(\sum_{i=n+1}^{\infty} \mathbf{u}_i^t R \mathbf{u}_i + \lambda_i(\mathbf{u}_i^t \mathbf{u}_i - 1)\right) = R\mathbf{u}_j + \lambda_j \mathbf{u}_j$$

Esto quiere decir que la base óptima está formada por los vectores propios de la matriz de autocorrelación. Observar que los x_i son no correlacionados ($E\{x_k x_j\} = \lambda_j \delta_{kj}$).

3.2 Aplicaciones

3.2.1 Problema no supervisado

Cuando no conocemos las clases (no tenemos un etiquetado de las muestras), para comprimir la información solo podemos usar:

$$\Sigma = E\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\}.$$

Si consideramos la matrices de vectores y valores propios de Σ, $U = [\mathbf{u}_1, ..., \mathbf{u}_n]$, $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$, $\Sigma U = U\Lambda$, y ordenamos los valores propios de mayor a menor, la matrix óptima $A = [\mathbf{u}_1, ..., \mathbf{u}_d]$.

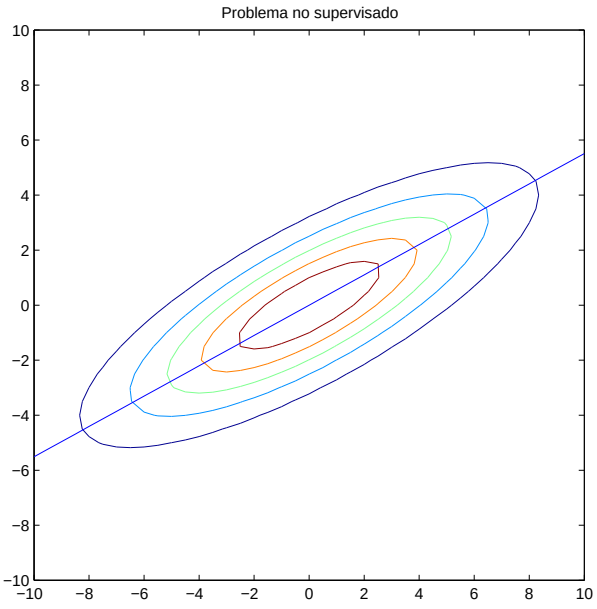
Ejemplo: Consideremos el siguiente problema con distribución gaussiana de parámetros $\boldsymbol{\mu} = [0, 0]^t$,

$$\Sigma = \begin{pmatrix} 19.5 & 9.5 \\ 9.5 & 7.5 \end{pmatrix}.$$

Los valores y vectores propios de Σ resultan: $\lambda_1 = 24.7$,$\lambda_2 = 2.3$,

$$U = \begin{pmatrix} 0.875 & -0.482 \\ 0.482 & 0.875 \end{pmatrix}.$$

De donde se deduce que la proyección óptima es $A = [0.875, 0.482]^t$.



3.2.2 Información de primer orden

Supongamos que la media global $\boldsymbol{\mu} = 0$. La proyección de información de las medias $\boldsymbol{\mu}_i$ a un eje arbitrario $\mathbf{u}_j$ es proporcional a:

$$\mathbf{u}_j^t \sum_{i=1}^{n} P(\omega_i)\boldsymbol{\mu}_i$$

o su cuadrado:

$$\mathbf{u}_j^t M \mathbf{u}_j$$

$$M = \sum_{i=1}^{n} P(\omega_i)\boldsymbol{\mu}_i\boldsymbol{\mu}_i^t$$

donde M es la matriz de dispersión de medias.

Esto solo no alcanza para determinar el poder de discriminación de la característica j, no podemos encontrar los ejes únicamente mirando la matriz M. Los ejes óptimos deben tener en cuenta la varianza de cada clase. Ahora, para poder juzgar correctamente la varianza, los ejes $\mathbf{u}_j$ deben estar decorrelacionados. Por tanto, aplicaremos KL a la matriz:

$$\Phi = \sum_{i=1}^{n} P(\omega_i)\phi_i$$

con

$$\phi_i = E\{(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^t\}, \;\; x \in \omega_I$$

La información de discriminación en el eje $\mathbf{u}_j$ sería la información de proyección de las medias dividida la varianza en ese eje:

$$\frac{\mathbf{u}_j^t M \mathbf{u}_j}{\lambda_j}.$$

Ahora ordenamos

$$\frac{\mathbf{u}_1^t M \mathbf{u}_1}{\lambda_1} \geq \frac{\mathbf{u}_2^t M \mathbf{u}_2}{\lambda_2} \geq ... \geq \frac{\mathbf{u}_n^t M \mathbf{u}_n}{\lambda_n}$$

y elegimos

$$A = [\mathbf{u}_1, ..., \mathbf{u}_d].$$

3.2.3 Ejemplo

Dados

$$P(\omega_1) = P(\omega_1) = 0.5 \quad \mu_1 = (4,2)^t \;\; \mu_2 = (-4,-2)^t$$

$$\Phi_1 = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \quad \Phi_2 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

Si encontramos los vectores y valores propios de la matriz:

$$\Phi = 0.5\Phi_1 + 0.5\Phi_2$$

llegamos a $\lambda_1 = 5$, $\lambda_2 = 2$,

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

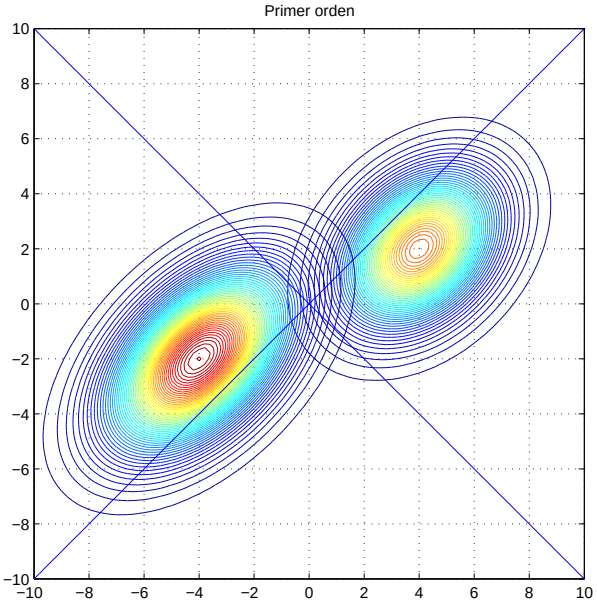
La matriz M es:

$$M = \sum_{i=1}^{2} P(\omega_i)\boldsymbol{\mu}_i\boldsymbol{\mu}_i^t = \begin{pmatrix} 16 & 8 \\ 8 & 4 \end{pmatrix}$$

Por tanto:

$$J(x_1) = \frac{\mathbf{u}_1^t M \mathbf{u}_1}{\lambda_1} = 3.6 \quad J(x_2) = \frac{\mathbf{u}_2^t M \mathbf{u}_2}{\lambda_2} = 1$$

De esta forma el eje óptimo es $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$



4 Mezcla de densidades e identificabilidad

$$p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)P(\omega_j) \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_c)^t$$

Una densidad $p(\mathbf{x}, \boldsymbol{\theta})$ se dice **identificable** si con $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ existe $\mathbf{x}$ tal que $p(\mathbf{x}, \boldsymbol{\theta}) \neq p(\mathbf{x}, \boldsymbol{\theta}')$

Ejemplo

$$\begin{aligned} P(x|\boldsymbol{\theta}) &= \frac{1}{2}\theta_1^x(1-\theta_1)^{1-x} + \frac{1}{2}\theta_2^x(1-\theta_2)^{1-x} \\ &= \begin{cases} \frac{1}{2}(\theta_1+\theta_2) & x=1 \\ 1-\frac{1}{2}(\theta_1+\theta_2) & x=0 \end{cases} \end{aligned}$$

5 Estimación de máxima verosimilitud

Sea $D = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ un conjunto de n muestras independientes, no etiquetadas, obtenidas de:

$$p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j) P(\omega_j)$$

donde el parámetro $\boldsymbol{\theta}$ es desconocido. La **verosimilitud** de las muestras es:

$$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

La estimación $\hat{\boldsymbol{\theta}}$ de máxima verosimilitud es aquella que maximiza $p(D|\boldsymbol{\theta})$.

Si asumimos que $p(D|\boldsymbol{\theta})$ es diferenciable, podemos encontrar las condiciones de optimalidad derivando el logaritmo de la verosimilitud:

$$l = \sum_{k=1}^{n} \log p(\mathbf{x}_k|\boldsymbol{\theta}).$$

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_i} l &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}_i} \left(p(\mathbf{x}_k|\boldsymbol{\theta})\right) \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}_i} \left(\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \right) \end{aligned}$$

Si asumimos que $\boldsymbol{\theta}_i$ y $\boldsymbol{\theta}_j$ son independientes con $i \neq j$:

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}_i} \left(p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) P(\omega_i)\right)$$

Observando que

$$p(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}) P(\omega_i)}{p(\mathbf{x}_k|\boldsymbol{\theta})}$$

se llega a que:

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)$$

Evaluando en el óptimo $\hat{\boldsymbol{\theta}}$:

$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0$$

Ejercicio

Mostrar que en el caso en que las probabilidades a priori son desconocidas, las estimaciones de máxima verosimilitud resultan:

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0$$

$$\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) = \frac{p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)}$$

6 Aplicación a mezclas de normales

6.1 Caso 1: Medias desconocidas

$$\log p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = -\log\left((2\pi)^{d/2}|\Sigma_i|^{1/2}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}_i} \log p(\mathbf{x}|\omega_i, \boldsymbol{\mu}_i) = \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

Resulta entonces que la estimación de máxima verosimilitud $\hat{\boldsymbol{\mu}}$ satisface:

$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}}) \Sigma_i^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_i) = 0$$

Alternativamente, multiplicando por Σ_i^{-1} se llega a:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})}$$

Esta ecuación se puede ver que es una ecuación (vectorial) de punto fijo. Por lo tanto, podemos intentar encontrar la solución del problema mediante un método iterativo a partir de una estimación inicial $\hat{\boldsymbol{\mu}}(0)$:

$$\hat{\boldsymbol{\mu}}_i(s+1) = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}}(s))\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}}(s))}$$

Ejemplo

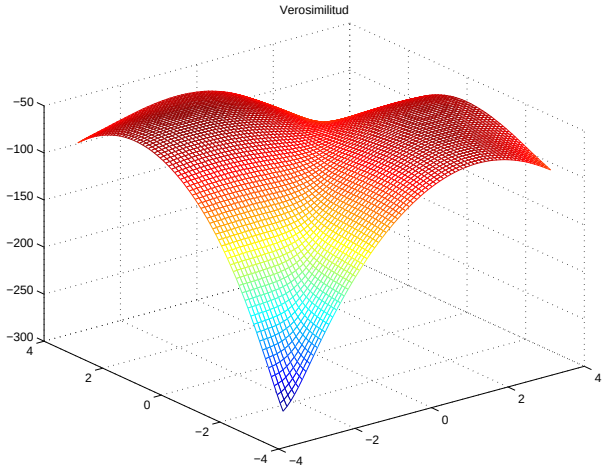
Consideremos la siguiente mezcla de dos gaussianas unidimensionales:

$$p(x|\mu_1,\mu_2) = \frac{1}{3\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_1)^2} + \frac{2}{3\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_2)^2}$$

Se obtuvieron 25 muestras utilizando $\mu_1 = -2$ y $\mu_2 = 2$. Utilizando estas muestras se calculo la verosimilitud

$$l(\mu_1,\mu_2) = \sum_{k=1}^{n} \log p(x_k|\mu_1,\mu_2)$$

para varios valores de μ_1 y μ_2. El resultado aparece en la figura siguiente:



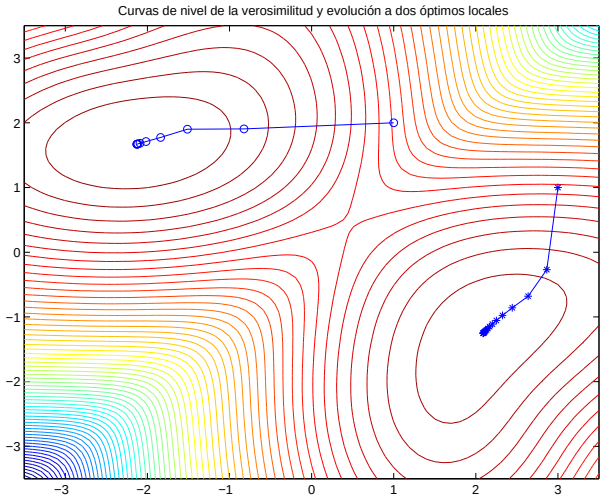
Como se puede ver, la función de verosimilitud tiene dos máximos muy similares. Dependiendo de dónde comience el algoritmo de maximización de la verosimilitud encontraremos una u otra solución (ver figuras).

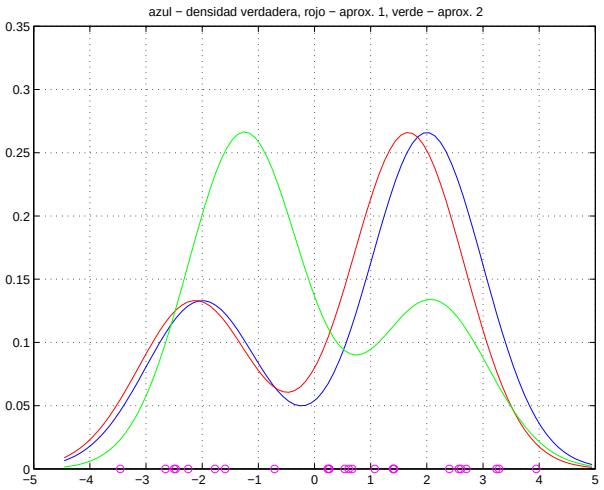
	caso 1	caso 2
$(\mu_1(0),\mu_2(0))$	(1,2)	(3,1)
(μ_1,μ_2)	(-2.13,1.67)	(2.09, -1.26)
Verosimilitud	-52.2	-56.7

6.2 Caso 2: Todos los parámetros desconocidos

Para deducir las ecuaciones correspondientes a este caso alcanza con encontrar las derivadas de $p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)$ respecto a los parámetros de media $\boldsymbol{\mu}_i$, y los elementos de las matrices de covarianza σ_i^{pq}:

$$\nabla_{\boldsymbol{\mu}_i} \log p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i) = \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$


Curvas de nivel de la verosimilitud y evolución a dos óptimos locales
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


azul − densidad verdadera, rojo − aprox. 1, verde − aprox. 2
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−5
−4
−3
−2
−1
0
1
2
3
4
5

$$\frac{\partial \log p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)}{\partial \sigma_i^{pq}} = \left(1 - \frac{\delta_{pq}}{2}\right)[\sigma_i^{pq} - (x_k^p - \mu_i^p)(x_k^q - \mu_i^q)]$$

donde x_k^p y μ_k^p son los elmentos p-ésimos de los vectores $\mathbf{x}_k$ y $\boldsymbol{\mu}_i$ respectivamente y δ_{pq} es la delta de Kronecker. Operando se llega a las siguientes condiciones para un máximo local de la verosimilitud.

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\theta}) \tag{2}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\theta})\mathbf{x}_k}{\hat{P}(\omega_i|\mathbf{x}_k, \hat{\theta})} \tag{3}$$

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\theta})(\mathbf{x}_k - \hat{\mu}_i)(\mathbf{x}_k - \hat{\mu}_i)^t}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\theta})} \tag{4}$$

Ejercicio

1. Deduzca las ecuaciones (2), (3), y (4).

2. Implemente (en `Matlab`, `Octave`, `Scilab`, u otro) un algoritmo que permita encontrar los parámetros $P(\omega_i)$, μ_i y σ_i de una mezcla de dos gaussianas unidimensionales. Para evaluar lo implementado utilice datos sintéticos.