Aprendizaje no supervisado paramétrico

Reconocimiento de Patrones

Departamento de Procesamiento de Señales
Instituto de Ingeniería Eléctrica
Facultad de Ingeniería, UdelaR

2018

Duda, Hart, Stork, cap. 3 y 10 – Dempster, Laird, Rubin, Journal of the Royal Statistical Society, 1977 – Wu, Annals of Statistics, 1983.

Introducción

Aprendizaje no supervisado: conjunto de muestras sin ningún tipo de etiquetado de las clases a las que pertenecen.

- Qué se puede hacer con un conjunto de muestras "mezcladas", cuyas categorías se desconocen?
- Porqué es relevante el problema?
  - Costo de etiquetar las muestras (e.g. speech to text)
  - Grandes cantidades de datos: distinguir primero grupos, luego etiquetar (data mining)
  - Búsqueda de características relevantes para categorizar las muestras
  - Descubrir o entender más sobre la estructura de los datos (grupos o clusters, subgrupos, sub-subgrupos, etc)

Introducción

Dos enfoques:

- *paramétrico:* mezcla de densidades de probabilidad de formas funcionales conocidas (HOY)
  1. Se estiman los parámetros de la mezcla de densidades
  2. Se asigna cada patrón a la clase que maximiza la probabilidad a posteriori.
- *Clustering o Agrupamiento*: particionar los datos en subgrupos o clusters (LA PRÓXIMA)
  - Tantos clusters como clases
  - Puede haber formas funcionales de densidades implícitas en el proceso.

1. Mezcla de Densidades

Asumimos:

1. Cantidad de clases c, conocida
2. Priors $P(\omega_j), j = 1, \ldots, c$ conocidos
3. Verosimilitudes paramétricas: dados $\boldsymbol{\theta}_j, \omega_j$, las $p(\mathbf{x}|\omega_j; \boldsymbol{\theta}_j)$ totalmente determinadas
4. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c$ desconocidos
5. Etiquetas de las muestras desconocidas.

Más adelante levantaremos los supuestos 2) y 1).

1. Mezcla de Densidades

Sea $\Theta = \{\theta_1, \ldots, \theta_c\}$.

Evidencia de $\mathbf{x}$: mezcla de densidades

$$p(\mathbf{x}; \Theta) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j; \theta_j) P(\omega_j).$$

Objetivo:

1. Determinar Θ, y así cada una de las densidades de la mezcla;
2. Una vez estimadas las $p(\mathbf{x}|\omega_j; \theta_j)$, conociendo las $P(\omega_j)$, clasificar las muestras (MAP).

Nota: en lo que sigue suponemos $p(\mathbf{x}; \Theta)$ identificable:

$$\text{Si } \Theta \neq \Theta' \Rightarrow \exists\, \mathbf{x} \text{ t.q. } p(\mathbf{x}; \Theta) \neq p(\mathbf{x}; \Theta').$$

1.1. Estimación de Θ por MLE

$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ muestras no etiquetadas, realización de proceso i.i.d. con densidad

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

Verosimilitud: $p(\mathcal{D}; \mathbf{\Theta}) \overset{\text{i.i.d.}}{=} \prod_{k=1}^{n} p(\mathbf{x}_k; \mathbf{\Theta})$

Log-verosimilitud: $l(\mathbf{\Theta}) = \log p(\mathcal{D}; \mathbf{\Theta}) = \sum_{k=1}^{n} \log p(\mathbf{x}_k; \mathbf{\Theta})$

MLE: $\widehat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} p(\mathcal{D}; \mathbf{\Theta}) = \arg\max_{\mathbf{\Theta}} l(\mathbf{\Theta})$.

1.1. Estimación de Θ por MLE

Si $p(\mathbf{x};\boldsymbol{\Theta})$ diferenciable con respecto a $\boldsymbol{\Theta}$:

$\nabla_{\boldsymbol{\theta}_i} l(\widehat{\boldsymbol{\Theta}}) = \mathbf{0},\ i = 1, 2, \ldots, c$

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

$(*)$ Suponemos θ_i, θ_j funcionalmente indep. $\forall\ i \neq j$.

1.1. Estimación de Θ por MLE

Ahora buscamos introducir las probas de clase a posteriori $\rightarrow$ Bayes:

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

$$\Leftrightarrow \frac{P(\omega_i)}{p(\mathbf{x}_k;\boldsymbol{\Theta})} = \frac{P(\omega_i|\mathbf{x}_k;\boldsymbol{\Theta})}{p(\mathbf{x}_k|\omega_i;\boldsymbol{\theta}_i)}$$

de dónde

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

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

Finalmente, condición necesaria para el estimador ML: $\forall\, i = 1, \ldots, c$,

$$\nabla_{\boldsymbol{\theta}_i} l(\widehat{\boldsymbol{\Theta}}) = \sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k;\widehat{\boldsymbol{\Theta}}) \nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{x}_k|\omega_i;\widehat{\boldsymbol{\theta}}_i) = \mathbf{0}.$$

1.2. Estimación de Θ y $P(\omega_i)$ por MLE

Qué pasa si los $P(\omega_i)$ son desconocidos?

Se estiman por ML, junto con Θ:

$$\{\widehat{\boldsymbol{\Theta}}, \widehat{P(\omega_1)}, \ldots, \widehat{P(\omega_c)}\} = \arg\max_{\boldsymbol{\theta}_i, P(\omega_i), i=1,\ldots,c} l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c, P(\omega_1), \ldots, P(\omega_c))$$

$$\text{sujeto a: } \sum_{i=1}^{c} P(\omega_i) = 1 \tag{1}$$

$$P(\omega_i) \geq 0, \ i = 1, \ldots, c \tag{2}$$

Impondremos (1) con multiplicadores de Lagrange.
Luego verificaremos que (2) se cumple.

1.2. Estimación de Θ y $P(\omega_i)$ por MLE

Lagrangeano $\mathcal{L} = l(\boldsymbol{\Theta}, P(\omega_1), \ldots, P(\omega_c)) - \lambda \left(\sum_{i=1}^{c} P(\omega_i) - 1 \right)$.

$\forall\, i = 1, \ldots, c$:

$$\nabla_{\boldsymbol{\theta}_i} \mathcal{L} = 0 \Leftrightarrow \nabla_{\boldsymbol{\theta}_i} l(\widehat{\boldsymbol{\Theta}}, \widehat{P(\omega_1)}, \ldots, \widehat{P(\omega_c)}) = 0 \qquad \text{(a)}$$

$$\nabla_{P(\omega_i)} \mathcal{L} = 0 \Leftrightarrow \sum_{k=1}^{n} \frac{p(\mathbf{x}_k|\omega_i;\widehat{\boldsymbol{\theta}}_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j;\widehat{\boldsymbol{\theta}}_j)\widehat{P(\omega_j)}} - \lambda = 0 \qquad \text{(b)}$$

$$\frac{d\mathcal{L}}{d\lambda} = 0 \Leftrightarrow \sum_{i=1}^{c} \widehat{P(\omega_i)} = 1 \qquad \text{(c)}$$

Multiplicando (b) por $\widehat{P(\omega_i)}$ y sumando en i se obtiene: $\lambda = n$.

Multiplicando (b) por $\widehat{P(\omega_i)}$ y usando $\lambda = n$:

$$\widehat{P(\omega_i)} = \frac{1}{n} \sum_{k=1}^{n} \frac{p(\mathbf{x}_k|\omega_i;\widehat{\boldsymbol{\theta}}_i)\widehat{P(\omega_i)}}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j;\widehat{\boldsymbol{\theta}}_j)\widehat{P(\omega_j)}} = \frac{1}{n} \sum_{k=1}^{n} \widehat{P(\omega_i|\mathbf{x}_k;\widehat{\boldsymbol{\Theta}})}$$

1.2. Estimación de Θ y $P(\omega_i)$ por MLE

Resumiendo, $\forall\ i = 1, \ldots, c$:

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

$$\widehat{P(\omega_i)} = \frac{1}{n} \sum_{k=1}^{n} \widehat{P(\omega_i|\mathbf{x}_k; \boldsymbol{\Theta})}, \qquad (2)$$

donde

$$\widehat{P(\omega_i|\mathbf{x}_k; \boldsymbol{\Theta})} = \frac{p(\mathbf{x}_k|\omega_i; \boldsymbol{\theta}_i)\widehat{P(\omega_i)}}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j; \boldsymbol{\theta}_j)\widehat{P(\omega_j)}}. \qquad (3)$$

Obs.: De (2), la estimación $\widehat{P(\omega_i)}$ se obtiene como el promedio sobre los datos, de la estimación del prior que da cada dato.

2. Aplicación a mezcla de Gaussianas

2.1. Ejemplo didáctico: μ_i desconocidas; Σ_i, $P(\omega_i)$, c conocidas

$\boldsymbol{\theta}_i = \boldsymbol{\mu}_i,\ \boldsymbol{\Theta} = \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c\}$

$$\log p(\mathbf{x}_k|\omega_i; \boldsymbol{\mu}_i) = -\log((2\pi)^{d/2}(\det \Sigma_i)^{1/2}) - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_i)$$

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

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

$$\Leftrightarrow \boxed{\widehat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k; \widehat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k; \widehat{\boldsymbol{\mu}})}}$$

con

$$P(\omega_i|\mathbf{x}_k; \widehat{\boldsymbol{\mu}}) = \frac{p(\mathbf{x}_k|\omega_i; \widehat{\boldsymbol{\mu}}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j; \widehat{\boldsymbol{\mu}}_j)P(\omega_j)} \quad \text{y} \quad p(\mathbf{x}_k|\omega_i; \widehat{\boldsymbol{\mu}}_i) = \mathcal{N}(\mathbf{x}_k; \widehat{\boldsymbol{\mu}}_i, \Sigma_i).$$

2. Aplicación a mezcla de Gaussianas

- Sistema de c ecuaciones vectoriales, no lineales, de punto fijo que se puede resolver iterativamente:

$$\text{Semilla: } \widehat{\boldsymbol{\mu}_1}(0), \ldots, \widehat{\boldsymbol{\mu}_c}(0)$$

$$\widehat{\boldsymbol{\mu}}_i(t+1) = \frac{\sum_{k=1}^{n} P(\omega_i | \mathbf{x}_k; \widehat{\boldsymbol{\mu}}(t)) \mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i | \mathbf{x}_k; \widehat{\boldsymbol{\mu}}(t))}$$

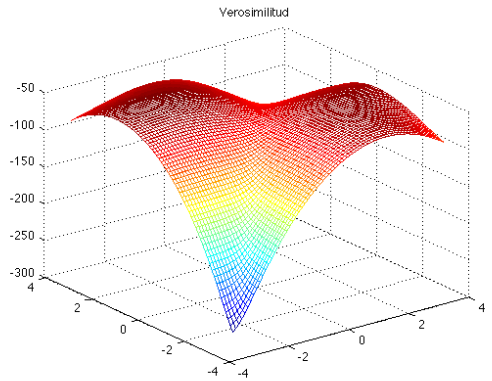
- La solución en general no es única, hay varios óptimos locales. Hay que inicializar varias veces, encontrarlos y quedarse con el de mayor valor.

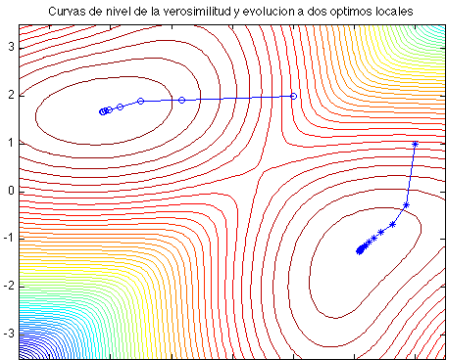
2. Aplicación a mezcla de Gaussianas

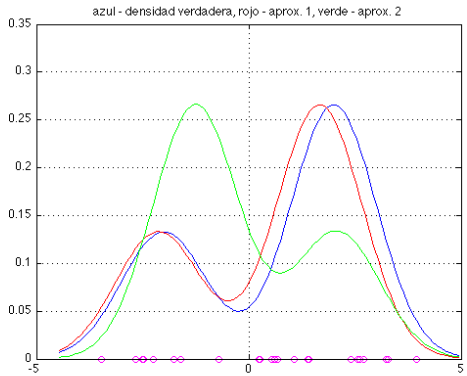
Ejemplo: se simulan 25 muestras según la densidad

$$p(x;\mu_1,\mu_2) = \frac{1}{3\sqrt{2\pi}}\exp\left(\frac{-(x-\mu_1)^2}{2}\right) + \frac{2}{3\sqrt{2\pi}}\exp\left(\frac{-(x-\mu_2)^2}{2}\right).$$

Asumiendo esta forma funcional conocida, estimar μ_1 y μ_2.







2. Aplicación a mezcla de Gaussianas

2.2. Mezcla de Gaussianas donde solo se conoce el número de clases

$$\log p(\mathbf{x}_k|\omega_i;\boldsymbol{\theta}_i) = \log\left(\frac{1}{(2\pi)^{d/2}}(\det\Sigma_i^{-1})^{1/2}\right) - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i)$$

Es más sencillo parametrizar con respecto a Σ_i^{-1}:

$$\boldsymbol{\theta}_i = \left\{\boldsymbol{\mu}_i, \Sigma_i^{-1}\right\} := \left\{\mu_i^1,\ldots,\mu_i^d,((\Sigma_i^{-1}))_{pq}, q\geq p, p=1,\ldots,d\right\}.$$

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

$$\nabla_{\Sigma_i^{-1}}\log p(\mathbf{x}_k|\omega_i;\boldsymbol{\mu}_i) \overset{(*)}{=} \frac{1}{2}\left(\Sigma_i - (\mathbf{x}_k-\boldsymbol{\mu}_i)(\mathbf{x}_k-\boldsymbol{\mu}_i)^T\right)$$

$$(*): \frac{\partial}{\partial A}\log\det A = A^{-T}, \quad \frac{\partial}{\partial A}\text{trace}(AB) = B^T.$$

2. Aplicación a mezcla de Gaussianas

De las Ecs. (1),(2),(3):

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

$$\widehat{\Sigma}_i = \frac{\sum_{k=1}^{n} \widehat{P(\omega_i|\mathbf{x}_k;\boldsymbol{\Theta})}(\mathbf{x}_k - \widehat{\boldsymbol{\mu}}_i)(\mathbf{x}_k - \widehat{\boldsymbol{\mu}}_i)^T}{\sum_{k=1}^{n} \widehat{P(\omega_i|\mathbf{x}_k;\boldsymbol{\Theta})}}$$

$$\widehat{P(\omega_i)} = \frac{1}{n}\sum_{k=1}^{n} \widehat{P(\omega_i|\mathbf{x}_k;\boldsymbol{\Theta})}$$

con $\widehat{P(\omega_i|\mathbf{x}_k;\boldsymbol{\Theta})} = \frac{p(\mathbf{x}_k|\omega_i;\widehat{\boldsymbol{\mu}}_i,\widehat{\Sigma}_i)\widehat{P(\omega_i)}}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j;\widehat{\boldsymbol{\mu}}_j,\widehat{\Sigma}_j)\widehat{P(\omega_j)}}$ y $p(\mathbf{x}_k|\omega_i;\widehat{\boldsymbol{\mu}}_i,\widehat{\Sigma}_i) = \mathcal{N}(\mathbf{x}_k;\widehat{\boldsymbol{\mu}}_i,\widehat{\Sigma}_i)$.

Se resuelve mediante un esquema iterativo como el anterior.

3. Qué pasa cuando el número de clases es desconocido?

Dos criterios muy usados (derivados de teo. de la información):

- *Criterio de Información de Akaike (AIC):*
  $(\widehat{\boldsymbol{\Theta}}, \widehat{c}) = \arg\min_{\boldsymbol{\Theta},c} \{-\log p(\mathcal{D}; \boldsymbol{\Theta}, c) + m_c\}$
- *MDL (Minimum Description Length, Rissannen):*
  $(\widehat{\boldsymbol{\Theta}}, \widehat{c}) = \arg\min_{\boldsymbol{\Theta},c} \left\{-\log p(\mathcal{D}; \boldsymbol{\Theta}, c) + \frac{m_c}{2} \log n\right\}$,

dónde m_c: cantidad de parámetros libres del problema. En mezcla de Gaussianas:

$$m_c = \underbrace{(c-1)}_{\text{priors}} + \underbrace{c \cdot d}_{\text{medias}} + \underbrace{c \cdot \frac{d(d+1)}{2}}_{\text{covarianzas}}.$$

4. El Algoritmo EM (Expectation-Maximization)

4.1. EM: caso general

- Permite encontrar máximos locales de la verosimilitud o el posterior, aún cuando los datos tienen características faltantes
- IDEA: iterar
  1. Un paso de estimación de la verosimilitud con los datos disponibles en la iteración en curso
  2. Un paso de maximización del estimado anterior
- Aún con datos completos, EM puede dar lugar a iteraciones de maximización más sencillas de hallar.

4. El Algoritmo EM (Expectation-Maximization)

$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$, $\mathbf{x}_k \in \mathbb{R}^d$.

$\mathbf{x}_k = (\mathbf{y}_k, \mathbf{z}_k)$, $\mathbf{y}_k$ características buenas, $\mathbf{y}_k$ características malas o faltantes.

Y, Z: conjunto de todas las características buenas y faltantes de $\mathcal{D}$, resp.

$$
\begin{aligned}
Q(\theta, \theta^i) :=& \mathbb{E}_Z \left[\log p(Y, Z; \theta) | Y; \theta^i \right] \\
&= \int \log p(Y, Z; \theta) p(Z|Y; \theta^i) dZ \\
&= \int \log p(Y, Z; \theta) \frac{p(Y, Z; \theta^i)}{p(Y; \theta^i)} dZ \\
&= \int \log p(Y, Z; \theta) \frac{p(Y, Z; \theta^i)}{\int p(Y, Z; \theta^i) dZ} dZ
\end{aligned}
$$

4. El Algoritmo EM (Expectation-Maximization)

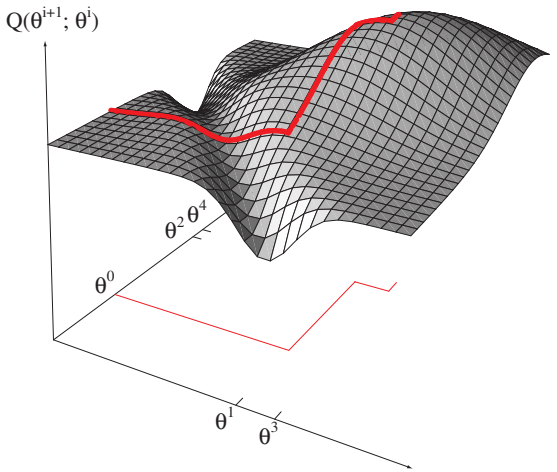
Interpretación de $Q(\theta, \theta^i)$:

- θ^i: mejor estimación de θ a la iteración i- esima
- θ: candidato a mejorar la Estimación
- Marginalización: integra en Z para eliminar la dependencia en los datos desconocidos o malos
- La marginalización se hace con respecto a la mejor distribución que conocemos al momento (dada por θ^i).

Algoritmo:

INIT: $\theta^0,\ T,\ i \leftarrow 0$
repeat
$i \leftarrow i + 1$
(E): compute $Q(\theta, \theta^i)$
(M): $\theta^{i+1} \leftarrow \arg\max_\theta Q(\theta, \theta^i)$
until $Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^{i-1}) \leq T$
return $\widehat{\theta} \leftarrow \theta^{i+1}$

4. El Algoritmo EM (Expectation-Maximization)



Observaciones:

- EM es particularmente útil cuando $Q(\theta;\theta^{i})$ es más fácil de optimizar que $l(\theta)$
- EM garantiza que la log-verosimilitud *con respecto a los datos buenos* crece de forma monótona con las iteraciones (ejercicio del práctico).
- Convergencia a un máximo local asegurada.

4. El Algoritmo EM (Expectation-Maximization)

Ejercicio (Duda & Hart, Sección 3.8, ejemplo 2):

$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \{(0,2), (1,0), (2,2), (x_{4,1}, 4)\}$ muestra iid con distribución Gaussiana de:

- Media (μ_1, μ_2)
- Matriz de covarianza diagonal $\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$.

La característica $x_{4,1}$ es faltante.

Hallar $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

4. El Algoritmo EM (Expectation-Maximization)

4.2. EM: Aplicación a mezcla de densidades

$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ realización de proceso i.i.d. con densidad

$$p(\mathbf{x}_k; \boldsymbol{\Theta}) = \sum_{j=1}^{c} p(\mathbf{x}_k | \Omega_k = \omega_j; \boldsymbol{\theta}_j) P(\Omega_k = \omega_j).$$

IDEA: $\Omega_k \in \{\omega_1, \ldots, \omega_c\}$, $k = 1, \ldots, n$ variables "latentes".

$$\begin{aligned}
Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) &= \mathbb{E}_{\Omega_1,\ldots,\Omega_n} \left[\log p(\mathcal{D}, \Omega_k; \boldsymbol{\Theta}) | \mathcal{D}; \boldsymbol{\Theta}^i\right] \\
&\overset{\mathbf{x}_k \text{i.i.d.}}{=} \mathbb{E}_{\Omega_1,\ldots,\Omega_n} \left[\sum_{k=1}^{n} \log p(\mathbf{x}_k, \Omega_k; \boldsymbol{\Theta}) | \mathcal{D}; \boldsymbol{\Theta}^i\right] \\
&= \sum_{k=1}^{n} \mathbb{E}_{\Omega_1,\ldots,\Omega_n} \left[\log p(\mathbf{x}_k, \Omega_k; \boldsymbol{\Theta}) | \mathcal{D}; \boldsymbol{\Theta}^i\right] \\
&\overset{\Omega_k \text{indeps}}{=} \sum_{k=1}^{n} \mathbb{E}_{\Omega_k} \left[\log p(\mathbf{x}_k, \Omega_k; \boldsymbol{\Theta}) | \mathcal{D}; \boldsymbol{\Theta}^i\right]
\end{aligned}$$

4. El Algoritmo EM (Expectation-Maximization)

$$Q(\Theta, \Theta^i) = \sum_{k=1}^{n} \sum_{j=1}^{c} \log p(\mathbf{x}_k, \Omega_k = \omega_j; \Theta) \underbrace{P(\Omega_k = \omega_j | \mathcal{D}; \Theta^i)}_{P(\Omega_k = \omega_j | \mathbf{x}_k; \Theta^i)}$$

Usando $p(\mathbf{x}_k, \omega_j; \Theta) = p(\mathbf{x}_k | \omega_j; \Theta) P(\omega_j)$

$= p(\mathbf{x}_k | \omega_j; \boldsymbol{\theta}_j) P(\omega_j)$ (ω_j sólo depende de $\boldsymbol{\theta}_j$),

$$\boxed{Q(\Theta, \Theta^i) = \sum_{k=1}^{n} \sum_{j=1}^{c} P(\omega_j | \mathbf{x}_k; \Theta^i) \left(\log p(\mathbf{x}_k | \omega_j; \boldsymbol{\theta}_j) + \log P(\omega_j) \right)}$$

Maximización: $\forall\, j = 1, \ldots, c$,

$$\boxed{\mathbf{0} = \nabla_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c; \Theta^i) = \sum_{k=1}^{n} P(\omega_j | \mathbf{x}_k; \Theta^i) \nabla_{\boldsymbol{\theta}_j} \log p(\mathbf{x}_k | \omega_j; \boldsymbol{\theta}_j)} \qquad \text{(a)}$$

4. El Algoritmo EM (Expectation-Maximization)

Si además los priors son desconocidos, se maximiza bajo las restricciones $P(\omega_j) \geq 0 \; \forall \, j = 1, \ldots, c$ y $\sum_{j=1}^{c} P(\omega_j) = 1$:

$\mathcal{L} = Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) - \lambda \left(\sum_{j=1}^{c} P(\omega_j) - 1 \right)$. (Verificamos positividad a posteriori).

$$0 = \frac{\partial L}{\partial P(\omega_j)} = \sum_{k=1}^{n} P(\omega_j | \mathbf{x}_k; \boldsymbol{\Theta}^i) \frac{1}{P(\omega_j)} - \lambda \quad \text{(b)}$$

$$0 = \frac{\partial L}{\partial \lambda} = 1 - \sum_{j=1}^{c} P(\omega_j) \quad \text{(c)}$$

Multiplicando (b) por $P(\omega_j)$ y sumando en j:

$$\lambda = n, \quad P(\omega_j) = \frac{1}{n} \sum_{k=1}^{n} P(\omega_j | \mathbf{x}_k; \boldsymbol{\Theta}^i) \quad \text{(d)}$$

Obs.: (a) y (d) idem que por ML; se resuelven iterativamente.