TRATAMIENTO DIGITAL DE SEÑALES
Ingeniería de Telecomunicación (4º, 2º c)

Unidad 5ª: El algoritmo EM. Mezclas de gaussianas

Aníbal R. Figueiras Vidal
Jesús Cid Sueiro
Ángel Navia Vázquez

Área de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

Un problema de estimación ML

E: Se dispone de las observaciones $\left\{ x^{(k)} \right\}_{k=1}^{K}$, *tomadas independientemente una de otra, de una mezcla de gaussianas*

$$p(x) = \rho \frac{1}{\sqrt{2\pi v_1}} exp\left[-\frac{(x-m_1)^2}{2v_1} \right] + (1-\rho)\frac{1}{\sqrt{2\pi v_2}} exp\left[-\frac{(x-m_2)^2}{2v_2} \right] = \rho p_1(x) + (1-\rho) p_2(x)$$

siendo los parámetros $\mathbf{s} = [\rho\ m_1\ v_1\ m_2\ v_2]^T$ *valores deterministas desconocidos.*

Establezca las ecuaciones ML para la estimación de dichos parámetros. Discuta la dificultad de resolución de dichas ecuaciones.

$$p\left(\left\{ x \right\}^{(k)} \mid \mathbf{s} \right) = \prod_{k=1}^{K} p\left(x^{(k)} \mid \mathbf{s} \right) = \prod_{k=1}^{K} \left\{ \rho\ p_1\left(x^{(k)} \mid m_1, v_1 \right) + (1-\rho) p_2\left(x^{(k)} \mid m_2, v_2 \right) \right\}$$

$$ln\ p\left(\left\{ x \right\}^{(k)} \mid \mathbf{s} \right) = \sum_{k=1}^{K} ln\left\{ \rho\ p_1\left(x^{(k)} \mid m_1, v_1 \right) + (1-\rho)\ p_2\left(x^{(k)} \mid m_2, v_2 \right) \right\}$$

y las ecuaciones son $\dfrac{\partial \ln}{\partial s_i} = 0$ *; es decir*

$$\frac{\partial \ln}{\partial \rho} = \sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1, v_1\right) - p_2\left(x^{(k)} \mid m_2, v_2\right)}{p\left(x^{(k)} \mid \mathbf{s}\right)} = 0$$

$$\frac{\partial \ln}{\partial m_1} = \sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1, v_1\right)}{p\left(x^{(k)} \mid \mathbf{s}\right)} (-2) \frac{x^{(k)} - m_1}{v_1} = 0$$

$$\frac{\partial \ln}{\partial v_1} = \sum_{k=1}^{K} \frac{1}{p\left(x^{(k)} \mid \mathbf{s}\right)} \rho \frac{1}{\sqrt{2\pi}} \left[-\frac{1}{2} v_2^{-3/2} + \frac{\left(x^{(k)} - m_1\right)^2}{2 v_1^2} \right] exp\left[- \frac{\left(x^{(k)} - m_1\right)^2}{2 v_1} \right] = 0$$

y análogamente para m_2 *y* v_2.

Es un sistema de ecuaciones no lineales: se necesita utilizar algún método de búsqueda para su solución, y el acoplo entre ecuaciones hace serios los problemas de convergencia, mínimos locales, etc.

E: *Considere el mismo caso anterior, pero incluyendo con cada muestra la observación de una variable indicadora que señale si $x^{(k)}$ ha sido generada por la primera gaussiana o por la segunda ($\mathbf{z}^{(k)}$: $[1, 0]^T$ si p_1, $[0, 1]^T$ si p_2).*

Indíquese cómo se procedería en esta situación.

Se dividiría el conjunto de muestras en $\left\{x^{(k)}\right\}_{k=1}^{K'}$ correspondientes a la primera gaussiana y $\left\{x^{(k)}\right\}_{k=K'+1}^{K}$ correspondientes a la segunda; con cada uno de ellos se estimaría m_1, v_1 y m_2, v_2, respectivamente, en la forma ya conocida: estimadores muestrales.

Quedaría entonces aplicar la primera ecuación anterior para la estimación de ρ:

$$\sum_{k=1}^{K} \frac{p_1(x^{(k)} \mid \hat{m}_1, \hat{v}_1) - p_2(x^{(k)} \mid \hat{m}_2, \hat{v}_2)}{\rho\, p_1(x^{(k)} \mid \hat{m}_1, \hat{v}_1) + (1-\rho) p_2(x^{(k)} \mid \hat{m}_2, \hat{v}_2)} = 0$$

$$\sum_{k=1}^{K} \frac{1}{\rho + \dfrac{p_2(x^{(k)} \mid \hat{m}_2, \hat{v}_2)}{p_1(x^{(k)} \mid \hat{m}_1, \hat{v}_1) - p_2(x^{(k)} \mid \hat{m}_2, \hat{v}_2)}} = 0$$

forma $\displaystyle\sum_{k=1}^{K} \frac{1}{\rho + f^{(k)}} = 0$ *, que se solucionaría mediante exploración directa.*

Discusión

Se ha visto que, en una situación moderadamente compleja, la resolución de las ecuaciones ML puede resultar complicada. Sin embargo, si se hubiese contado con la observación del indicador, las cosas serían elementales.

Lo que quiere decir que, si se tuviese acceso a observar una variable **y =** (x, **z**), se podría proceder sin dificultad. Pero no si tenemos observaciones **"incompletas"**, de x (la variable **observable**).

Ante situaciones de este tipo se abre la posibilidad de construir artificialmente una variable **"completa"** **y**; ganando ventaja computacional si sabemos qué hacer para, en definitiva, maximizar p(**x**|**s**) (que es el problema ML planteado).

A ello nos ayuda el algoritmo que se expone acto seguido.

El Algoritmo EM ("Expectation-Maximization")

Dempster, Laird y Rubin propusieron proceder así: establecida la expresión

$$p(\mathbf{y} \mid \mathbf{s}) = p(\mathbf{y} \mid \mathbf{x}, \mathbf{s})\, p(\mathbf{x} \mid \mathbf{s})$$

y sobre la forma

$$\ln p(\mathbf{y} \mid \mathbf{s}) = \ln p(\mathbf{y} \mid \mathbf{x}, \mathbf{s}) + \ln p(\mathbf{x} \mid \mathbf{s})$$

* inicialícese **s** en **s**(1)

* en el paso n, aplíquese

- una **etapa E**: promediando ln p(**y** | **s**) respecto a p(**y** | **x**, **s**(n)); es decir, calculando

$$\int \ln p(\mathbf{y} \mid \mathbf{x}, \mathbf{s})\, p(\mathbf{y} \mid \mathbf{x}, \mathbf{s}(n))\, d\mathbf{y} + \ln p(\mathbf{x} \mid \mathbf{s})$$

que indicaremos como

$$Q(\mathbf{s}, \mathbf{s}(n)) = E_{\mid \mathbf{s}(n)} \{\ln p(\mathbf{y} \mid \mathbf{s})\}$$

- una **etapa M**: $\mathbf{s}(n+1) = \arg \{\max_{s} Q(\mathbf{s}, \mathbf{s}(n))\}$

* itérese hasta la convergencia.

En el apéndice se prueba que este algoritmo maximiza p(**x** | **s**)

<u>Notas</u>

- El Algoritmo EM no está exento de padecer problemas de detención en máximos locales.

- Si se sustituye la etapa M (p. ej., por dificultad analítica) por un mero aumento de Q, el algoritmo resultante (**GEM**: “Generalized EM”) tiene igual sentido.

- Si en las expresiones anteriores se añade ln p(**s**) a ln p(**y** | **s**), se tiene la formulación EM para planteamiento MAP.

<u>Ejemplo de aplicación</u> (Modelado EM-GM)

A: Formúlese el Algoritmo EM para la mezcla de dos Gaussianas, utilizando la variable indicadora ***z*** *como no observable.*

Se tiene como variable completa $\mathbf{y} = \mathbf{z}x$; *cuya ddp será* $\rho\, p_1$ *si z toma el valor* $[1, 0]^T$, y $(1-\rho)p_2$ *si* $\mathbf{z}$ *es* $[0, 1]^T$; *es decir*

$$p\left(\left\{\mathbf{y}^{(k)}\right\}|\,\mathbf{s}\right)=\prod_{k=1}^{K}\mathbf{z}^{(k)T}\begin{bmatrix}\rho\, p_1\left(x^{(k)}\,|\,m_1,v_1\right)\\(1-\rho)p_2\left(x^{(k)}\,|\,m_2,v_2\right)\end{bmatrix}$$

$$\ln p\left(\left\{\mathbf{y}^{(k)}\right\}|\,\mathbf{s}\right)=\sum_{k=1}^{K}\mathbf{z}^{(k)T}\begin{bmatrix}\ln\rho\, p_1\left(x^{(k)}\,|\,m_1,v_1\right)\\\ln(1-\rho)p_2\left(x^{(k)}\,|\,m_2,v_2\right)\end{bmatrix}$$

con lo que el Algoritmo EM resulta:

- *etapa E*

$$Q(\mathbf{s},\mathbf{s}(n))=\sum_{k=1}^{K}\left[E_{|\mathbf{s}(n)}\left\{z_1^{(k)}\right\}\quad E_{|\mathbf{s}(n)}\left\{z_2^{(k)}\right\}\right]\begin{bmatrix}\ln\rho\, p_1\left(x^{(k)}\,|\,m_1,v_1\right)\\\ln(1-\rho)p_2\left(x^{(k)}\,|\,m_2,v_2\right)\end{bmatrix}$$

dado que $z_1(z_2)$ *es 1 con probabilidad* $\rho\, p_1/p$ $((1-\rho)p_2/p)$, *y si no 0:*

$$E_{|\mathbf{s}(n)}\left\{z_1^k\right\}=\frac{\rho(n)p_1\left(x^{(k)}\mid m_1(n),v_1(n)\right)}{p\left(x^{(k)}\mid \mathbf{s}(n)\right)}=w^{(k)}(n)$$

$$E_{|\mathbf{s}(n)}\left\{z_2^{(k)}\right\}=\frac{(1-\rho(n))\,p_2\left(x^{(k)}\mid m_2(n),v_2(n)\right)}{p\left(x^{(k)}\mid \mathbf{s}(n)\right)}=1-w^{(k)}(n)$$

con lo que resulta

$$Q(\mathbf{s},\mathbf{s}(n))=\sum_{k=1}^{K}\left[w^{(k)}(n)\quad 1-w^{(k)}(n)\right]\left[\begin{array}{c}\ln\rho\, p_1\left(x^{(k)}\mid m_1,v_1\right)\\ \ln(1-\rho)p_2\left(x^{(k)}\mid m_2,v_2\right)\end{array}\right]$$

- Etapa M

$$*\qquad \frac{\partial Q(\mathbf{s},\mathbf{s}(n))}{\partial\rho}=\sum_{k=1}^{K}\left[\frac{w^{(k)}(n)}{\rho}-\frac{1-w^{(k)}(n)}{1-\rho}\right]=0$$

$$\sum_{k=1}^{K}\left[(1-\rho)w^{(k)}(n)-\rho\left(1-w^{(k)}(n)\right)\right]=0\quad ;\qquad \sum_{k=1}^{K}\left(w^{(k)}(n)-\rho\right)=0$$

de donde

$$\rho(n+1)=\frac{1}{K}\sum_{k=1}^{K}w^{(k)}(n)=\frac{\rho(n)}{K}\sum_{k=1}^{K}\frac{p_1\left(x^{(k)}\mid m_1(n),v_1(n)\right)}{p\left(x^{(k)}\mid \mathbf{s}(n)\right)}$$

$$* \quad \frac{\partial Q(\mathbf{s},\mathbf{s}(n))}{\partial m_1} = \sum_{k=1}^{K} w^{(k)}(n) \, \frac{\partial \ln p_1\left(x^{(k)} \mid m_1, v_1\right)}{\partial m_1} = 0$$

$$\sum_{k=1}^{K} \frac{\rho(n)\, p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} \frac{x^{(k)} - m_1}{v_1} = 0$$

$$\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} \left(x^{(k)} - m_1\right) = 0$$

de donde

$$m_1(n+1) = \frac{\displaystyle\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} x^{(k)}}{\displaystyle\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)}}$$

y análogamente $m_2(n+1)$

$$* \quad \frac{\partial Q(\mathbf{s},\mathbf{s}(n))}{\partial v_1} = \sum_{k=1}^{K} w^{(k)}(n)\, \frac{\partial \ln p_1\left(x^{(k)} \mid m_1, v_1\right)}{\partial v_1} = 0$$

$$\sum_{k=1}^{K} \frac{\rho(n) p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} \left[-\frac{1}{2v_1} + \frac{1}{2v_1^2}\left(x^{(k)} - m_1\right)^2 \right] = 0$$

$$\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} \left[\left(x^{(k)} - m_1\right)^2 - v_1 \right] = 0$$

de donde

$$v_1(n+1) = \frac{\displaystyle\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} \left(x^{(k)} - m_1(n+1)\right)^2}{\displaystyle\sum_{k=1}^{K} \frac{p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)}}$$

y análogamente $v_2(n+1)$

Notando que

$$\sum_{k=1}^{K} \frac{\rho(n)p_1\left(x^{(k)} \mid m_1(n), v_1(n)\right)}{p\left(x^{(k)} \mid \mathbf{s}(n)\right)} = \sum_{k=1}^{K} w^{(k)}(n)$$

y lo mismo con subíndices 2 y 1-$w^{(k)}(n)$, son posibles expresiones más compactas:

$$\rho(n+1) = \frac{1}{K}\sum_{k=1}^{K} w^{(k)}(n)$$

$$m_1(n+1) = \frac{\sum_{k=1}^{K} w^{(k)}(n)\, x^{(k)}}{\sum_{k=1}^{K} w^{(k)}(n)}$$

$$v_1(n+1) = \frac{\sum_{k=1}^{K} w^{(k)}(n)\left(x^{(k)} - m_1(n+1)\right)^2}{\sum_{k=1}^{K} w^{(k)}(n)}$$

y análogamente para m_2, v_2, con 1-$w^{(k)}(n)$.

La generalización de lo anterior a una mezcla de G gaussianas multidimensionales (**GM,** "Gaussian Mixture") es inmediata: si la ddp es

$$p(\mathbf{x}\,|\,\mathbf{s}) = \sum_{j=1}^{G} \rho_j p_j\left(\mathbf{x}\,|\,\mathbf{m}_j, V_j\right)$$

se tiene

$$\rho_j(n+1) = \frac{1}{K}\sum_{k=1}^{K} w_j^{(k)}(n)$$

$$m_j(n+1) = \frac{\sum_{k=1}^{K} w_j^{(k)}(n)\mathbf{x}^{(k)}}{\sum_{k=1}^{K} w_j^{(k)}(n)}$$

$$V_j(n+1) = \frac{\sum_{k=1}^{K} w_j^{(k)}(n)\left(\mathbf{x}^{(k)} - \mathbf{m}_j(n+1)\right)^T\left(\mathbf{x}^{(k)} - \mathbf{m}_j(n+1)\right)}{\sum_{k=1}^{K} w_j^{(k)}(n)}$$

Apéndice: convergencia del Algoritmo EM

Se tiene que

$$Q(\mathbf{s},\mathbf{s}(n)) = E_{|\mathbf{s}(n)}\{\ln p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s})\} + \ln p(\mathbf{x}\,|\,\mathbf{s})$$
$$= H(\mathbf{s},\mathbf{s}(n)) + \ln p(\mathbf{x}\,|\,\mathbf{s})$$

y es

$$H(\mathbf{s},\mathbf{s}(n)) - H(\mathbf{s}(n),\mathbf{s}(n)) = E_{|\mathbf{s}(n)}\left\{\ln \frac{p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s})}{p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s}(n))}\right\} =$$

$$= \int_Y \ln\left[\frac{p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s})}{p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s}(n))}\right] p(\mathbf{y}\,|\,\mathbf{x},\mathbf{s}(n))\,d\mathbf{y} \qquad \leq\ 0 \quad (\text{si y sólo si } \mathbf{s} = \mathbf{s}(n))$$

debido a la versión límite de la desigualdad de Jensen: para una función convexa f y unos pesos $p_j \geq 0$ tales que $\sum_j p_j = 1$

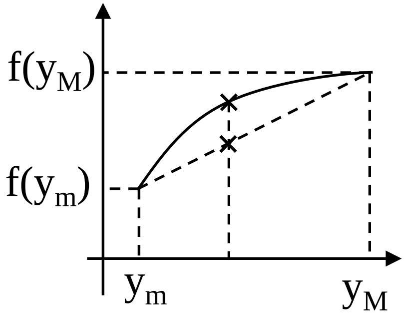
$$f\left(\sum_j p_j y_j\right) \geq \sum_j p_j f(y_j)$$

tomando ln como f y p($\mathbf{y}$ | $\mathbf{x}$, $\mathbf{s}$(n)) en el papel de las p_j

$$H(\mathbf{s},\mathbf{s}(n)) - H(\mathbf{s}(n),\mathbf{s}(n)) \leq \ln \int_Y \left[\frac{p(\mathbf{y}\mid\mathbf{x},\mathbf{s})}{p(\mathbf{y}\mid\mathbf{x},\mathbf{s}(n))}\right] p(\mathbf{y}\mid\mathbf{x},\mathbf{s}(n))\,d\mathbf{y} =$$

$$= \ln \int_Y p(\mathbf{y}\mid\mathbf{x},\mathbf{s})\,d\mathbf{y}$$

de modo que en cada paso M, al decrecer H, se incrementa p($\mathbf{x}$ | $\mathbf{s}$).



La prueba de la desigualdad de Jensen es elemental: f es convexa si, con $0 \leq p \leq 1$

$$f\,[p y_m + (1-p)\, y_M] \geq p f(y_m) + (1-p) f(y_M)$$

ya que $p y_m + (1-p)\, y_M$ es un punto del intervalo $[y_m, y_M]$, y la función supera a la cuerda.

Probada la desigualdad para dos sumandos, se supone cumplida para J; como para J+1 se tiene

$$p_{J+1}y_{J+1} + \sum_{j=1}^{J} p_j y_j = p_{J+1}y_{J+1} + (1-p_{J+1})\sum_{j=1}^{J}\frac{p_j}{1-p_{J+1}}\, y_j = p_{J+1}y_{J+1} + (1-p_{J+1})\sum_{j=1}^{J} p'_j y_j$$

con $p'_j \geq 0, \ \sum_{j=1}^{J} p'_j = 1$; aplicando la convexidad

$$f\left(\sum_{j=1}^{J+1} p_j y_j\right) \geq p_{J+1} f(y_{J+1}) + (1-p_{J+1}) f\left(\sum_{j=1}^{J} p'_j y_j\right)$$

y por la desigualdad para J

$$f\left(\sum_{j=1}^{J+1} p_j y_j\right) \geq p_{J+1} f(y_{J+1}) + (1-p_{J+1})\sum_{j=1}^{J} p'_j f(y_j) = p_{J+1} f(y_{J+1}) + \sum_{j=1}^{J} p_j f(y_j)$$

que es, en definitiva

$$f\left(\sum_{j=1}^{J+1} p_j y_j\right) \geq \sum_{j=1}^{J+1} p_j f\left(y_j\right)$$

y cuya versión límite es

$$f\left(\int y(x)\, p(x)\, dx\right) \geq \int f\left[y(x)\, p(x)\right] dx$$

con p(x) ≥ 0 e $\int p(x)\, dx = 1$