El Algoritmo E-M

José Antonio Camarena Ibarrola

Introducción

- Método para encontrar una estimación de máxima verosimilitud para un parámetro ɵ de una distribución

Ejemplo simple

- Si $x \in \Re^{24}$ tiene las temperaturas del jardín de c/u de las 24 horas del día
- Sabemos que la distribución de probabilidad x depende de la estación ɵ (Primavera, Verano, Otoño, Invierno)
- Suponga que lo único con lo que en realidad contamos es con la temperatura promedio del día $y = \overline{x}$
- Se quiere hacer una estimación acerca de ɵ
- Un estimador de máxima verosimilitud maximiza $p(y|\theta)$ para encontrar ɵ pero puede ser un problema intratable
- El algoritmo EM hace suposiciones acerca de los datos completos x e iterativamente encuentra el ɵ que maximiza $p(x|\theta)$

Que se necesita para usar EM

- Los datos observados y
- Una densidad paramétrica que describa los datos observados $p(y|\theta)$
- Una descripción de los datos completos x
- Una densidad paramétrica de los datos completos $p(x|\theta)$ considerando que el soporte de X no depende de ɵ

X no se observa directamente

- Lo que se observa son realizaciones y de la variable aleatoria Y=T(X)
- Ej. T puede mapear X a su media
- Ej. Si X es un número complejo Y es solo su magnitud
- Ej. T devuelve la norma de un vector X

Estimador de máxima verosimilitud

- Dado que solo contamos con y, la estimación de ɵ deseable es $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} p(y \mid \theta)$
- Frecuentemente es mas fácil maximizar la log-verosimilitud $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \log p(y \mid \theta)$
- En muchas ocasiones, es muy complicado maximizar cualquiera de las dos, entonces EM surge como una mejor opción
- EM hacemos una suposición acerca de X (los datos completos), luego encontramos el ɵ que maximiza el valor esperado de la log-verosimilitud de X, una vez que tenemos el nuevo ɵ, podemos elegir una mejor elección respecto a X e iteramos

El paso E

1. Elegir un valor inicial para ɵ $\theta^{(m=0)}$
2. Dados los datos observados y y pretendiendo por ahora que la suposición actual de $\theta^{(m)}$ es correcta, calcular que tan probable es que los datos completos valgan x, es decir, determinar la distribución condicional $p(x \mid y, \theta^{(m)})$
3. Desechar $\theta^{(m)}$, sin embargo, preservar $p(x \mid y, \theta^{(m)})$
4. Quisiéramos maximizar $\log p(x \mid \theta)$, pero como no conocemos realmente x, vamos a maximizar su valor esperado a lo cual llamamos "Función Q".

$$Q(\theta \mid \theta^{(m)}) = E_{X|y,\theta^{(m)}}[\log p(X \mid \theta)] = \int_{\mathcal{X}(y)} \log p(x \mid \theta) p(x \mid y, \theta^{(m)}) dx$$

- Se integra sobre todo el soporte $\mathcal{X}(y) = \{x \mid p(x \mid y) > 0\}$
- Observe que la función Q depende de ɵ pero también de la suposición inicial de $\theta^{(m)}$

El paso M

5. Elegir un nuevo valor para ɵ $\theta^{(m+1)}$ seleccionando aquel que maximiza a la función Q

$$\theta^{(m+1)} = \arg\max_{\theta \in \Theta} Q(\theta \mid \theta^{(m)})$$

6. Hacer m=m+1
7. Volver a (2)

Expresión mas conveniente para la función Q

- Vimos que

$$Q(\theta \mid \theta^{(m)}) = E_{X|y,\theta^{(m)}} [\log p(X \mid \theta)] = \int_{\mathcal{X}(y)} \log p(x \mid \theta) p(x \mid y, \theta^{(m)}) dx$$

- Pero de acuerdo a Bayes $p(x \mid y, \theta) = \frac{p(x, y \mid \theta)}{p(y \mid \theta)}$

- Y como Y=T(X) es una función determinista $p(x \mid y, \theta) = \frac{p(x \mid \theta)}{p(y \mid \theta)}$

- Decir que Y=T(X) es una función determinista significa que una realización x determina a y de manera única y por lo tanto la probabilidad de que ocurra el par (x,y) es igual a la probabilidad de que ocurra x, por ejemplo la probabilidad de que llueva es igual a la probabilidad de que llueva y haya nubes en el cielo.

- Finalmente: $$Q(\theta \mid \theta^{(m)}) = \int_{\mathcal{X}(y)} \log p(x \mid \theta) \frac{p(x \mid \theta^{(m)})}{p(y \mid \theta^{(m)})} dx$$

Ejemplo

- A n niños les dan a escoger un juguete de entre 4 opciones
- El histograma de elecciones es $Y = [Y_1 \quad \dots \quad Y_4]^T$

Entonces Y_1 es el número de niños que escogieron el juguete 1, etc

- Podemos modelar el histograma aleatorio Y mediante la distribución multinomial la cual tiene 2 parámetros, el número de ensayos (n) y la probabilidad de que los niños elijan cada uno de los 4 juguetes, el vector $p \in (0,1)^4$; $p_1 + p_2 + p_3 + p_4 = 1$
- La probabilidad de observar un histograma en particular es

$$P(y \mid \theta) = \frac{n!}{y_1! y_2! y_3! y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}$$

Ejemplo (continuación)

- Suponga que en este caso p depende de un parámetro ɵ de manera que

$$p_\theta = \left[\tfrac{1}{2} + \tfrac{1}{4}\theta \quad \tfrac{1}{4}(1-\theta) \quad \tfrac{1}{4}(1-\theta) \quad \tfrac{1}{4}\theta\right]^T,\ \theta \in (0,1)$$

- El problema consiste en estimar el valor de ɵ que maximiza la verosimilitud del histograma observado
- La probabilidad de observar el histograma $y = [y_1 \quad y_2 \quad y_3 \quad y_4]$

es entonces:

$$P(y \mid \theta) = \frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}.$$

- Este ejemplo puede resolverse mediante un maximización de la log-verosimilitud directamente, sin embargo lo resolveremos mediante EM

El truco

- Para utilizar EM necesitamos especificar a que llamamos X (los datos completos) y hacerlo de manera que su distribución también se pueda expresar en términos del mismo parámetro ɵ. Para nuestro ejemplo, una solución es considerar $X = \begin{bmatrix} X_1 & \dots & X_5 \end{bmatrix}^T$

con distribución multinomial donde las probabilidades de cada evento son $\begin{bmatrix} \frac{1}{2} & \frac{1}{4}\theta & \frac{1}{4}(1-\theta) & \frac{1}{4}(1-\theta) & \frac{1}{4}\theta \end{bmatrix}^T, \ \theta \in (0,1)$

- Y la relación entre los datos observados Y con los datos completos X es: $Y = T(X) = \begin{bmatrix} X_1 + X_2 & X_3 & X_4 & X_5 \end{bmatrix}^T$
- Por tanto, la verosimilitud de una realización x de los datos completos es:

$$P(x \mid \theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

La función Q del ejemplo

- Sabemos que la función Q está definida por

$$Q(\theta \mid \theta^{(m)}) \equiv E_{X|y,\theta^{(m)}} [\log p(X \mid \theta)]$$

- En nuestro caso

$$P(x \mid \theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

- Para maximizar Q solo necesitamos los términos que dependan de ɵ, los demás son irrelevantes (para efectos de la maximización sobre ɵ)

$$\theta^{(m+1)} = \arg \max_{\theta \in (0,1)} E_{X|y,\theta^{(m)}} [(X_2 + X_5) \log \theta + (X_3 + X_4) \log(1-\theta)]$$

$$\equiv \arg \max_{\theta \in (0,1)} \left(\log \theta \left(E_{X|y,\theta^{(m)}}[X_2] + E_{X|y,\theta^{(m)}}[X_5]\right) + \log(1-\theta) \left(E_{X|y,\theta^{(m)}}[X_3] + E_{X|y,\theta^{(m)}}[X_4]\right)\right)$$

Obteniendo la expresión a maximizar

- Para resolver

$$\theta^{(m+1)} = \arg\max_{\theta\in(0,1)} E_{X|y,\theta^{(m)}}\left[(X_2+X_5)\log\theta + (X_3+X_4)\log(1-\theta)\right]$$

$$\equiv \arg\max_{\theta\in(0,1)} \left(\log\theta\left(E_{X|y,\theta^{(m)}}[X_2] + E_{X|y,\theta^{(m)}}[X_5]\right) + \log(1-\theta)\left(E_{X|y,\theta^{(m)}}[X_3] + E_{X|y,\theta^{(m)}}[X_4]\right)\right)$$

- Necesitamos la esperanza condicional de los datos completos X condicionada a la realización de datos completos conocidos (y) lo cual solo nos deja incertidumbre acerca de X_1, X_2 dado que $X_1 + X_2 = y_1$
- Dado y_1, el par X_1, X_2 se distribuye binomialmente

$$P(x\,|\,y,\theta) = \frac{y_1!}{x_1!x_2!}\left(\frac{2}{2+\theta}\right)^{x_1}\left(\frac{\theta}{2+\theta}\right)^{x_2} 1_{\{x_1+x_2=y_1\}}\prod_{i=3}^{5} 1_{\{x_i=y_{i-1}\}}$$

$1_{\{\cdot\}}$ es la función indicadora

- El valor esperado es la media, en este caso $P(x\,|\,y,\theta)$ es binomial por tanto $E_{X|y,\theta}[X] = \left[\frac{2}{2+\theta}y_1 \quad \frac{\theta}{2+\theta}y_1 \quad y_2 \quad y_3 \quad y_4\right]^T$

Obteniendo la expresión a maximizar y comenzamos a maximizar de una vez

- Sustituyendo $E_{X|y,\theta}[X] = \left[\frac{2}{2+\theta}y_1 \quad \frac{\theta}{2+\theta}y_1 \quad y_2 \quad y_3 \quad y_4\right]^T$

en

$$\theta^{(m+1)} = \arg\max_{\theta\in(0,1)} \left(\log\theta\left(E_{X|y,\theta^{(m)}}[X_2] + E_{X|y,\theta^{(m)}}[X_5]\right) + \log(1-\theta)\left(E_{X|y,\theta^{(m)}}[X_3] + E_{X|y,\theta^{(m)}}[X_4]\right)\right)$$

obtenemos

$$\theta^{(m+1)} = \arg\max_{\theta\in(0,1)} \left(\log\theta\left(\frac{\theta^{(m)}y_1}{2+\theta^{(m)}} + y_4\right) + \log(1-\theta)(y_2+y_3)\right)$$

- Derivando e igualando a cero obtenemos

$$\left[\frac{\theta^{(m)}y_1}{2+\theta^{(m)}} + y_4\right]\frac{1}{\theta} + \frac{y_2+y_3}{1-\theta}(-1) = 0$$

Maximizando

$$\left[\frac{\theta^{(m)}y_1}{2+\theta^{(m)}}+y_4\right]\frac{1}{\theta}+\frac{y_2+y_3}{1-\theta}(-1)=0$$

$$\left[\frac{\theta^{(m)}y_1}{2+\theta^{(m)}}+y_4\right](1-\theta)=(y_2+y_3)\theta$$

$$\frac{\theta^{(m)}y_1}{2+\theta^{(m)}}+y_4=\left[y_2+y_3+\frac{\theta^{(m)}y_1}{2+\theta^{(m)}}+y_4\right]\theta$$

$$\theta^{(m+1)}=\frac{\frac{\theta^{(m)}}{2+\theta^{(m)}}y_1+y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}}y_1+y_2+y_3+y_4}$$

Ejecutando el algoritmo EM

$$\theta^{(m+1)} = \frac{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_2 + y_3 + y_4}$$

Si iniciamos con $\theta^{(0)} = 0.5$

Si los datos observados (el histograma) fueran y=[55 20 20 5]

Aplicando sucesivamente :

t=((t/(2+t))*y(1)+y(4))/((t/(2+t))*y(1)+y(2)+y(3)+y(4))

m	$\theta^{(m)}$
0	**0.5**
1	0.2857
2	0.2289
3	0.2102
4	0.2037
5	0.2013
6	0.2005
7	0.2002
8	0.2001

Cuando realmente se trata de un problema de datos incompletos

- Si los datos completos X consisten de datos observados Y y de datos perdidos (u ocultos) de manera que X=(Y,Z) se puede escribir la función Q como una integral sobre todo el dominio de Z dado que es la única parte aleatoria de X
- Un ejemplo de esto es cuando se estiman los parámetros de una mezcla de gaussianas
- Otro ejemplo es cuando se estiman los parámetros de un modelo oculto de Markov

$$
\begin{aligned}
Q(\theta \mid \theta^{(m)}) &= E_{X|y,\theta^{(m)}} \left[\log p_X(X \mid \theta)\right] \\
&= E_{Z|y,\theta^{(m)}} \left[\log p_X(y, Z \mid \theta)\right] \\
&= \int_{\mathcal{Z}} \log p_X(y, z \mid \theta) p_{Z|Y}(z \mid y, \theta^{(m)}) dz
\end{aligned}
$$

Mezcla de gaussianas

Una mezcla de gaussianas es una suma ponderada de k gaussianas

$$p(y \mid \theta) = \sum_{i=1}^{k} w_i \phi(\mu_i, \Sigma_i)$$

$$= \sum_{i=1}^{k} w_i \frac{\exp\left(-\frac{1}{2}(y-\mu_i)^T \Sigma_i^{-1}(y-\mu_i)\right)}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}$$

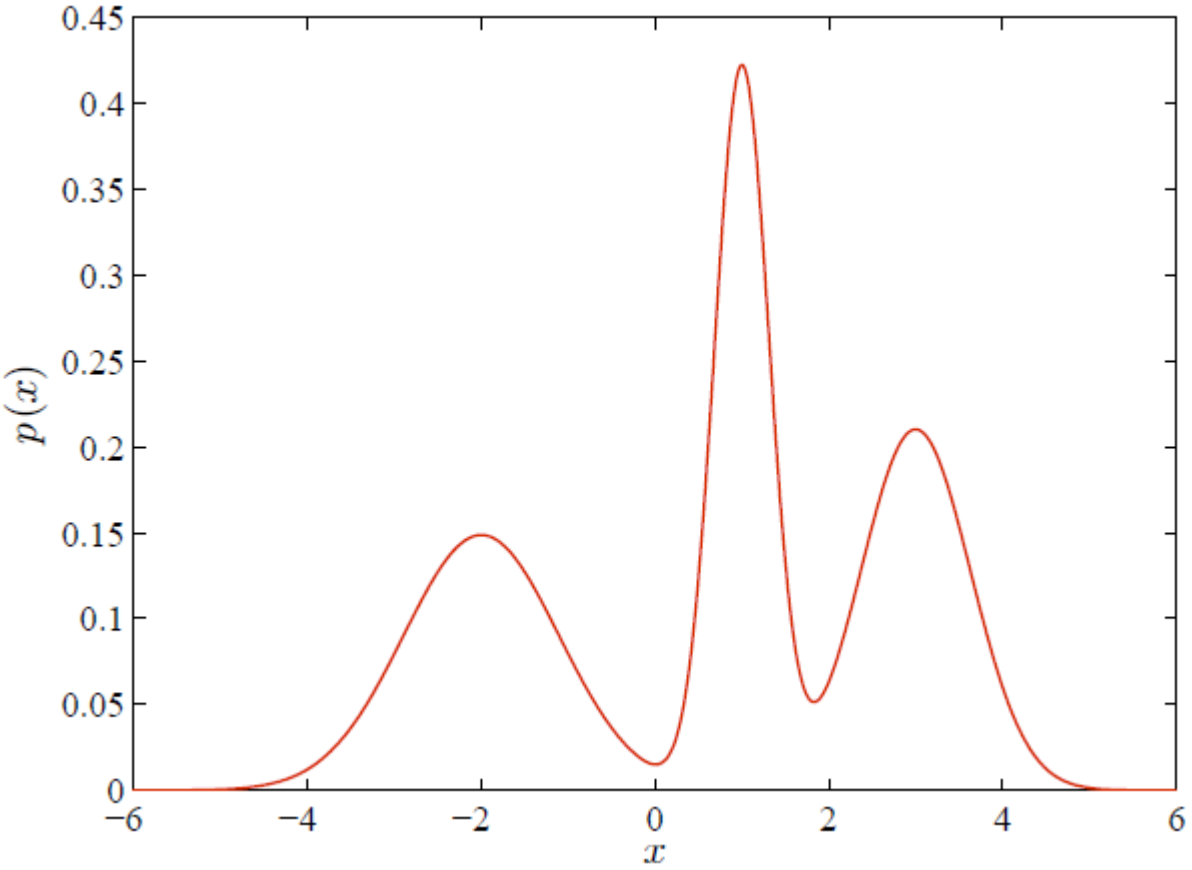
Donde:

$$w_i > 0,\ i = 1, \ldots, k, \text{ and } \sum_{i=1}^{k} w_i = 1$$

Dadas n observaciones de dimensión d, deseamos estimar los parámetros:

$$\theta = \{(w_i, \mu_i, \Sigma_i)\}_{i=1}^{k}$$

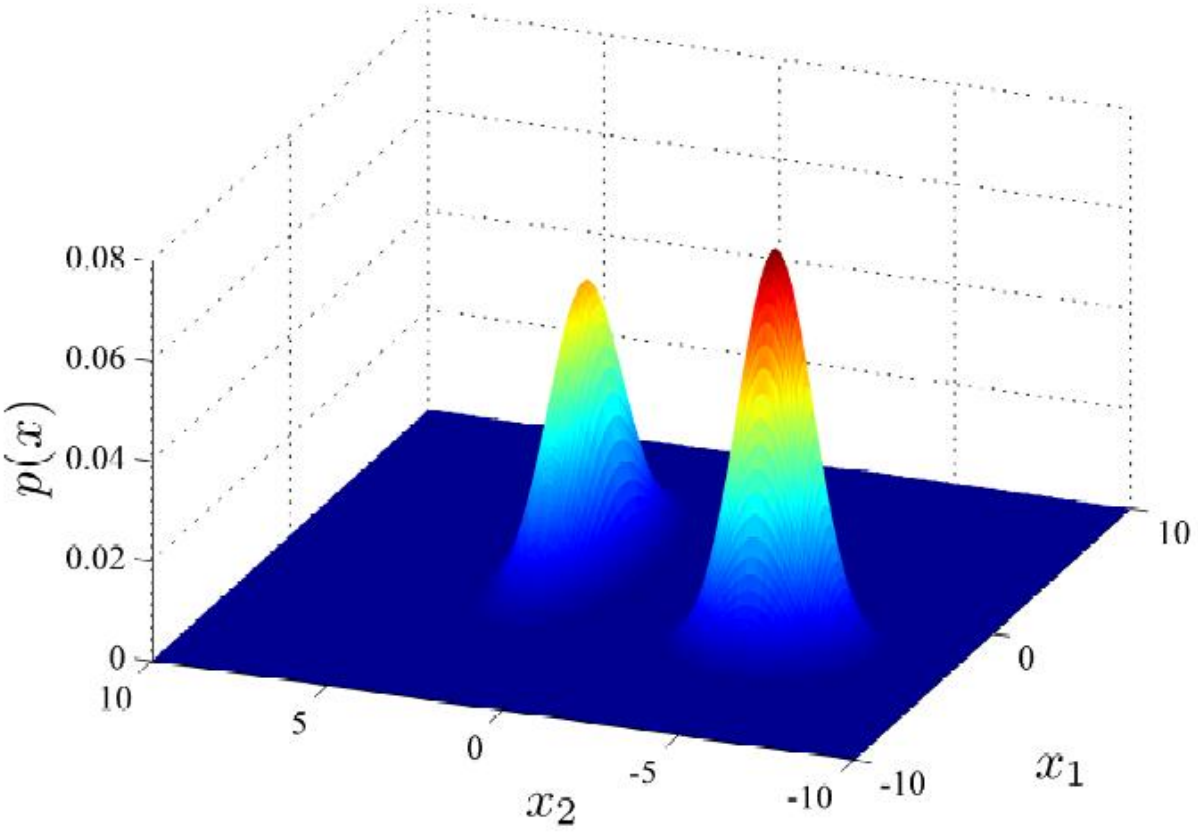
Mezcla de 3 gaussianas 1D



$$\mu_1 = -2,\ \mu_2 = 1,\ \mu_3 = 3,$$

$$\sigma_1^2 = 0.8,\ \sigma_2^2 = 0.1,\ \sigma_3^2 = 0.4,\ w_1 = w_2 = w_3 = 1/3$$

Mezcla de 2 gaussianas 2D



$$\mu_1 = [1 \;\; 2]^T, \mu_2 = [-3 \;\; -5]^T, \Sigma_1 = \mathrm{diag}(4, 0.5), \Sigma_2 = I_2, w_1 = w_2 = 0.5$$

Una proposición útil

Si X consiste de n muestras $X_1, \ldots, X_n$ i.i.d., es decir:

$$p(x \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

Entonces

$$Q(\theta \mid \theta^{(m)}) = \sum_{i=1}^{n} Q_i(\theta \mid \theta^{(m)})$$

donde

$$Q_i(\theta \mid \theta^{(m)}) = E_{X_i \mid y_i, \theta^{(m)}} \left[\log p(X_i \mid \theta) \right], \quad i = 1, \ldots, n$$

Demostración

$$
\begin{aligned}
Q(\theta \mid \theta^{(m)}) &= E_{X|y,\theta^{(m)}}[\log p(X \mid \theta)] \\
&= E_{X|y,\theta^{(m)}}\left[\log \prod_{i=1}^{n} p(X_i \mid \theta)\right] \\
&= E_{X|y,\theta^{(m)}}\left[\sum_{i=1}^{n} \log p(X_i \mid \theta)\right] \\
&= \sum_{i=1}^{n} E_{X_i|y,\theta^{(m)}}[\log p(X_i \mid \theta)] \\
&= \sum_{i=1}^{n} E_{X_i|y_i,\theta^{(m)}}[\log p(X_i \mid \theta)]
\end{aligned}
$$

Ya que las muestras son i.i.d.

Ya que

$$p(x_i \mid y, \theta^{(m)}) = p(x_i \mid y_i, \theta^{(m)})$$

Y esto es porque x_i no depende de y_j excepto cuando i=j

Derivación de EM para mezcla de gaussianas

Dadas n muestras i.i.d. $y_1, y_2, \ldots, y_n \in \mathbb{R}^d$

tomadas de una mezcla de gaussianas con parámetros $\theta = \{(w_j, \mu_j, \Sigma_j)\}_{j=1}^k$

Sea: $$\phi(y \mid \mu, \Sigma) \triangleq \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

Definimos a la probabilidad de que la i-ésima muestra pertenezca a la j-ésima gaussiana como

$$\gamma_{ij}^{(m)} \triangleq P(Z_i = j \mid Y_i = y_i, \theta^{(m)}) = \frac{w_j^{(m)}\phi(y_i \mid \mu_j^{(m)}, \Sigma_j^{(m)})}{\sum_{l=1}^k w_l^{(m)}\phi(y_i \mid \mu_l^{(m)}, \Sigma_l^{(m)})}$$

que satisface: $\sum_{j=1}^k \gamma_{ij}^{(m)} = 1$

El paso E

$$
\begin{aligned}
Q_i(\theta \mid \theta^{(m)}) &= E_{Z_i|y_i,\theta^{(m)}}\left[\log p_X(y_i, Z_i \mid \theta)\right] \\
&= \sum_{j=1}^{k} \gamma_{ij}^{(m)} \log p_X(y_i, j \mid \theta) \\
&= \sum_{j=1}^{k} \gamma_{ij}^{(m)} \log w_j \phi(y_i \mid \mu_j, \Sigma_j) \\
&= \sum_{j=1}^{k} \gamma_{ij}^{(m)} \left(\log w_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(y_i - \mu_j)^T \Sigma_j^{-1}(y_i - \mu_j) \right) + \mathrm{c}
\end{aligned}
$$

Por tanto:

$$
Q(\theta \mid \theta^{(m)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(m)} \left(\log w_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(y_i - \mu_j)^T \Sigma_j^{-1}(y_i - \mu_j) \right)
$$

El paso M

Definiendo $$n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)}$$

Podemos reescribir la función Q como

$$Q(\theta \mid \theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)} \left(\log w_j - \frac{1}{2} \log|\Sigma_j| \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(m)} (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)$$

El paso M consiste en maximizar $Q(\theta \mid \theta^{(m)})$

sujeto a $$\sum_{j=1}^{k} w_j = 1,\ w_j \geq 0,\ j = 1, \ldots, k,$$

$$\Sigma_j \succ 0,\ j = 1, \ldots, k,$$

Multiplicadores de Lagrange

- Para maximizar una función no lineal $f(x_1, x_2, \ldots, x_n)$

- Sujeta a restricciones dadas por

$$g_1(x_1, x_2, \ldots, x_n) = b_1$$
$$g_2(x_1, x_2, \ldots, x_n) = b_2$$
$$\vdots$$
$$g_m(x_1, x_2, \ldots, x_n) = b_m$$

- Formamos el Lagrangiano asociando un multiplicador de Lagrange λ con cada restricción

$$L(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{m} \lambda_i [b_i - g_i(x_1, x_2, \ldots, x_n)]$$

- Luego se resuelve el sistema de n+m ecuaciones

$$\frac{\partial L}{\partial x_j} = 0 \quad \forall j = 1,2,\ldots, n$$

$$\frac{\partial L}{\partial \lambda_j} = 0 \quad \forall j = 1,2,\ldots, m$$

Aplicando Lagrange para encontrar los pesos

Maximizar la función

$$Q(\theta \mid \theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)} \left(\log w_j - \frac{1}{2} \log|\Sigma_j| \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(m)} (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)$$

Sujeto a la restricción $\sum_{j=1}^{k} w_j = 1$

Formamos el Lagrangiano: $J(w, \lambda) = \sum_{j=1}^{k} n_j^{(m)} \log w_j + \lambda \left(\sum_{j=1}^{k} w_j - 1 \right)$

Observe que se han eliminado los términos que no depende de ω

Derivando respecto a w_j
Obtenemos k ecuaciones $\frac{\partial J}{\partial w_j} = \frac{n_j^{(m)}}{w_j} + \lambda = 0, \quad j = 1, \ldots, k.$

Derivando respecto a λ
Obtenemos otra ecuación $\sum_{j=1}^{k} w_j = 1$

Finalmente, resolvemos el sistema de k+1 ecs de donde: $w_j^{(m+1)} = \frac{n_j^{(m)}}{\sum_{j=1}^{k} n_j^{(m)}} = \frac{n_j^{(m)}}{n}, \quad j = 1, \ldots, k$

Para encontrar las medias:

$$Q(\theta \mid \theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)} \left(\log w_j - \frac{1}{2} \log|\Sigma_j| \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(m)} (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)$$

hacemos:

$$\frac{\partial Q(\theta \mid \theta^{(m)})}{\partial \mu_j} = \Sigma_j^{-1} \left(\sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i - \mu_j \sum_{i=1}^{n} \gamma_{ij}^{(m)} \right)$$

j=1,…,k

pero

$$n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)}$$

por lo tanto:

$$\frac{\partial Q(\theta \mid \theta^{(m)})}{\partial \mu_j} = \Sigma_j^{-1} \left(\sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i - n_j^{(m)} \mu_j \right) = 0, \quad j = 1, \ldots, k$$

Lo cual conduce a:

$$\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i, \quad j = 1, \ldots, k.$$

Para encontrar la matriz de covarianzas

$$Q(\theta \mid \theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)} \left(\log w_j - \frac{1}{2} \log|\Sigma_j| \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(m)} (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j)$$

Derivando:

$$\begin{aligned} \frac{\partial Q(\theta \mid \theta^{(m)})}{\partial \Sigma_j} &= -\frac{1}{2} n_j^{(m)} \frac{\partial}{\partial \Sigma_j} \log|\Sigma_j| - \frac{1}{2} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \frac{\partial}{\partial \Sigma_j} (y_i - \mu_j)^T \Sigma_j^{-1} (y_i - \mu_j) \\ &= -\frac{1}{2} n_j^{(m)} \Sigma_j^{-1} + \frac{1}{2} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \Sigma_j^{-1} (y_i - \mu_j)(y_i - \mu_j)^T \Sigma_j^{-1} \\ &= 0, \quad j = 1, \ldots, k, \end{aligned}$$

De ahí:

$$\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \left(y_i - \mu_j^{(m+1)} \right) \left(y_i - \mu_j^{(m+1)} \right)^T, \quad j = 1, \ldots, k$$

Algoritmo EM para mezcla de gaussianas

1. **Initialization:** Choose the initial estimates $w_j^{(0)}$, $\mu_j^{(0)}$, $\Sigma_j^{(0)}$, $j = 1, \ldots, k$, and compute the initial log-likelihood

$$L^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} w_j^{(0)} \phi(y_i \,|\, \mu_j^{(0)}, \Sigma_j^{(0)}) \right).$$

2. **E-step:** Compute

$$\gamma_{ij}^{(m)} = \frac{w_j^{(m)} \phi(y_i \,|\, \mu_j^{(m)}, \Sigma_j^{(m)})}{\sum_{l=1}^{k} w_l^{(m)} \phi(y_i \,|\, \mu_l^{(m)}, \Sigma_l^{(m)})}, \; i = 1, \ldots, n, \; j = 1, \ldots, k,$$

and

$$n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)}, \; j = 1, \ldots, k.$$

3. **M-step:** Compute the new estimates

$$w_j^{(m+1)} = \frac{n_j^{(m)}}{n}, \; j = 1, \ldots, k,$$

$$\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i, \; j = 1, \ldots, k,$$

$$\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \left(y_i - \mu_j^{(m+1)} \right) \left(y_i - \mu_j^{(m+1)} \right)^T, \; j = 1, \ldots, k.$$

4. **Convergence check:** Compute the new log-likelihood

$$L^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} w_j^{(m+1)} \phi(y_i \,|\, \mu_j^{(m+1)}, \Sigma_j^{(m+1)}) \right).$$

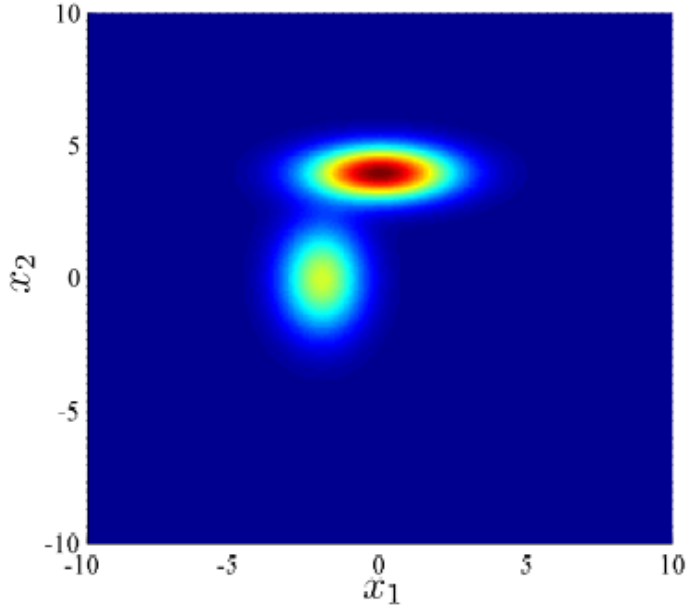
Return to step 2 if $|L^{(m+1)} - L^{(m)}| > \delta$ for a preset threshold δ; otherwise end the algorithm.

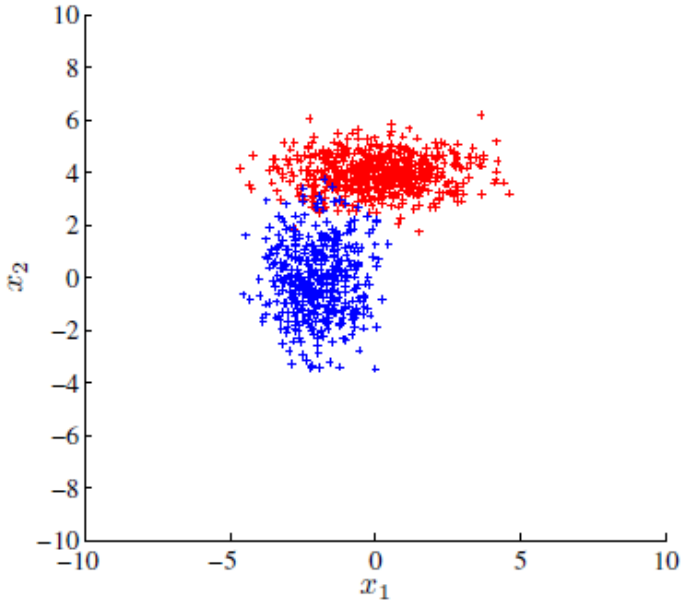
Ejemplo

Considere una mezcla de 2 gaussianas 2D con los siguientes parámetros:

$$\mu_1 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \ \Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \ w_1 = 0.6, \ w_2 = 0.4$$

Se generan 1000 valores aleatorios de esta distribución





Solución

Usamos k-medias para determinar los centroides y usarlos como primera aproximación

$$\mu_1^{(0)} = \begin{bmatrix} 0.0823 \\ 3.9189 \end{bmatrix}, \quad \mu_2^{(0)} = \begin{bmatrix} -2.0706 \\ -0.2327 \end{bmatrix}$$

Elegimos $w_1^{(0)} = w_2^{(0)} = 0.5$ y $\Sigma_1^{(0)} = \Sigma_2^{(0)} = I_2$

también $\delta = 10^{-3}$ Y después de solo tres iteraciones encontramos:

$$\mu_1^{(3)} = \begin{bmatrix} 0.0806 \\ 3.9445 \end{bmatrix}, \quad \mu_1^{(3)} = \begin{bmatrix} -2.0181 \\ -0.1740 \end{bmatrix}, \quad \Sigma_1^{(3)} = \begin{bmatrix} 2.7452 & 0.0568 \\ 0.0568 & 0.4821 \end{bmatrix}, \quad \Sigma_2^{(3)} = \begin{bmatrix} 0.8750 & -0.0153 \\ -0.0153 & 1.7935 \end{bmatrix}$$

$$w_1^{(3)} = 0.5966 \qquad w_2^{(3)} = 0.4034$$

Tres iteraciones