Apéndice B

Algoritmo EM

El algoritmo EM, inicialmente propuesto por Dempster (Dempster, Laird, y Rubin 1977), presenta un técnica iterativa general para realizar una estimación de máxima verosimilitud de parámetros de problemas en los que existen ciertos *datos ocultos*. Presentamos a continuación el algoritmo, y un ejemplo de su aplicación a la resolución de un problema concreto, tomando (Mitchel 1997) como referencia.

B.1 Descripción del algoritmo EM

El algoritmo EM puede aplicarse en muchas situaciones en las que se desea estimar un conjunto de parámetros θ que describen una distribución de probabilidad subyacente, dada únicamente una parte observada de los datos completos producidos por la distribución. En general, supongamos que en cada realización del experimento aleatorio se observa un parámetro z_i y existe un parámetro oculto x_i. Denotamos entonces por $Z = \{z_1, \ldots, z_m\}$ al conjunto de datos observados en m realizaciones del experimento, por $X = \{x_1, \ldots, x_m\}$ al conjunto de datos no observados y por $Y = Z \cup X$ al conjunto completo de datos. Los datos X pueden considerarse una variable aleatoria cuya distribución de probabilidad depende de los parámetros a estimar θ y de los datos observados Z. De la misma forma, Y es una variable aleatoria ya que está definida en términos de la variable aleatoria X. Llamemos h a la hipótesis actual de los valores de los parámetros θ, y denotemos por h' la hipótesis revisada que se estima en cada iteración del algoritmo EM.

El algoritmo EM busca la hipótesis h' que maximiza la esperanza $E[\ln p(Y \mid h')]$, siendo $p(Y \mid \theta)$ la distribución de probabilidad que define Y y que depende de los parámetros desconocidos θ. Esta distribución de probabilidad define la verosimilitud de los datos completos Y dada una hipótesis h' de los parámetros ocultos. Al maximizar el logaritmo de la distribución se está maximizando la verosimilitud. Se introduce el valor esperado $E[\ln p(Y \mid h')]$

debido a que el conjunto completo de datos Y es una variable aleatoria. Dado que el conjunto completo de datos Y contiene datos X no observados, se deben considerar todos los posibles valores de X, ponderándolos según su probabilidad. En otras palabras, se calcula el valor esperado $E[\ln p(Y \mid h')]$ sobre la distribución de probabilidad que gobierna la variable aleatoria Y. Esta distribución está determinada por los valores observados Z más por la distribución de los valores no observados X.

En general, se desconoce la distribución de Y, porque está determinada por los parámetros θ que se intenta estimar. Por ello, el algoritmo EM usa la hipótesis actual h para estimar la distribución de Y. Se define entonces una función $Q(h \mid h')$ que proporciona $E[\ln p(Y \mid h')]$ como una función de h', bajo la suposición de que $\theta = h$ y dada el conjunto de observaciones Z del conjunto completo de datos Y

$$Q(h' \mid h) = E[\ln p(Y \mid h') \mid h, Z].$$

En la función $Q(h' \mid h)$ se supone que la hipótesis h y los datos observados Z tienen unos valores fijos y que éstos definen la distribución de probabilidad de las variables ocultas X (y, por tanto, sus valores esperados). La distribución de probabilidad de Y definida por Z y h es, entonces, la que se utiliza para calcular $E[\ln p(Y \mid h')]$ para una hipótesis cualquiera h'. En su forma general, el algoritmo EM repite la siguiente pareja de pasos hasta que converge.

Paso 1: *Paso de estimación (E)*: Calcular $Q(h' \mid h)$ utilizando la hipótesis actual h y los datos observados Z para estimar la distribución de probabilidad de Y

$$Q(h' \mid h) \leftarrow E[\ln p(Y \mid h') \mid h, Z]. \tag{B.1}$$

Paso 2: *Paso de maximización (M)*: Sustituir h por la hipótesis h' que maximiza la función Q

$$h \leftarrow \arg\max_{h'} Q(h' \mid h). \tag{B.2}$$

B.2 Aplicación a la estimación de k medias

Para ilustrar el funcionamiento del algoritmo EM, vamos a utilizarlo para derivar un algoritmo que estime las medias de una mezcla de k distribuciones normales $\theta = (\mu_1, \ldots, \mu_k)$ con igual desviación típica σ, que se supone conocida. Los datos observados $Z = \{z_j\}$ son los datos producidos por la distribución. Los datos no observados

$$X = \{(x_{1j}, \ldots, x_{kj})\}, \qquad x_{ij} = (0, 1), \qquad \sum_{i=1}^{k} x_{ij} = 1$$

indican cuál de las k distribuciones normales se ha utilizado para obtener el dato z_j.

Para aplicar EM primero se necesita derivar una expresión de $Q(h \mid h')$ para el problema. Derivemos primero la formulación de $\ln p(Y \mid h')$. Para un único conjunto de datos $y_j = (z_j, x_{1j}, \ldots, x_{kj})$, la verosimilitud de que estos datos hayan sido obtenidos con una hipótesis $h' = (\mu'_1, \ldots, \mu'_k)$ se puede escribir como

$$p(y_j \mid h') = p(z_j, x_{1j}, \ldots, x_{kj} \mid h') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{k} x_{ji}(z_j - \mu'_i)^2\right). \quad \text{(B.3)}$$

Esta expresión proporciona la probabilidad de que el valor z_j haya sido generado por la distribución normal seleccionada por los datos ocultos. La probabilidad para todos las instancias m de los datos es

$$\begin{aligned} \ln p(Y \mid h') &= \ln \prod_{i=j}^{m} p(y_j \mid h') = \\ &= \sum_{j=1}^{m} \ln p(y_j \mid h') = \\ &= \sum_{j=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} x_{ij}(z_j - \mu'_i)^2\right) \end{aligned} \quad \text{(B.4)}$$

Por último, se debe calcular el valor esperado de esta expresión $\ln p(Y \mid h')$ sobre toda la distribución de probabilidad que gobierna Y o, de forma equivalente, sobre la distribución de los datos ocultos de Y, x_{ij}. Al ser la expresión anterior una expresión linear en función de estos datos, es posible derivar la siguiente expresión

$$\begin{aligned} E[\ln p(Y \mid h')] &= E\left[\sum_{j=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} x_{ij}(z_j - \mu'_i)^2\right)\right] \\ &= \sum_{j=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} E[x_{ij}](z_j - \mu'_i)^2\right) \end{aligned} \quad \text{(B.5)}$$

Para resumir, la función $Q(h \mid h')$ del problema de las k medias es

$$Q(h' \mid h) = \sum_{j=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} E[x_{ij}](z_j - \mu'_i)^2\right), \quad \text{(B.6)}$$

donde $h' = (\mu'_1, \ldots, \mu'_k)$ y donde los valores esperados de los datos ocultos $E[x_{ij}]$ se calculan a partir de la hipótesis actual y a los datos observados Z. Este valor $E[x_{ij}]$ es simplemente la probabilidad de que la muestra z_j haya sido generada por la distribución normal i

$$\begin{aligned} E[x_{ij}] &= \frac{p(x = z_j \mid \mu = \mu_i)}{\sum_{n=1}^{k} p(x = z_j \mid \mu = \mu_n)} = \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2}(z_j - \mu_i)^2\right)}{\sum_{n=1}^{k} \exp\left(-\frac{1}{2\sigma^2}(z_j - \mu_n)^2\right)} \end{aligned} \tag{B.7}$$

Esta ecuación completa el primer paso del algoritmo EM, en el que se define la función Q a partir de los datos ocultos esperados. El segundo paso (maximización) consiste en encontrar los valores $(\mu'_1, \ldots, \mu'_k)$ que maximizan la función Q así definida. En este caso,

$$\begin{aligned} &\arg\max_{h'} Q(h' \mid h) = \\ &\quad = \arg\max_{h'} \sum_{j=1}^{m} (\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} E[x_{ij}](z_j - \mu'_i)^2) = \\ &\quad = \arg\min_{h'} \sum_{i=1}^{k} \sum_{j=1}^{m} E[x_{ij}](z_j - \mu'_i)^2 \end{aligned} \tag{B.8}$$

Esto es, la hipótesis de máxima verosimilitud es la que minimiza la suma ponderada de los errores al cuadrado, donde la contribución de cada instancia z_j al error que define μ'_i está ponderada por $E[x_{ij}]$. Esta hipótesis se puede calcular de forma analítica con la siguiente expresión

$$\mu_j \leftarrow \frac{1}{m} \sum_{j=1}^{m} E[x_{ij}] z_j. \tag{B.9}$$