El algoritmo EM para las estimacion de parametros en mezclas gaussianas

Edgar Acuna.

Una mezcla de distribuciones con K componentes tiene la forma

$$f(x) = \pi_1 f_1(x) + ... + \pi_K f_K(x)$$

donde cada componente es una función de probabilidad (Poisson, Binomial, etc) o una funcion de densidad (Normal, Exponencial, Gamma, etc). Notar que
$\pi_1 + ... + \pi_K = 1.$

De todas las mezclas la mas usada es la mezcla de densidades normales, llamada también una Mezcla Gaussiana , que tiene la forma general

$$f(x) = \sum_{i=1}^{K} \pi_i \frac{\exp(-(x-\mu_i)^2/2\sigma_i^2)}{\sigma_i\sqrt{2\pi}} \qquad (1)$$

en esta mezcla habria que estimar 3K parámetros(K medias, K varianzas y K coeficientes). Como la suma de los coeficientes da 1. el numero de parámetros a estimar se puede reducir a 3K-1.

1. Estimacion de parámetros en una mezcla Gaussiana con dos componentes. (caso Unidimensional)

En este caso la mezcla es de la forma

$$f_Y(y) = (1-\pi)\frac{e^{-(x-\mu_1)^2/2\sigma_1^2}}{\sigma_1\sqrt{2\pi}} + \pi\frac{e^{-(x-\mu_2)^2/2\sigma_2^2}}{\sigma\sqrt{2\pi}}$$

sean $\theta_1=(\mu_1,\sigma_1^2)$, $\theta_2=(\mu_2,\sigma_2^2)$ y $\theta=(\pi,\theta_1,\theta_2)$ vectores de parámetros. Entonces, la mezcla anterior se puede escribir como

$$f_Y(y,\theta) = (1-\pi)N(y,\theta_1) + \pi N(y,\theta_2)$$

donde N representa a la función de densidad Normal o gaussiana. Notar que el vector de parámetros θ contiene 5 parametrso que deben ser estimados usando una muestra aleatoria de tamaño n de la mezcla gaussiana.
Sea $y_1,y_2,.....y_n$ la muestra aleatoria tomada de la distribución de Y. El metodo mas usado para estimar el vector de parámetros θ es el Maximum Likelihood (ML) (maxima

verosimilitud. Aquí se maximiza con respecto a θ, la función de verosimilitud definida por

$$L(\theta/Y)=f(y_1,\theta)f(y_2,\theta)\ldots\ldots f(y_n,\theta)=\prod_{i=1}^{n} f(y_i,\theta) \qquad (2)$$

Alguna gente usa solo la notación L(θ) y otros L(θ;y), pero en EM se prefiere la notación (2). Frecuentemente es mas facil trabajar con Log(L(θ/Y)), que debido a la monoticidad del la funcion Logaritmica, preserva el maximo. Así

$$\log(L(\theta/Y)) = \log[\prod_{i=1}^{n} f(y_i,\theta) = \sum_{i=1}^{n} \log f(y_i,\theta)$$

En el caso de la mezcla gaussiana con dos componentes, se tendrá

$$\log(L(\theta/Y)) = \sum_{i=1}^{n} \log[(1-\pi)N(y_i,\theta_1) + \pi N(y_i,\theta_2)] \quad (3)$$

donde $\log(N(y,\theta_1)) = -\frac{(y-\mu_1)^2}{2\sigma_1^2} - .5\log(\sigma_1^2) - .5\log(2\pi)$, similarmente para log(N(y,θ_2)). Habria que maximizar (3) con respecto a θ=(π,θ_1,θ_2) =(π,μ_1,σ_1^2,μ_2,σ_2^2), hallando las respectivas derivadas parciales, pero la presencia de la funcion logaritmo hace que esto sea complicado. La alternativa mas usada es el algoritmo EM introducido por Dempster, Laird, and Rubi (1977).

Notar que la variable aleatoria Y correspondiente a la mezcla Gausiana puede ser obtenida como $Y=(1-T)*Y_1+T*Y_2$, donde Y1~N(μ_1,σ_1^2) , Y2~N(μ_2,σ_2^2) y T es una variable aleatoria que asume los valores 0 y 1 con P[T=1]=π. Pero notar que los valores de T no son observados en la muestra.

El algoritmo EM se basa en el siguiente hecho: Log(L(θ/Y) es complicado de resolver, pero log(L(θ/Y,T) si se puede resolver. Y es llamada la data observada, T es llamada la data no observada y Z=(Y,T) es llamada la data completa. Consideremos que para la i-esima observación y_i de la muestra, T_i=1 si dicha observación es extraída de la población Y_2 y T_i=0 si la observación es extraída de la población Y_1. Entonces la función de verosimilitud correspondiente, puede ser escrita de la forma

$$L(\theta/Y,T) = \prod_{i=1}^{n} [N(y_i,\theta_1)]^{1-T_i}[N(y_i,\theta_2)]^{T_i} \quad (4)$$

Tomando logaritmos se tendría

$$LogL(\theta/Y,T) = \sum_{i=1}^{n} (1-T_i)\log(N(y_i,\theta_1)) + T_i \log(N(y_i,\theta_2)) \quad (5)$$

La ecuación (5) si puede ser maximizada mas fácilmente que la ecuación (3). Pero el proceso debe ser iterativo, porque en cada paso el valor de T_i depende de los valores de π, θ_1 y θ_2.

El paso E (calculo del valor esperado)
Notar que si en (4) tomamos valor esperado condicionado a (θ,Y), resulta

$$E[Q(\theta)/Y,\theta] = \sum_{i=1}^{n}(1 - E(T_i/\theta,Y))\log(N(y_i,\theta_1)) + E(T_i/\theta,Y)\log(N(y_i,\theta_2))$$

donde Q(θ)=log(L(θ)/Y,T). Pero,

$E(T_i/\theta,Y)=0*P[T_i=0/\theta,Y]+1*P[T_i=1/\theta,Y]$
$= P[T_i=1/\theta,Y]=Prob(caer\ en\ Y2)/Prob(caer\ Y1\ o\ Y2)$

Consideremos ahora el proceso iterativo, que comienza considerando $\pi^{(0)}$=.5, $\mu_1^{(0)} = random\ (y_i), \mu_2^{(0)} = random\ (y_i), \sigma_1^2 = S^2, \sigma_2^2 = S^2)$, donde S^2 es la varianza muestral. Luego en el m-esimo paso,

$$\gamma_i^{(m)} = E[T_i/\theta^{(m)},Y] = \frac{\pi^{(m)}N(y_i,\theta_2^{(m)})}{(1-\pi^{(m)})N(y_i,\theta_1^{(m)}) + \pi^{(m)}N(y_i,\theta_2^{(m)})}$$

Estos valores son llamados las *responsabilidades* y se usan para asignar las observaciones a un cluster. Asi al acabar el proceso iterativo si $\gamma_i \geq .5$ se asigna la observación al cluster 2 de lo contrario iria al cluster 1.
Luego lo que habría que maximizar se reduce a

$$Q(\theta/\theta^{(m)}) = \sum_{i=1}^{n}(1-\gamma_i^{(m)})\log(N(y_i,\theta_1)) + \gamma_i^{(m)}\log(N(y_i,\theta_2) \quad (6)$$

Alguna gente prefiere la notación $Q(\theta,\theta^{(m)})$. Aquí termina la etapa E,

La etapa M (Maximización)

Aquí se maximiza la expresión (6) con respecto a $\theta=(\pi,\theta_1,\theta_2) =(\pi,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$, esto implica calcula derivadas parciales de Q con respecto a $\pi,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2$ e igualarlas a 0.

Así derivando Q con respecto a μ_1 se obtiene.

$\frac{\partial Q}{\partial \mu_1} = \sum_{i=1}^{n}(1-\gamma_i^{(m)})\frac{2(y_i-\mu_1)}{2\sigma_1^2} = 0$, lo cual produce

$\hat{\mu}_1^{(m+1)} = \frac{\sum_{i=1}^{n}(1-\gamma_i^{(m)})y_i}{\sum_{i=1}^{n}(1-\gamma_i^{(m)})}$. Similarmente, $\hat{\mu}_2^{(m+1)} = \frac{\sum_{i=1}^{n}\gamma_i^{(m)}y_i}{\sum_{i=1}^{n}\gamma_i^{(m)}}$

Derivando ahora con respecto a σ_1^2 se obtiene que

$\frac{\partial Q}{\partial \sigma_1^2} = \frac{\sum_{i=1}^{n}(1-\gamma_i^{(m)})(y_i-\hat{\mu}_i^{(m+1)})^2}{2\sigma_1^4} - \frac{\sum_{i=1}^{n}(1-\gamma_i^{(m)})}{2\sigma_1^2} = 0$, de donde resulta

$\sigma_1^{2(m+1)} = \frac{\sum_{i=1}^{n}(1-\gamma_i^{(m)})(y_i-\hat{\mu}_1^{(m+1)})^2}{\sum_{i=1}^{n}(1-\gamma_i^{(n)})}$. Similarmente,

$\sigma_2^{2(m+1)} = \frac{\sum_{i=1}^{n}\gamma_i^{(m)}(y_i-\hat{\mu}_2^{(m+1)})^2}{\sum_{i=1}^{n}\gamma_i^{(n)}}$.

Finalmente, por definición del γi, se tiene que $\hat{\pi}^{(m+1)} = \bar{\gamma}^{(m)} = \frac{\sum_{i=1}^{n}\gamma_i^{(m)}}{n}$.

El algoritmo se repite hasta que $\pi^{(m)}, \mu_1^{(m)}, \sigma_1^{2(m)}, \mu_2^{(m)}, \sigma_2^{2(m)}$ no cambie mucho con respecto al paso anterior. Habria que establecer un nivel de tolerancia pequeño.

2. Estimacion de parámetros en una mezcla Gaussiana con K componentes. (caso Unidimensional)

La funcion de densidad esta dada en la ecuación (1). Aquí introducimos K variables aleatorias Tj (j=1…K) tal que Tj= 1 si la observación es extraida de la población Yj, y Tj=0 en otro caso, y P[Tj=1]=πj. Notar que $\pi_K = 1 - \sum_{i=1}^{K-1}\pi_i$.

Luego, $Y=T_1Y_1+\ldots\ldots\ldots T_KY_K$. La función de log likelihood, condicionada a la data completa, similar a la ecuación 5 puede se escrita como

$$LogL(\theta / Y,T) = \sum_{i=1}^{n}\sum_{j=1}^{K} T_{ij} \log(N(y_i,\theta_j)) \quad (7)$$

donde T_{ij}=1 si la i-ésima observación yi cae en la componente j y T_{ij}=0 en otro caso.

Paso E.

Aquí las T_{ij} que aparecen en la ecuación (7) deben ser sustituidas iterativamente por

$\gamma_{ij}^{(m)} = E[T_{ij} / \theta^{(m)}, Y] = \dfrac{\hat{\pi}_j^{(m)} N(y_i, \theta_j^{(m)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(m)} N(y_i, \theta_l^{(m)})}$. Los valores iniciales de los parámetros son los mismos que antes, excepto que $\pi_j^{(0)} = \dfrac{1}{K}$ para j=1….K.

Luego habría que maximizar la función

$$Q(\theta / \theta^{(m)}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}^{(m)} \log(N(y_i,\theta_j)) \quad (8)$$

Paso M:

Derivando parcialmente (8) con respecto a μ_j y σ_j^2 para j=1,…,K, se obtiene

$$\hat{\mu}_j^{(m+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i}{\sum_{i=1}^{n} \gamma_{ij}^{(m)}} \quad \text{y} \quad \sigma_j^{2(m+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)} (y_i - \hat{\mu}_j^{(m+1)})^2}{\sum_{i=1}^{n} \gamma_{ij}^{(n)}} \quad \text{y}$$

$$\hat{\pi}_j^{(m+1)} = \bar{\gamma}_j^{(m)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)}}{n}$$

3. Caso Multivariado.

Consideremos que tomamos p variables distribuidas conjuntamente con una normal Multivariada con vector de media **μ** y matriz de covarianza **Σ**. Esto es que la función de densidad del vector aleatorio **x** tiene función de densidad

$$f(x) = \frac{\exp[-(x-\mu)'\Sigma^{-1}(x-\mu)/2]}{|\Sigma|^{1/2} (2\pi)^{p/2}}$$

y se escribe **x**~NM_p(**μ**, **Σ**) Existen muchas estructuras que se pueden considerar para Σ. Entre ellas que sea una matriz diagonal con una constante. Es decir, $\Sigma=\sigma^2 I$, o que sea solamente diagonal con diferente entradas en ella.
Sea la variable aleatoria Y que se distribuye como una mezcla de K Gaussianas multivariadas. Esto es,

$$f(y) = \pi_1 NM(\mu_1, \Sigma_1) + ... + \pi_K NM(\mu_K, \Sigma_K)$$

donde $\sum_{l=1}^{K} \pi_l = 1$

Introduciendo el mismo tipo de variables T_j como en el caso 2, se obtiene

$$LogL(\theta / Y, T) = \sum_{i=1}^{n} \sum_{j=1}^{K} T_{ij} \log(NM(y_i, \theta_j)) \quad (9)$$

donde $\theta_j = (\mu_j, \Sigma_j)$ para j=1,….K.

Paso E.
Aquí las T_{ij} que aparecen en la ecuación (9) deben ser sustituidas iterativamente por

$\gamma_{ij}^{(m)} = E[T_{ij} / \theta^{(m)}, Y] = \dfrac{\hat{\pi}_j^{(m)} NM(y_i, \theta_j^{(m)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(m)} NM(y_i, \theta_l^{(m)})}$. Los valores iniciales para las k medias μj son k observaciones de las y_i's elegidas al azar. Los valores iniciales de las matrices de covarianza depende de la estructura que se halla decidido para la matriz. El programa mclust de R considera 6 tipos distintos de estructuras. También, $\pi_j^{(0)} = \dfrac{1}{K}$ para j=1….K.
Luego habría que maximizar la función

$$Q(\theta / \theta^{(m)}) = \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_{ij}^{(m)} \log(NM(y_i, \theta_j)) \quad (10)$$

Paso M.

Derivando parcialmente (10) con respecto a μ_j y **Σj** para j=1,…,K, se obtiene

$$\hat{\mu}_j^{(m+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i}{\sum_{i=1}^{n} \gamma_{ij}^{(m)}} \quad \text{y} \quad \Sigma_j^{(m+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)} (y_i - \hat{\mu}_j^{(m+1)})'(y_i - \hat{\mu}_j^{(m+1)})}{p \sum_{i=1}^{n} \gamma_{ij}^{(n)}} \text{ y}$$

$$\hat{\pi}_j^{(m+1)} = \bar{\gamma}_j^{(m)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(m)}}{n}$$