Datos incompletos - algoritmo EM

Escuela Santaló, Julio 2012

Isaac Meilijson, Tel Aviv University

Desigualdad de Jensen:
Para f cóncava y v.a. X, $E[f(X)] \leq f(E[X])$

Desigualdad de Información:
Para densidades f y g,
$\int f(x) \log(g(x))dx \leq \int f(x) \log(f(x))dx$

Kullback-Leibler Divergence KLD:
$KLD = \int f(x) \log(\frac{f(x)}{g(x)})dx$

Estimador de Máxima Verosimilitud MLE

$$
\begin{aligned}
LIK = \prod_x p(x)^{n(x)} &= \exp\{n \sum_x \frac{n(x)}{n} \log(p(x))\} \\
&= \exp\{n \sum_x \hat{p}(x) \log(p(x))\} \\
&\approx \exp\{n \sum_x p_0(x) \log(p(x))\} \\
&= K * \exp\{-nKLD(p_0, p)\}
\end{aligned}
$$

Vemos el papel del KLD y vemos por qué MLE es **consistente**:
LIK es cercano a una función determinística que tiene su único máximo en **"la verdad"** p_0.

Vemos Y (**dato incompleto**) con modelo complicado pero $Y = Y(X)$ para X (**dato completo**) con modelo sencillo.

Ejemplo 1: Riesgos en competencia
Longevidad de componentes $X = (X_1, X_2, \ldots, X_k)$ son independientes
Longevidad de la máquina y **causa de fallo**
$Y = (min_i X_i,$ nombre del minimizador$)$
Problema: Estimar las distribuciones de los X_i

Ejemplo 2: Mezcla de distribuciones
X =(altura, sexo), Y=altura
Altura de persona de sexo $i \sim N(\mu_i, \sigma^2)$
Problema: Estimar $\mu_1, \mu_2, \sigma, \alpha$ (**proporción** de mujeres)

Siendo Y función de X, $\forall x$ tal que $Y(x) = y$,
$f_X(x;\theta) = f_Y(y;\theta) f_{X|Y=y}(x;\theta)$
de lo cual
$\log f_Y(y;\theta) = \log f_X(x;\theta) - \log f_{X|Y=y}(x;\theta)$
Como no depende de x, podemos tomar esperanza c/r a cualquier distribución sobre los x consistentes con y, e.g. la distribución condicional de X dado $Y = y$ para valor arbitrario θ_0 del parámetro θ.

$$
\begin{aligned}
\log f_Y(y;\theta) &= E[\log f_X(X;\theta)|Y = y;\theta_0] \\
&- E[\log f_{X|Y=y}(X;\theta)|Y = y;\theta_0] \\
&= Q(\theta,\theta_0) - H(\theta,\theta_0)
\end{aligned}
$$

La diferencia entre el caso θ y el caso θ_0

$$\begin{aligned}
\log f_Y(y;\theta) - \log f_Y(y;\theta_0) &= [Q(\theta,\theta_0) - Q(\theta_0,\theta_0)] \\
&- [H(\theta,\theta_0) - H(\theta_0,\theta_0)] \\
&\geq Q(\theta,\theta_0) - Q(\theta_0,\theta_0)
\end{aligned}$$

debido a la desigualdad de la información.

Si logramos encontrar θ con $Q(\theta,\theta_0) > Q(\theta_0,\theta_0)$, el (no calculado) log verosimilitud escala aún más. Se reemplaza θ_0 por el nuevo θ y se itera el reemplazo:

Algoritmo EM (Expectation - Maximization)

$$\frac{\partial Q(\theta, \theta_n)}{\partial \theta}|_{\theta=\theta_{n+1}} = 0$$

El (?) límite $\hat{\theta} = \theta_\infty$ es MLE de θ, porque

$$\frac{\partial Q(\theta, \hat{\theta})}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

y para todo θ_0, debido a la desigualdad de información,

$$\frac{\partial Q(\theta, \theta_0)}{\partial \theta}|_{\theta=\theta_0} = \frac{\partial \log f_Y(y; \theta)}{\partial \theta}|_{\theta=\theta_0}$$

Distribuciones de tipo exponencial

$$f_X(x;\theta) = h(x)\exp\{\theta t(x) - b(\theta)\}$$

$$\frac{\partial \log f_X(x;\theta)}{\partial\theta} = t(x) - b'(\theta)$$

$$\frac{\partial Q(\theta,\theta_0)}{\partial\theta} = E[t(X)|Y = y;\theta_0] - b'(\theta)$$

Se reemplaza el no observado $t(X_i)$ por su esperanza $E[t(X)|Y = y_i;\theta_0]$ dado lo que sabemos y se lo trata como si fuera $t(X_i)$

Pero no en forma final sino como etapa iterativa

Mezcla de distribuciones

LIK dato **incompleto**: $\alpha f_0(y) + (1-\alpha) f_1(y)$

LIK dato **completo**: $(\alpha f_0(y))^I ((1-\alpha) f_1(y))^{(1-I)}$

cuyo logaritmo acumulado es

$log(\alpha)\, \Sigma_{i=1}^{n} I_i + log(1-\alpha)\, \Sigma_{i=1}^{n}(1-I_i)$

$+\, \Sigma_{i=1}^{n} I_i \log f_0(y_i) + \Sigma_{i=1}^{n}(1-I_i) \log f_1(y_i)$

Lo único que necesitaremos es

$E[I|Y=y;\theta_0] = \frac{\alpha_0 f_0(y;\theta_0)}{\alpha_0 f_0(y;\theta_0) + (1-\alpha_0) f_1(y;\theta_0)}$

El nuevo α es sencillamente

$$\frac{1}{n}\sum_{i=1}^{n} E[I|Y=y_i;\theta_0]$$

y los demás parámetros se actualizan resolviendo

$$\begin{aligned}
&\sum_{i=1}^{n} E[I|Y=y_i;\theta_0]\frac{\partial}{d\theta}\log f_0(y_i)\\
+&\sum_{i=1}^{n}(1-E[I|Y=y_i;\theta_0])\frac{\partial}{d\theta}\log f_1(y_i)=0
\end{aligned}$$

References

[1] Fernández-Busnadiego, R., Zuber, B., Maurer,U. E., Cyrklaff, M., Baumeister, W. & Lučić, V. (2010). Quantitative analysis of the native presynaptic cytomatrix by cryoelectron tomography. *J. Cell. Biol.* **188(1)**, 145–156.

[2] Dempster, A.P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from Incomplete Data via the EM algorithm. *J. R. Stat. Soc.: Series B*, **39**, 1–38.

[3] Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Stat. Soc.: Series B*, **51**, 127–138.

[4] Nitzany, E., Hammel, I. & Meilijson, I. (2010). Quantal basis of vesicle growth and information content, a unified approach. *J. Theor. Biol.* **266**, 202–209.

[5] Hammel, I. & Meilijson, I. (2012). Function suggests nano-structure: electrophysiology suggests that granule membranes play dice. *J. R. Soc. Interface.*