

# Algoritmo EM

Tópicos de Investigación: Machine Learning

Luis Alberto Evangelista

Universidad Nacional de Ingeniería

05 de Octubre

# Introducción

El algoritmo EM (Expectation-Maximization) es una técnica de optimización originalmente introducida por Dempster, Laird y Rubin (1997), en su publicación *Maximun Likelihood from Imcomplete Data via the EM algorithm*.

Se utiliza en estadística para encontrar estimadores de verosimilitud (VM) de parámetros en modelos probabilísticos que dependen de variables no observables (datos perdidos).

# Introducción

- El algoritmo EM alterna pasos de esperanza (paso E), donde se calcula la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables,
- y un paso de Maximización (paso M), donde se calculan estimadores de VM de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentra en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite.

# Preliminares: Desigualdad De Jensen

Sea  $f$  una función cuyo dominio es el conjunto de los números reales. Sabemos que  $f$  es función convexa si  $f''(x) \geq 0$ , para todo  $x \in \mathbb{R}$ . En caso que  $f$  sea función real en varias variables, la función  $f$  es convexa si su Hessiana  $H$  es semidefinida positiva ( $H \geq 0$ ).

Si  $f''(x) > 0$  para todo  $x \in \mathbb{R}$  ( Hessiana  $H$  de  $f$  es definida positiva:  $H > 0$ ),decimos que  $f$  es estrictamente convexa.

La desigualdad de Jensen puede expresarse de la siguiente forma, escrito como el siguiente teorema

# Desigualdad de Jensen

## Teorema

*Sea  $f$  una función y  $X$  una variable aleatoria entonces*

$$f(E[X]) \leq E[f(X)]$$

*Además si  $f$  es estrictamente convexa entonces:*

*$E[f(X)] = f(E[X])$  si y sólo si  $X = E[X]$  con probabilidad 1 (i.e  $X$  es constante).*

Proof.

Proof.

Veamos en el caso que

## Proof.

Veamos en el caso que

1  $X$  sea finito, en efecto:

Sea  $p_i = P\{X = x_i\}$  entonces  $E[X] = \sum_{i=1}^n p_i x_i$  es una combinación convexa de valores de  $X$ .

Como  $X$  es una función convexa entonces

$$f(E[X]) = f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i) = E[f(X)]$$



## Proof.

Veamos en el caso que

1  $X$  sea finito, en efecto:

Sea  $p_i = P\{X = x_i\}$  entonces  $E[X] = \sum_{i=1}^n p_i x_i$  es una combinación convexa de valores de  $X$ .

Como  $X$  es una función convexa entonces

$$f(E[X]) = f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i) = E[f(X)]$$

2  $X$  sea numerable, en efecto:

Como  $X$  toma una cantidad numerable de valores  $x_i$  con probabilidad  $p_i$ .

Definimos  $s_n = \sum_{i=1}^n p_i$ , para cada  $n \in \mathbb{N}$



Notamos que  $\sum_{i=1}^n \frac{p_i}{s_n} x_i$  es una combinación convexa.

Luego como  $f$  es convexa entonces

$$f\left(\sum_{i=1}^n \frac{p_i}{s_n} x_i\right) \leq \sum_{i=1}^n \frac{p_i}{s_i} f(x_i)$$

Luego tomando limite cuando  $n \rightarrow \infty$ ,  $s_n \rightarrow 1$  y de la continuidad de  $f$  tenemos:

$$f(E[X]) = f\left(\sum_{i=1}^{\infty} p_i x_i\right) \leq \sum_{i=1}^{\infty} p_i f(x_i) = E[f(X)]$$

## Recordar :

$f$  es concava (estrictamente concava) si y sólo si  $-f$  es convexa (estrictamente convexa) i.e.  $f(x) \leq 0, H \leq 0 (f(x) < 0, H < 0)$

El teorema puede reescribirse en términos de la función cóncava de la siguiente forma:

### Teorema

*Sea  $f$  una función cóncava y  $X$  una variable aleatoria entonces*

$$E[f(X)] \leq f(E[X])$$

# Algoritmo EM

Se supone que  $X = (X_1, X_2, \dots, X_n)$  es una variable aleatoria con distribución conjunta de  $g(X | \theta)$  y se quiere calcular

$$\hat{\theta} = \arg\max_{\theta} L(\theta | X)$$

Donde  $L(\theta | X) = g(X | \theta)$ .

Consideremos los datos completos  $W$  provenientes de una muestra aleatoria constituida por  $W = (X, Z)$ , donde  $W$  representa los datos completos,  $X$  los datos observados y  $Z$  datos perdidos.

La distribución conjunta de  $W$  es

$$f(W | \theta) = f(X, Z | \theta) = k(Z | \theta, X)g(X | \theta).$$

# Algoritmo EM

¿Cómo calculamos  $L^c(\theta | W) = L^c(\theta | X, Z)$  si no conocemos  $Z$ ?

**Respuesta:** No conocemos  $Z$  de  $L^c(\theta | X, Z)$ , así que la supondremos como variable aleatoria y calculamos una media.

Considerando  $g(X | \theta) = \int_Z f(X, Z | \theta) dZ$ , donde  $(X, Z) \sim f(X, Z | \theta)$ . Entonces la distribución condicional de los datos perdidos  $Z$ , dado los datos observados  $X$  es

$$k(Z | \theta, X) = \frac{f(X, Z | \theta)}{g(X | \theta)}$$

# Algoritmo EM

Además existe una relación entre la verosimilitud para los datos completos  $L^c(\theta | X, Z)$  y la verosimilitud para los datos observados  $L(\theta | X)$  dada por

$$L^c(\theta | X, Z) = k(Z | \theta, X)L(\theta | X)$$

Luego tomando logaritmo

$$\log L^c(\theta | X, Z) = \log k(Z | \theta, X) + \log L(\theta | X)$$

Es decir :

$$\log g(X | \theta) = \log L(\theta | X) = \log L^c(\theta | X, Z) - \log k(Z | \theta, X)$$

# Algoritmo EM

Para un valor de  $\theta_0$ , calculando la esperanza con respecto a  $k(Z \mid \theta, X)$  y utilizando la desigualdad de Jensen, se tiene

- $\log L(\theta \mid X)$ : datos observados
- $E_{\theta_0}[\log L^c(\theta \mid X, Z)]$ : datos completos
- $E_{\theta_0}[\log k(Z \mid \theta, X)]$ : datos perdidos, que cumplen :

$$\log L(\theta \mid X) = E_{\theta_0}[\log L^c(\theta \mid X, Z)] - E_{\theta_0}[\log k(Z \mid \theta, X)]$$

Al maximizar  $\log L(\theta \mid X)$  se debe ignorar el término asociado solo a los datos perdidos.

# Iteraciones

- 1 El valor esperado del log-verosimilitud se denota por

$$Q(\theta \mid \theta_0, X) = E_{\theta_0}[\log L^c(\theta \mid X, Z)].$$

- 2 El algoritmo EM comienza maximizando en cada iteración  $Q(\theta \mid \theta_0, X)$ .
- 3 Si  $\hat{\theta}_{(1)} = \operatorname{argm\acute{a}x} Q(\theta \mid \theta_0, X)$  entonces  $\hat{\theta}_{(0)} \rightarrow \hat{\theta}_{(1)}$ .
- 4 Se tienen secuencias de estimadores  $\{\hat{\theta}_{(j)}\}$ , donde  $\hat{\theta}_{(j)} = \operatorname{argm\acute{a}x} Q(\theta \mid \hat{\theta}_{(j-1)}, X)$
- 5 Este esquema iterativo, en cada paso contiene un cculo de esperanza y maximizacin.



# El Algoritmo

# El Algoritmo

Se comienza con un valor inicial  $\hat{\theta}_{(0)}$  fijado por el investigador.  
Repita

# El Algoritmo

Se comienza con un valor inicial  $\hat{\theta}_{(0)}$  fijado por el investigador.  
Repita

1 Calcule(Paso E)

$$Q(\theta \mid \hat{\theta}_{(m)}, X) = E_{\hat{\theta}_{(m)}}[\log L^c(\theta \mid X, Z)],$$

donde la esperanza es con respecto a  $k(Z \mid \hat{\theta}_{(m)}, X)$   
y establecer  $m = 0$ .

# El Algoritmo

Se comienza con un valor inicial  $\hat{\theta}_{(0)}$  fijado por el investigador.  
Repita

- 1 Calcule(Paso E)

$$Q(\theta \mid \hat{\theta}_{(m)}, X) = E_{\hat{\theta}_{(m)}}[\log L^c(\theta \mid X, Z)],$$

donde la esperanza es con respecto a  $k(Z \mid \hat{\theta}_{(m)}, X)$   
y establecer  $m = 0$ .


- 2 Maximizar  $Q(\theta \mid \hat{\theta}_{(m)}, X)$  en  $\theta$  y tomar (paso M)


$$\hat{\theta}_{(m+1)} = \operatorname{argm\acute{a}x}\{Q(\theta \mid \hat{\theta}_{(m)}, X) : \theta\}$$

y establecer  $m = m + 1$ .

Los parámetros que se encuentran en el paso M se usan para comenzar el paso E, y así el proceso se repite. Es decir se fija el punto  $\hat{\theta}_{(m+1)} = \hat{\theta}_{(m)}$

# Bibliografía

 Maya R. Gupta, Yihua Chen  
Theory and Use of the EM Algorithm.  
Now Publishers Inc, 2011

 Bickel and Docksum  
Mathematical Statistics  
A Chapman & Hall Book