Algoritmo de maximización de la esperanza (*Expectation Maximization algorithm*)

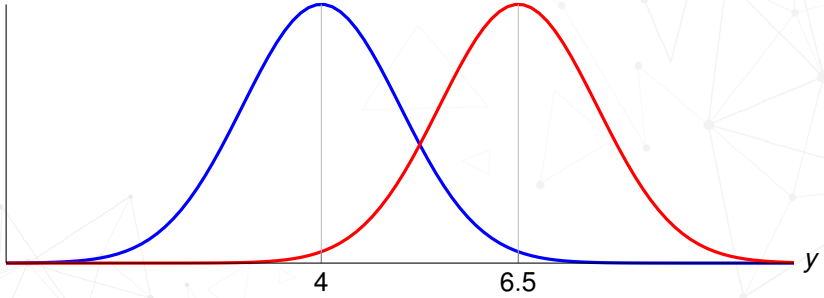
Cristian Guarnizo-Lemus
cristianguarnizo@itm.edu.co

Instituto Tecnológico Metropolitano

Contenido

Contenido

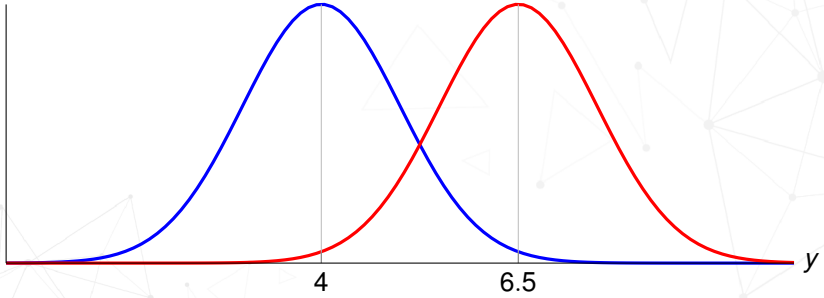
Definición del problema - Ejemplo



z_n

y_n

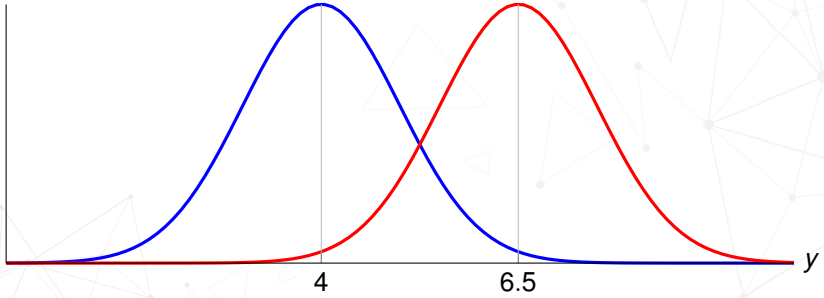
Definición del problema - Ejemplo



z_n 0

y_n

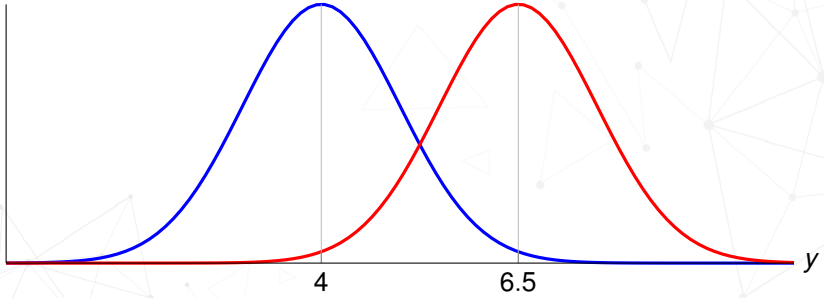
Definición del problema - Ejemplo



z_n 0

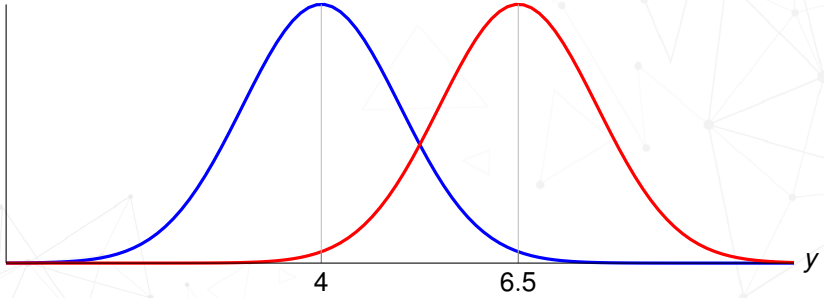
y_n 4.3

Definición del problema - Ejemplo



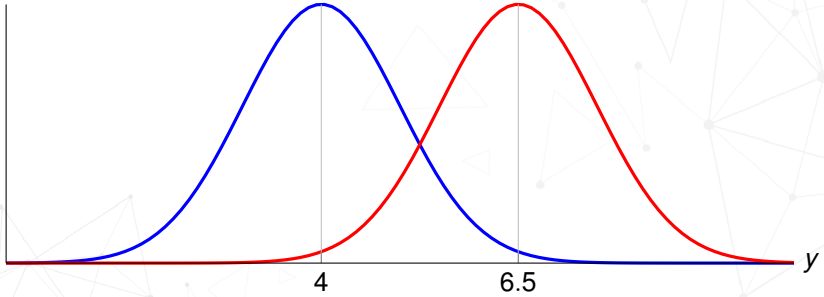
z_n	0	1
y_n	4.3	

Definición del problema - Ejemplo



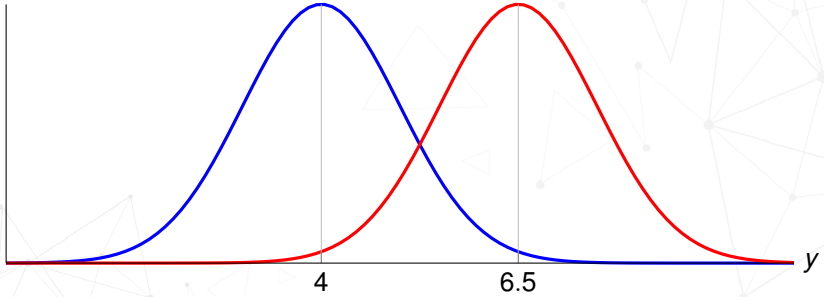
z_n	0	1
y_n	4.3	6.0

Definición del problema - Ejemplo



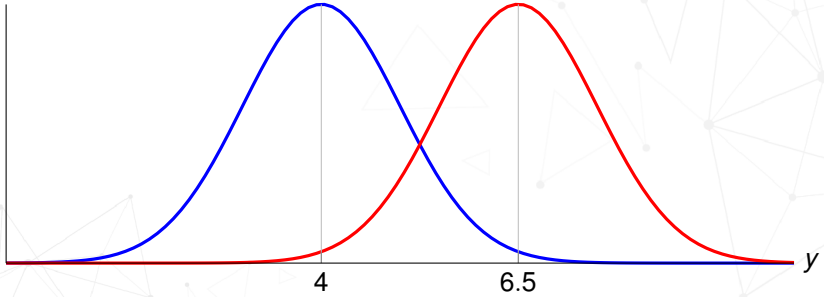
z_n	0	1	1
y_n	4.3	6.0	

Definición del problema - Ejemplo



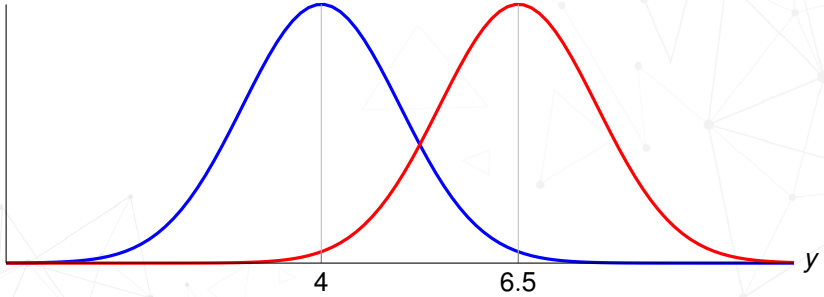
z_n	0	1	1
y_n	4.3	6.0	7.2

Definición del problema - Ejemplo



z_n	0	1	1	0
y_n	4.3	6.0	7.2	

Definición del problema - Ejemplo



z_n	0	1	1	0
y_n	4.3	6.0	7.2	3.8

Definición del problema - Ejemplo

Para z:

$$p(z) = \pi^z(1-\pi)^{1-z}, \quad p(z=1) = \pi, \quad p(z=0) = 1-\pi.$$

Para y:

$$p(y|z) = \left(\mathcal{N}\left(y|\mu_1, \sigma_1^2\right)\right)^z \left(\mathcal{N}\left(y|\mu_0, \sigma_0^2\right)\right)^{1-z}.$$

Entonces la probabilidad conjunta:

$$p(y,z) = \left(\pi\mathcal{N}\left(y|\mu_1, \sigma_1^2\right)\right)^z \left((1-\pi)\mathcal{N}\left(y|\mu_0, \sigma_0^2\right)\right)^{1-z}.$$

Definición del problema - Ejemplo

Asumamos que tenemos $(\mathbf{y}, \mathbf{z}) = \{y_n, z_n\}_{n=1}^{N}$.
Los parámetros $\theta = \{\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$, se pueden estimar como

$$\begin{aligned}
\theta_{\text{ML}} &= \arg\max_{\theta} \ln\left(p(\mathbf{y}, \mathbf{z}|\theta)\right), \\
&= \arg\max_{\theta} \ln\left(\prod_{n=1}^{N} p(y_n, z_n|\theta)\right), \\
&= \arg\max_{\theta} \sum_{n=1}^{N} \ln\left(p(y_n, z_n|\theta)\right).
\end{aligned}$$

Donde:

$$\begin{aligned}
\ln p(y_n, z_n|\theta) =& z_n \left(\ln(\pi) + \ln \mathcal{N}\left(y_n|\mu_1, \sigma_1^2\right)\right) \\
&+ (1 - z_n)\left(\ln(1-\pi) + \ln \mathcal{N}\left(y_n|\mu_0, \sigma_0^2\right)\right)
\end{aligned}$$

Definición del problema - Ejemplo

Asumamos que tenemos $\mathbf{y} = \{y_n\}_{n=1}^{N}$, y $\mathbf{z}$ es desconocida (latente).
Los parámetros $\theta = \{\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$, se pueden estimar como

$$\begin{aligned}\theta_{\text{ML}} &= \arg\max_{\theta} \sum_{n=1}^{N} \ln\left(p(y_n|\theta)\right), \\ &= \arg\max_{\theta} \sum_{n=1}^{N} \ln\left(\sum_{z} p(y_n, z_n|\theta)\right).\end{aligned}$$

Donde:

$$\ln\left(\sum_{z} p(y_n, z_n|\theta)\right) = \ln\left(\pi\mathcal{N}\left(y_n|\mu_1, \sigma_1^2\right) + (1-\pi)\mathcal{N}\left(y_n|\mu_0, \sigma_0^2\right)\right).$$

Definición del problema - General

Asumamos que el modelo consiste en observaciones **y** y una variable aleatoria **z** latente. Entonces

$$p(\mathbf{y}|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$
$$= \prod_{n=1}^{N} \sum_{\mathbf{z}} p(y_n, \mathbf{z}|\theta) = \prod_{n=1}^{N} \sum_{\mathbf{z}} p(y_n|\mathbf{z}, \theta) p(\mathbf{z}|\theta)$$

La estimación θ_{ML} esta dada por

$$\theta_{\text{ML}} = \arg\max_{\theta} \ln\left(p(\mathbf{y}|\theta)\right)$$
$$= \arg\max_{\theta} \sum_{n=1}^{N} \ln\left(\sum_{\mathbf{z}} p(y_n|\mathbf{z}, \theta) p(\mathbf{z}|\theta)\right)$$

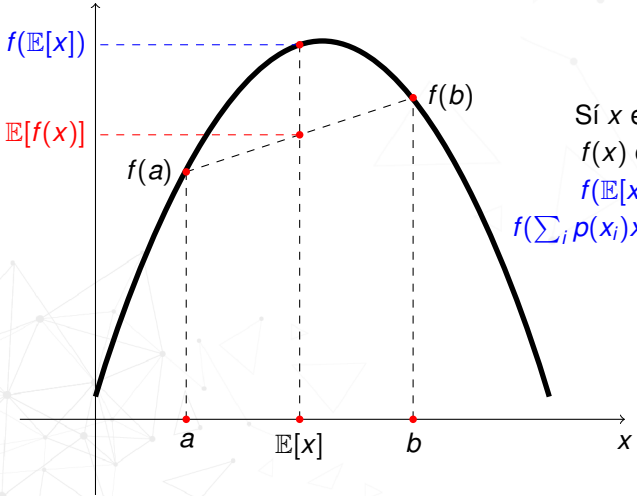
Definición del problema

$$\theta_{\text{ML}} = \arg\max_{\theta} \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} p(y_n, |\mathbf{z}, \theta) p(\mathbf{z}|\theta) \right)$$

- La sumatoria acopla los parámetros θ.
- Los gradientes no tienen forma cerrada.
- $\mathbf{z}$ y θ están acoplados.

Contenido

Desigualdad de Jensen



Sí x es una v.a. y $f(x)$ es concava:

$$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$$

$$f(\textstyle\sum_i p(x_i)x_i) \geq \sum_i p(x_i)f(x_i)$$

Desarrollo EM - Desigualdad de Jensen

$$\log p(\mathbf{y}|\theta) = \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} p(y_n, \mathbf{z}|\theta) \right)$$

Desarrollo EM - Desigualdad de Jensen

$$\log p(\mathbf{y}|\theta) = \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} p(y_n, \mathbf{z}|\theta) \right)$$
$$= \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(y_n, \mathbf{z}|\theta)}{q(\mathbf{z})} \right)$$

Desarrollo EM - Desigualdad de Jensen

$$
\begin{aligned}
\log p(\mathbf{y}|\theta) &= \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} p(y_n, \mathbf{z}|\theta) \right) \\
&= \sum_{n=1}^{N} \ln \left(\sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(y_n, \mathbf{z}|\theta)}{q(\mathbf{z})} \right) \\
&\geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(y_n, \mathbf{z}|\theta)}{q(\mathbf{z})} \right)
\end{aligned}
$$

Distribución

Asumiendo un valor para los parámetros como $\theta^{(t)}$, que forma debe tener $q(\mathbf{z})$ para maximizar la expresión

$$\ln p(\mathbf{y}|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}) p(y_n|\theta^{(t)})}{q(\mathbf{z})} \right)$$

Distribución

Asumiendo un valor para los parámetros como $\theta^{(t)}$, que forma debe tener $q(\mathbf{z})$ para maximizar la expresión

$$\ln p(\mathbf{y}|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}) p(y_n|\theta^{(t)})}{q(\mathbf{z})} \right)$$

$$\geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \left[\ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}}{q(\mathbf{z})} \right) + \ln p(y_n|\theta^{(t)}) \right]$$

Distribución

Asumiendo un valor para los parámetros como $\theta^{(t)}$, que forma debe tener $q(\mathbf{z})$ para maximizar la expresión

$$\ln p(\mathbf{y}|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}) p(y_n|\theta^{(t)})}{q(\mathbf{z})} \right)$$

$$\geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \left[\ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}}{q(\mathbf{z})} \right) + \ln p(y_n|\theta^{(t)}) \right]$$

reorganizando,

$$\sum_{n=1}^{N} \ln p(y_n|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)})}{q(\mathbf{z})} \right) + \sum_{n=1}^{N} \ln p(y_n|\theta^{(t)}),$$

Distribución

Asumiendo un valor para los parámetros como $\theta^{(t)}$, que forma debe tener $q(\mathbf{z})$ para maximizar la expresión

$$\ln p(\mathbf{y}|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}) p(y_n|\theta^{(t)})}{q(\mathbf{z})} \right)$$

$$\geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \left[\ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)}}{q(\mathbf{z})} \right) + \ln p(y_n|\theta^{(t)}) \right]$$

reorganizando,

$$\sum_{n=1}^{N} \ln p(y_n|\theta^{(t)}) \geq \sum_{n=1}^{N} \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|y_n, \theta^{(t)})}{q(\mathbf{z})} \right) + \sum_{n=1}^{N} \ln p(y_n|\theta^{(t)}),$$

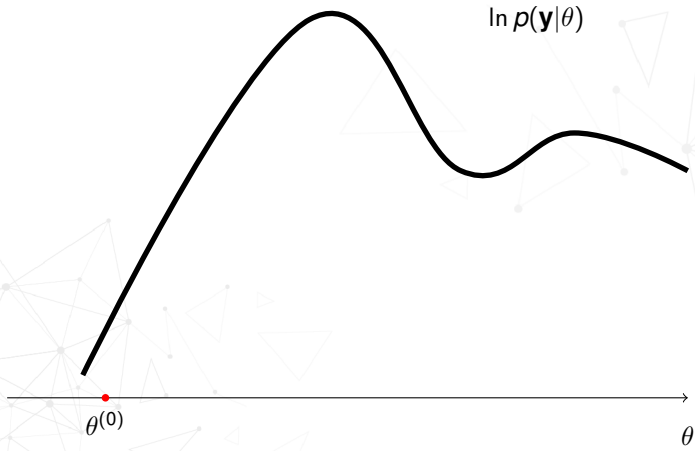
entonces, $q(\mathbf{z}) = p(\mathbf{z}|y_n, \theta^{(t)})$.

Estimación de los parámetros

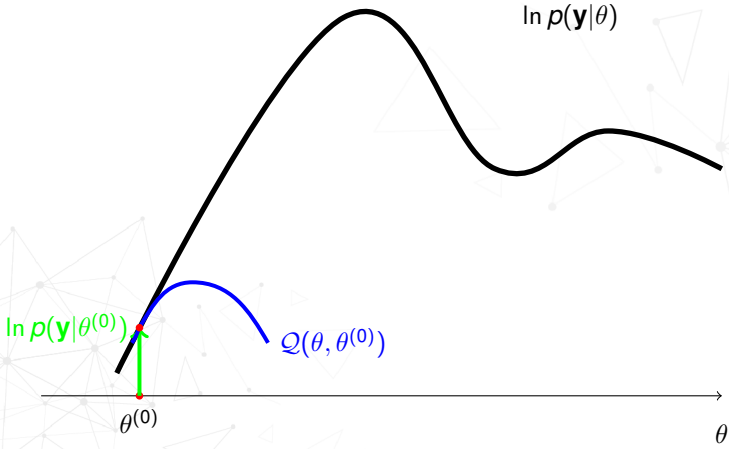
Sí $q(\mathbf{z}) = p(\mathbf{z}|y_n, \theta^{(t)})$,

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|y_n, \theta^{(t)}) \ln\left(\frac{p(y_n, \mathbf{z}|\theta)}{p(\mathbf{z}|y_n, \theta^{(t)})}\right) \\ &= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{z})}\left[\ln\left(p(y_n, \mathbf{z}|\theta)\right)\right] \\ &= \arg\max_{\theta} \mathcal{Q}\left(\theta, \theta^{(t)}\right)\end{aligned}$$

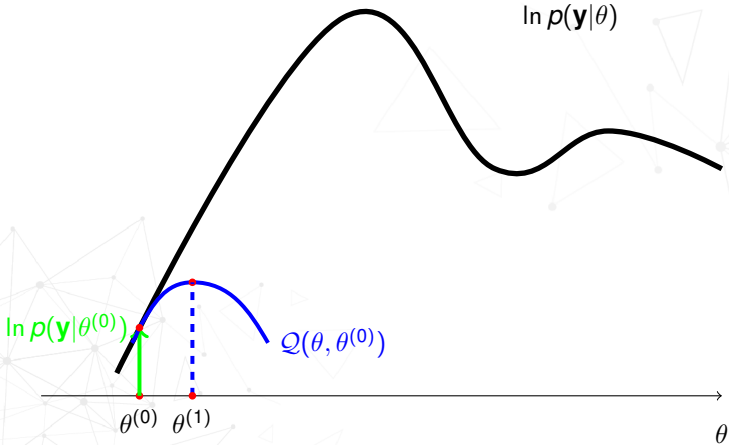
Algoritmo EM - Visual



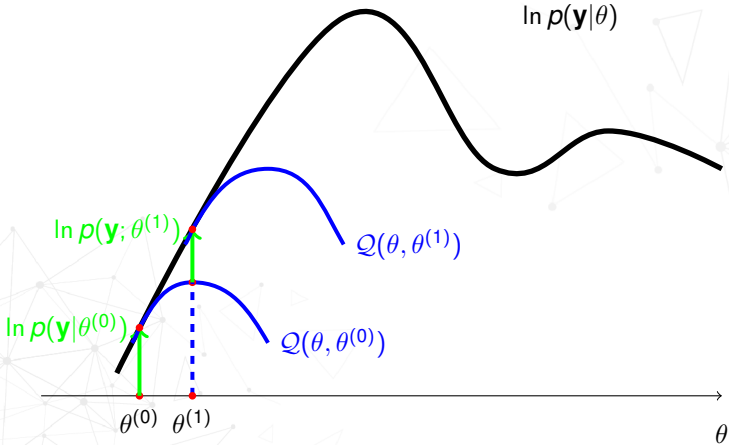
Algoritmo EM - Visual



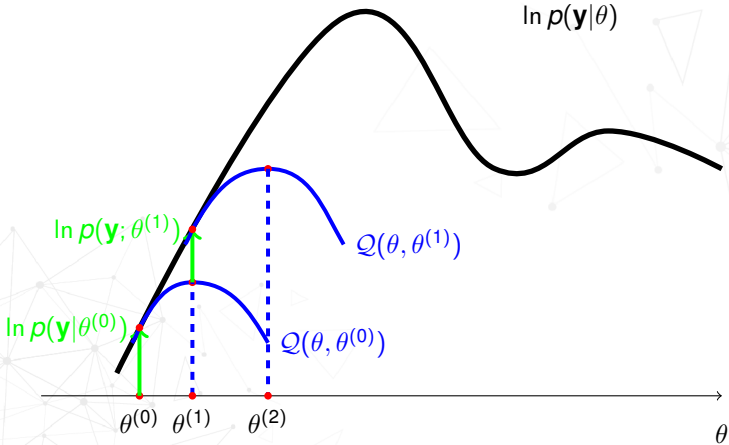
Algoritmo EM - Visual



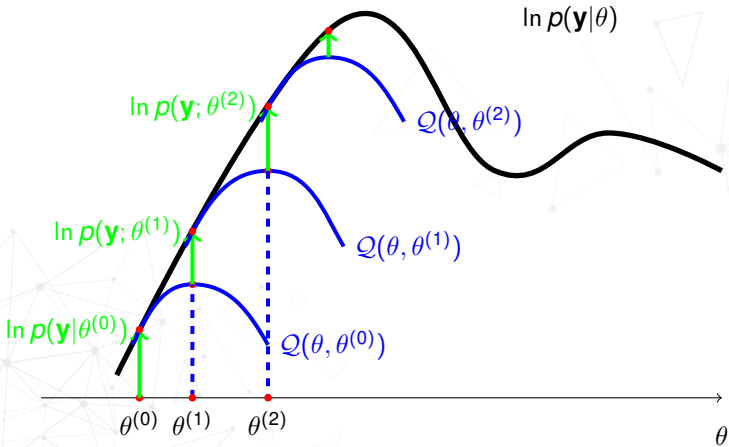
Algoritmo EM - Visual



Algoritmo EM - Visual



Algoritmo EM - Visual



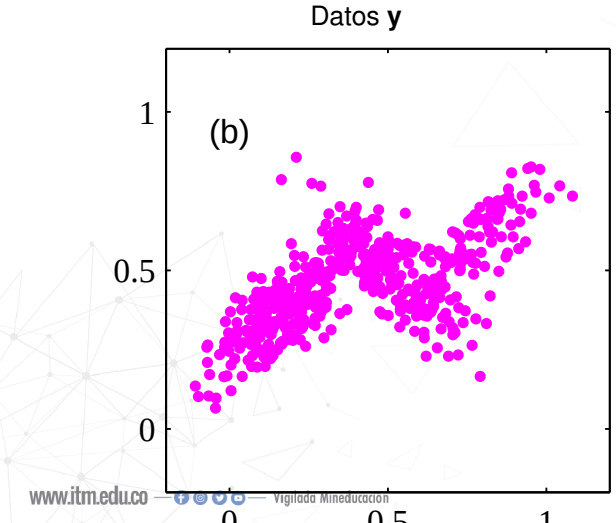
Algoritmo EM - Pseudo-código

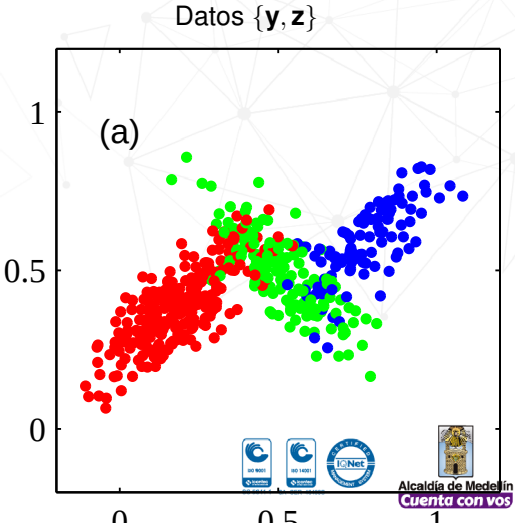
Dada la distribución conjunta $p(\mathbf{y}, \mathbf{z}|\theta)$, el objetivo es maximizar $p(\mathbf{y}|\theta)$ con respecto a θ.

1. Seleccionar los parámetros iniciales $\theta^{(t)} \leftarrow \theta^{(0)}$.
2. Evaluar el paso E, $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{y}, \theta^{(t)})$.
3. Evaluar paso M,
$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{Q}\left(\theta, \theta^{(t)}\right)$$
4. Verificar convergencia. Sí no se satisface, entonces
$$\theta^{(t)} \leftarrow \theta^{(t+1)},$$
y regresar al paso 2.

Contenido

Mezclas de Gaussianas





Mezclas de Gaussianas - Modelo

Observaciones $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, $\mathbf{Y} \in \mathbb{R}^{N \times D}$.
Para la variable latente $\mathbf{z} = \{z_1, \ldots, z_K\}$:

$$p(z_k = 1) = \pi_k$$
$$p(\mathbf{y}_n | z_k = 1) = \mathcal{N}(\mathbf{y}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Modelo probalístico:

$$p(\mathbf{y}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{Y}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Objetivo: Determinar $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

Mezclas de Gaussianas - Paso E

Definimos $\gamma(z_{nk}) = p(z_k = 1|\mathbf{y}_n, \theta^{(t)})$,

$$\begin{aligned}\gamma(z_{nk}) &= \frac{p(\mathbf{y}_n|z_k = 1)p(z_k = 1)}{p(\mathbf{y}_n)} \\ &= \frac{\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})\pi_k^{(t)}}{\sum_{k=1}^{K} \pi_j^{(t)}\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}\end{aligned}$$

Mezclas de Gaussianas - Paso M

La función objetivo:

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|y_n, \theta^{(t)}) \ln \left(p(\mathbf{y}_n, |\mathbf{z}, \theta) p(\mathbf{z}|\theta) \right)$$

Mezclas de Gaussianas - Paso M

La función objetivo:

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|y_n, \theta^{(t)}) \ln \left(p(\mathbf{y}_n, |\mathbf{z}, \theta) p(\mathbf{z}|\theta)\right)$$

$$= \arg\max_{\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln \left(\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

Mezclas de Gaussianas - Paso M

La función objetivo:

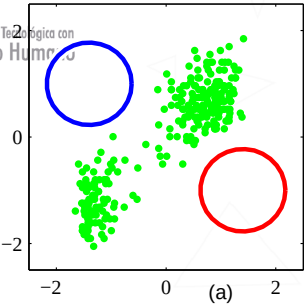
$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|y_n, \theta^{(t)}) \ln \left(p(\mathbf{y}_n, |\mathbf{z}, \theta) p(\mathbf{z}|\theta)\right)$$

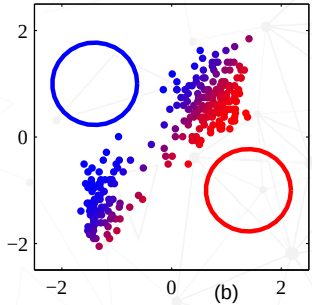
$$= \arg\max_{\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln \left(\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k\right)$$

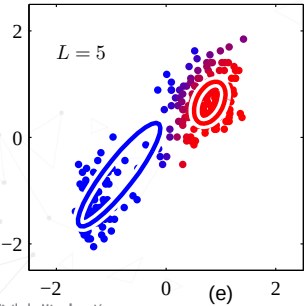
$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{y}_n, \quad \pi_k^{(t+1)} = \frac{N_k}{N}$$

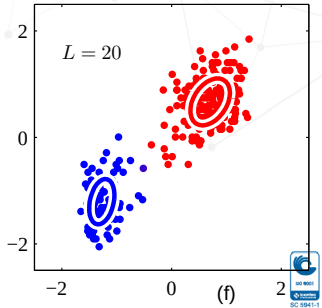
$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)^{\top},$$

donde $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.


2
0
−2
−2
0
2
(a)


2
0
−2
−2
0
2
(b)


$L = 5$
2
0
−2
−2
0
2
(e)


$L = 20$
2
0
−2
−2
0
2
(f)

Contenido

Datos faltantes

y_1	y_2	y_3
1.5	3.2	2.
2.2	2.7	
1.7		
1.5	3.2	2.2
2.2	2.9	
	3.0	

Datos faltantes

y_1	y_2	y_3
1.5	3.2	2.
2.2	2.7	1.8
1.7	2.8	2.1
1.5	3.2	2.2
2.2	2.9	2.3
1.9	3.0	2.1

$$\hat{\mathbf{z}} = \mathbb{E}_{p(\mathbf{z}|\mathbf{Y},\theta^{(t)})}[p(\mathbf{Y},\mathbf{z}|\theta)]$$

Datos faltantes

y_1	y_2	y_3
1.5	3.2	2.
2.2	2.7	1.8
1.7	2.8	2.1
1.5	3.2	2.2
2.2	2.9	2.3
1.9	3.0	2.1

$$\hat{\mathbf{z}} = \mathbb{E}_{p(\mathbf{z}|\mathbf{Y},\theta^{(t)})}[p(\mathbf{Y},\mathbf{z}|\theta)]$$
$$\theta^{(t+1)} = \arg\max_{\theta} p(\mathbf{Y},\hat{\mathbf{z}}|\theta)$$

Contenido

Resumen

- El algoritmo EM permite realizar estimación de máxima verosimilitud sobre modelos con variables latentes.
- El paso E construye la función $\mathcal{Q}(\theta, \theta^{(t)})$ (conjunta) a partir del valor esperado (posterior) de las variables latentes.
- La función $\mathcal{Q}(\theta, \theta^{(t)})$ que es más fácil de maximizar que $\ln p(\mathbf{y}|\theta)$.

Referencias I


Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.