

# **MÉTODOS BASADOS EN MEZCLAS**

ALAN REYES-FIGUEROA

INTRODUCCIÓN A LA CIENCIA DE DATOS

(AULA 24) 25.ABRIL.2022

# Verosimilitud

Consideremos una observación  $\mathbf{x} \in \mathbb{R}^d$  de una variable aleatoria  $X$ .

Supongamos que  $X$  sigue una distribución  $f$ , donde  $f$  pertenece a una familia de distribuciones  $\{f_\theta\}_{\theta \in \Theta}$ , parametrizadas por  $\theta \in \Theta \subseteq \mathbb{R}^k$ . De todas estas distribuciones, queremos hallar aquella que maximiza la probabilidad de observar un cierto dato  $\mathbf{x}$ .

## Definición

La función de **verosimilitud**  $\mathcal{L}$  mide la bondad de ajuste de un modelo o distribución  $f_\theta$  con respecto de un conjunto de observaciones. Se define por

- Para distribuciones discretas,  $\mathcal{L}(\theta \mid \mathbf{x}) = \mathbb{P}_\theta(X = \mathbf{x})$ .
- Para distribuciones continuas,  $\mathcal{L}(\theta \mid \mathbf{x}) = f_\theta(\mathbf{x})$ .

# Verosimilitud

Consideremos una muestra  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$  de datos provenientes de una distribución  $f_\theta$ . Esto es, las  $\mathbf{x}_i$  son v.a. i.i.d. con  $\mathbf{x}_i \sim \mathbf{x}_1$  y  $\mathbf{x}_1 \sim f_\theta$ .

Com las  $\mathbf{x}_i$  son i.i.d., en este caso, la función de verosimilitud se calcula como

- Para distribuciones discretas

$$\mathcal{L}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \mathbb{P}_\theta(X = \mathbf{x}_i).$$

- Para distribuciones continuas

$$\mathcal{L}(\theta \mid \mathbf{x}) = f_{(\theta, \dots, \theta)}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n f_\theta(\mathbf{x}_i).$$

# Verosimilitud

Recordemos que uno de los métodos más útiles para estimar parámetros de una distribución se debe a Fisher, el **método de máxima verosimilitud**. Este consiste en determinar el **estimador de máxima verosimilitud** como aquel que maximiza la función  $\mathcal{L}$ , esto es

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} h(\mathbf{x}_1, \dots, \mathbf{x}_n) \mathcal{L}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n), \quad (1)$$

En general, conviene trabajar con algún múltiplo de la función de verosimilitud

$$\mathcal{L}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \prod_{i=1}^n \mathbb{P}_{\theta}(X = \mathbf{x}_i),$$

donde  $h$  es una función conveniente que sólo depende de la muestra observada.

En la práctica, usualmente se trabaja con la función de **log-verosimilitud**

$$\ell(\theta) = \log \mathcal{L}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n).$$

En este caso, el problema (1) se escribe como

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta). \quad (2)$$

Otras funciones útiles que sirven para encontrar el estimador máximo verosímil son la **función de Score**:

$$S(\theta) = \frac{\partial \ell}{\partial \theta}(\theta) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n).$$

y la **función de Información**:

$$I(\theta) = -\frac{\partial^2 \ell}{\partial \theta^2}(\theta).$$

(que son los análogos del criterio de la primera y segunda derivadas para hallar óptimos locales):

- el estimador máximo verosímil  $\hat{\theta}$  debe satisfacer  $S(\hat{\theta}) = \frac{\partial \ell}{\partial \theta}(\hat{\theta}) = 0$ .
- Si se cumple lo anterior, entonces  $\hat{\theta}$  es un
  - máximo local, si  $I(\hat{\theta}) \succ 0$ .
  - mínimo local, si  $I(\hat{\theta}) \prec 0$ .
  - punto silla, en caso contrario.

# Verosimilitud

Ejemplo: (Estimadores para una distribución normal). Sea  $x_1, \dots, x_n \in \mathbb{R}$  una muestra aleatoria proveniente de una distribución normal  $\mathcal{N}(\mu, \sigma^2)$ . Denotamos  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$

Queremos estimar  $\mu$  (asumimos  $\sigma$  conocida). En este caso, la función de verosimilitud para  $\mu$  es

$$\mathcal{L}(\mu \mid \mathbf{x}) = \prod_{i=1}^n f_{\theta}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^n \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Tomamos  $h(\mathbf{x}) = (\sqrt{2\pi}\sigma)^n$  y nos queda

$$\mathcal{L}(\mu \mid \mathbf{x}) = \prod_{i=1}^n \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

# Verosimilitud

La función de log-verosimilitud es

$$\ell(\mu) = \log \mathcal{L}(\mu \mid \mathbf{x}) = - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}.$$

Luego,

$$S(\mu) = \frac{\partial \ell}{\partial \mu} = \sum_{i=1}^n \frac{2(x_i - \mu)}{2\sigma^2} = \frac{1}{2\sigma^2} \left( \sum_{i=1}^n x_i - n\mu \right).$$

De ahí que  $S(\mu) = 0 \Rightarrow \sum_{i=1}^n x_i - n\mu = 0$ , y

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$



# Verosimilitud

Podemos verificar que, en efecto,  $\hat{\mu}$  es un máximo local. Tomamos la función de información

$$I(\mu) = -\frac{\partial^2 \ell}{\partial \mu^2}(\hat{\mu}) = \frac{n}{2\sigma^2} > 0.$$

Esto muestra que  $\hat{\mu}$  es un máximo local, y portanto es el estimador máximo verosímil para  $\mu$ .

(la media muestra es el estimador máximo verosímil de la media  $\mu$ , para una normal).

**Ejercicio.** Asumiendo  $\hat{\mu}$  como parámetro de la media para la normal, mostrar que el estimador máximo verosímil para la varianza  $\sigma^2$  es la varianza muestral

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

# Métodos basados en mezclas

Punto de partida: K-medias.

Haremos un cambio de notación,  $k$  en lugar de  $g$  y  $\mu_k$  en lugar de  $c_g$ .

Algoritmo: (K-medias)

- Elegir al azar  $k$  representantes
- Repetir hasta convergencia:
  - Asignar cada dato  $\mathbf{x}_i$  al representante  $\mu_k$  más cercano según la métrica  $d(\cdot, \cdot)$ .
  - Tomar como nuevos representantes, los centroides  $\mu_k$  de los datos asociados a un mismo representante.

**Observación:** minimizar  $\sum_k \|\mathbf{x}_i - \mu_k\|^2 \iff$  maximizar  $\sum_k \exp(-\frac{1}{2}\|\mathbf{x}_i - \mu_k\|^2)$ .  
( $X \sim \mathcal{N}_d(\mu_k, I)$ ).

# Métodos basados en mezclas

Pensemos ahora que los elementos de un grupo provienen de  $\mathbb{P}_k \sim \mathcal{N}(\mu_k, I)$ , para  $k = 1, 2, \dots, K$ .

Asignamos cada dato  $\mathbf{x}_i$  al índice  $k$  que maximiza  $\mathbb{P}_k(X = \mathbf{x}_i)$  (de alguna manera, estamos maximizando una especie de verosimilitud).

## Ideas nuevas:

- Queremos generalizar a  $\mathbb{P}_k(X = \mathbf{x}_i)$ .
- Primer paso:  $\mathbb{P}_k \sim \mathcal{N}(\mu_k, I)$ .
- Definimos v.a.  $X$  y  $Y$ , donde  $Y$  denota el grupo. Trabajamos con  $\mathbb{P}_k(Y \mid X = \mathbf{x}_i)$ .

# Métodos basados en mezclas

Primer intento: definir la distribución de  $X$  directamente.

Modelo de mezclas:

Por ejemplo, la mezcla de dos gaussianas

$$\mathbb{P}(X = \mathbf{x}) = (1 - \alpha_1)\mathbb{P}_0(X = \mathbf{x}) + \alpha_1\mathbb{P}_1(X = \mathbf{x}),$$

con  $\mathbb{P}_0, \mathbb{P}_1$  distribuciones gaussianas.

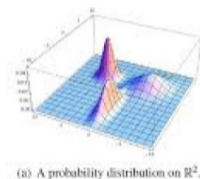
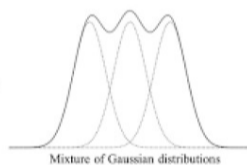
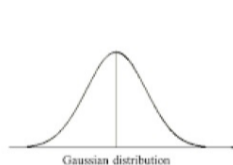
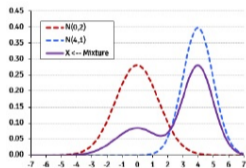
El agrupamiento consiste en estimar los parámetros  $\theta$  de la distribución de  $X$ . La log-verosimilitud es de la forma

$$\ell(\theta) = \sum_i \log \left( (1 - \alpha_1)\mathbb{P}_0(X = \mathbf{x}_i) + \alpha_1\mathbb{P}_1(X = \mathbf{x}_i) \right).$$

# Métodos basados en mezclas

Caso general: se supone  $X$  proviene de una mezcla de  $K$  gaussianas

$$\mathbb{P}(X = \mathbf{x}) = \sum_{k=1}^K \alpha_k \mathbb{P}_k(X = \mathbf{x}).$$



(a) A probability distribution on  $\mathbb{R}^2$ .

## Problema:

El modelo es difícil de estimar (la función de log-verosimilitud es difícil de optimizar).

# Métodos basados en mezclas

Segundo intento: definir una v.a.  $Y$  que indica la categoría (grupo) de  $X$ ,  
 $\Rightarrow Y \sim \text{Ber}(\alpha_1)$ .

- antes  $\mathbb{P}(X = \mathbf{x}) = P_0(X = \mathbf{x})(1 - \alpha_1) + \mathbb{P}_1(X = \mathbf{x})\alpha_1$ .

- ahora

$$\mathbb{P}(X = \mathbf{x}) = \mathbb{P}(X = \mathbf{x} \mid Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X = \mathbf{x} \mid Y = 1)\mathbb{P}(Y = 1).$$

Conociendo  $\{(X_i, Y_i)\}$ , la log-verosimilitud es

$$\begin{aligned} \ell(\theta) &= \sum_i \log \mathbb{P}(X_i = \mathbf{x}_i, Y_i = y_i) = \sum_i \log (\mathbb{P}(X_i = \mathbf{x}_i, Y_i = y_i) \mathbb{P}(Y_i = y_i)) \\ &= \sum_i \log \mathbb{P}(X_i = \mathbf{x}_i, Y_i = y_i) + \sum_i \log \mathbb{P}(Y_i = y_i) \end{aligned}$$

Como  $Y_i$  es binaria, entonces

# Métodos basados en mezclas

$$\sum_i \log \mathbb{P}(Y_i = y_i) = n_0 \log \mathbb{P}(Y = 0) + n_1 \log \mathbb{P}(Y = 1),$$

con  $n_k = \#\{i : y_i = k\}$ . Luego

$$\ell(\theta) = \sum_{i:y_i=0} \log \mathbb{P}(Y = 0) + \sum_{i:y_i=1} \log \mathbb{P}(Y = 1) + n_0 \log(1 - \alpha_1) + n_1 \log \alpha_1.$$

Con esto, podemos obtener problemas de optimización desacoplados: más simple de obtener estimadores de máxima verosimilitud.

Por ejemplo, si  $\mathbb{P}_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ , entonces

- $\hat{\mu}_k$  es el promedio muestral de  $M_k = \{\mathbf{x}_i : y_i = k\}$ ,
- $\hat{\sigma}_k$  es la desviación estándar muestral de  $M_k$ ,  $\hat{\sigma}_k = \frac{n_1}{n_0 + n_1}$ .

**Problema:** no conocemos  $\{Y_i\}$ .

# El algoritmo EM

EM = *Expectation Maximization*. Idea:

- Si conocemos  $\{Y_i\}$ , hay una solución cerrada, dada por

$$\hat{\mu}_0 = \frac{\sum_i (1 - y_i) \mathbf{x}_i}{n_0}, \quad \hat{\mu}_1 = \frac{\sum_i y_i \mathbf{x}_i}{n_1}. \quad (3)$$

similarmente para  $\hat{\sigma}_0$  y  $\hat{\sigma}_1$ ; y  $\hat{\alpha}_1 = \frac{n_0}{n_0 + n_1} = \frac{\sum_i y_i}{n}$ .

- Si conocemos los parámetros  $\theta$ , podemos calcular

$$\mathbb{E}(Y_i | \{\mathbf{x}_i\}, \theta) = \mathbb{E}_\theta(Y_i | \{\mathbf{x}_i\}) = \mathbb{P}_\theta(Y_i = 1 | \mathbf{x}_i) = \frac{\mathbb{P}_\theta(\mathbf{x}_i | Y_i = 1) \mathbb{P}_\theta(Y_i = 1)}{\mathbb{P}_\theta(\mathbf{x}_i)}.$$

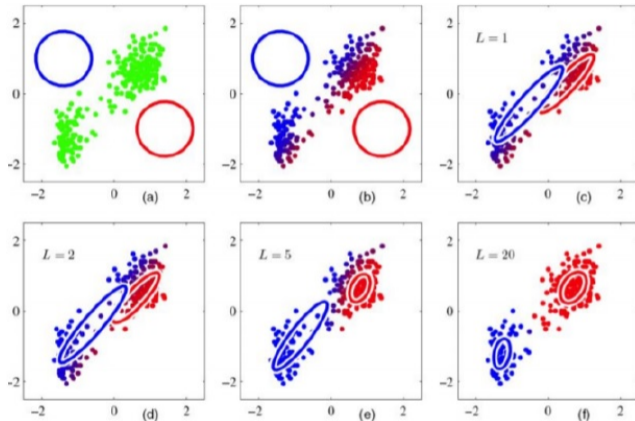
La idea es iterar lo anterior, usando en (3) en lugar de  $y_i$ ,  $\mathbb{E}_\theta(Y_i | \{\mathbf{x}_i\})$ ,

$$\hat{\mu}_0 = \frac{\sum_i \mathbb{E}_\theta[(1 - y_i) | \{\mathbf{x}_i\}]}{n_0}, \quad \hat{\mu}_1 = \frac{\sum_i \mathbb{E}_\theta[y_i | \{\mathbf{x}_i\}]}{n_1}.$$

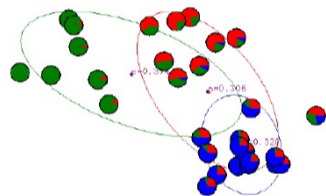
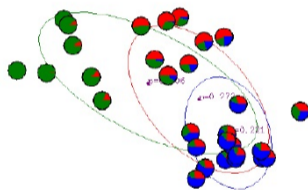
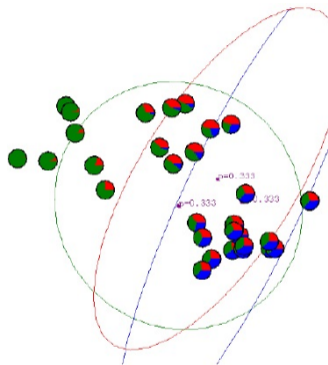


# El algoritmo EM

En general, para una mezcla de  $K$  distribuciones, para cada  $\mathbf{x}_i$ , tenemos un vector  $\gamma_i \in \mathbb{R}^K$ , con  $\gamma_i(k) = \mathbb{P}(Y_i = k \mid \mathbf{x}_i, \theta)$ .



# El algoritmo EM



# El algoritmo EM

Algoritmo: (EM, forma general).

Sea  $T = (Z, Z^m)$ , donde  $Z^m$  se refiere a la parte faltante.

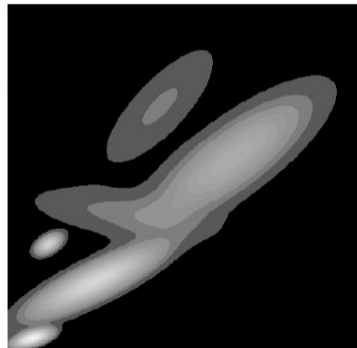
1. Definir  $\ell_o(\theta, T)$  como la log-verosimilitud basada en los datos completos.
2. Adivinar inicialmente  $\hat{\theta}_0$ .
3. Repetir, hasta convergencia:
  - Calcular  $Q(\theta | \hat{\theta}^t) = \mathbb{E}_{\hat{\theta}^t}(\ell_o(\theta, T) | Z)$ .
  - Definir  $\hat{\theta}^{t+1}$  como el máximo de  $Q(\theta | \hat{\theta}^t)$ .

En el caso de la mezcla de gaussianas,  $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1, \gamma)$ ,  $T = (Z, Z^m)$  es  $(X, Y)$ . De ahí

$$\ell_o(\theta, T) = \sum_{i:y_i=0} \log \mathbb{P}_{0,\theta}(X = \mathbf{x}_i) + \sum_{i:y_i=1} \log \mathbb{P}_{1,\theta}(X = \mathbf{x}_i) + \left(1 - \sum_i \frac{y_i}{n}\right) \log(1 - \alpha_1) + \sum_i \frac{y_i}{n} \log \alpha_1.$$

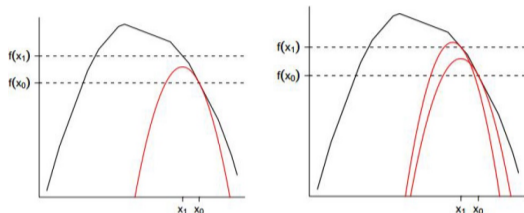
# El algoritmo EM

En el caso de una mezcla de gaussianas, si se compara EM con  $k$ -medias se observa que EM usa asignación *fuzzy*.



# El algoritmo EM

En el trasfondo, EM es un algoritmo de *maxmin* (o *minimax*).



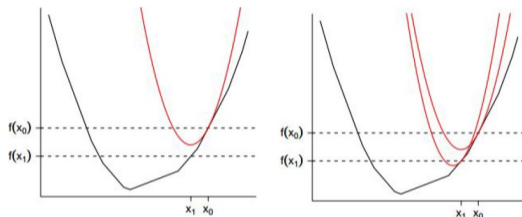
Algoritmo MM: Maximizar un minorizador. Para resolver  $\operatorname{argmax} f(\cdot)$ , construye secuencia de aproximaciones donde

$$\theta^{n+1} = \operatorname{argmax}_{\theta} g(\theta | \theta^n),$$

donde  $g$  es tal que:  $f(\theta^n) = g(\theta^n | \theta^n)$ ,  $f(\theta) \geq g(\theta | \theta^n)$ .

# El algoritmo EM

En el trasfondo, EM es un algoritmo de *maxmin* (o *minimax*).



Algoritmo MM: También se refiere a minimizar un mayorizador. Para resolver  $\operatorname{argmin} f(\cdot)$ , construye secuencia de aproximaciones donde

$$\theta^{n+1} = \operatorname{argmin}_{\theta} g(\theta \mid \theta^n),$$

donde  $g$  es tal que:  $f(\theta^n) = g(\theta^n \mid \theta^n)$ ,  $f(\theta) \leq g(\theta \mid \theta^n)$ .

# El algoritmo EM

En este contexto de minimizar un mayorizador (minimax), tenemos la siguiente

## Propiedad

$$f(\theta^{n+1}) \leq f(\theta^n), \forall n \in \mathbb{N}.$$

### Prueba:

Como  $\theta^{n+1} = \operatorname{argmin}_{\theta} g(\theta | \theta^n)$  y  $f(\theta) \leq g(\theta | \theta^n)$ , entonces

$$\begin{aligned} f(\theta^{n+1}) &= g(\theta^{n+1} | \theta^n) + f(\theta^{n+1}) - g(\theta^{n+1} | \theta^n) \\ &\leq g(\theta^n | \theta^n) \\ &\leq f(\theta^n). \end{aligned}$$

# El algoritmo EM

Ejemplo: Dada muestra  $\{y_i\}$  calcular la mediana, minimizando

$$f(\theta) = \sum_i |y_i - \theta|.$$

Se puede mostrar que la función  $h_i(\theta | \theta^n) = \frac{1}{2} \frac{(y_i - \theta)^2}{|y_i - \theta^n|} + \frac{1}{2} |y_i - \theta^n|$  mayoriza a  $|y_i - \theta|$  en  $\theta^n$ . Definimos

$$g(\theta | \theta^n) = \sum_i h_i(\theta | \theta^n), \quad (4)$$

que mayoriza a  $f(\theta)$ .

Hay una solución explícita para el mínimo en (4):

$$\theta^{n+1} = \frac{\sum_i w_i^n y_i}{\sum_i w_i^n}, \quad \text{com } w_i^n = |y_i - \theta^n|^{-1}.$$