Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

Modelos Mixtos

Definiciones

Definición: decimos que una distribución F es mixtura de K distribuciones $F_1 \ldots, F_K$ si

$$F(x) = \sum_{k=1}^{K} \pi_k F_k, \tag{1}$$

donde π_j es el peso de la mixtura y cumple que,

- $\pi_j > 0, \ 1 \leq j \leq K.$
- $\sum_{j=1}^{K} \pi_j = 1.$

Definiciones

Es inmediato ver que si las distribuciones $F_1, \ldots, F_k$ tienen asociada una función de densidad f_j o probabilidad puntual p_j, entonces

$$p(x) = \sum_{j=1}^{K} \pi_j p_j(x).$$

$$f(x) = \sum_{j=1}^{K} \pi_j f_j(x).$$

Inclusive pueden obtenerse rapidamente relaciones para la esperanza y varianza (ejercicio).

Definiciones

De (1) se desprende un modelo estocástico que genera los datos con los que estemos trabajando.

$$Z \sim Cat(\pi_1, \pi_2, \ldots, \pi_K). \tag{2}$$

$$X|Z = k \sim F_k. \tag{3}$$

Observar que Z es discreta en (2), es decir, $P(Z = k) = \pi_k$ para $1 \leq k \leq K$.

Ejemplo

Veamos un ejemplo simple, consideremos una mixtura de dos gaussianas con la misma varianza.

$$f(x; \mu_1, \mu_2, \sigma^2) = (1 - \alpha) f_1(x; \mu_1, \sigma^2) + \alpha f_2(x; \mu_2, \sigma^2),$$

donde

$$f_1(x; \mu_1, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}.$$

$$f_2(x; \mu_2, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}.$$

$$\alpha \in [0, 1]$$

Dibujito Mixtura Normales Univariadas

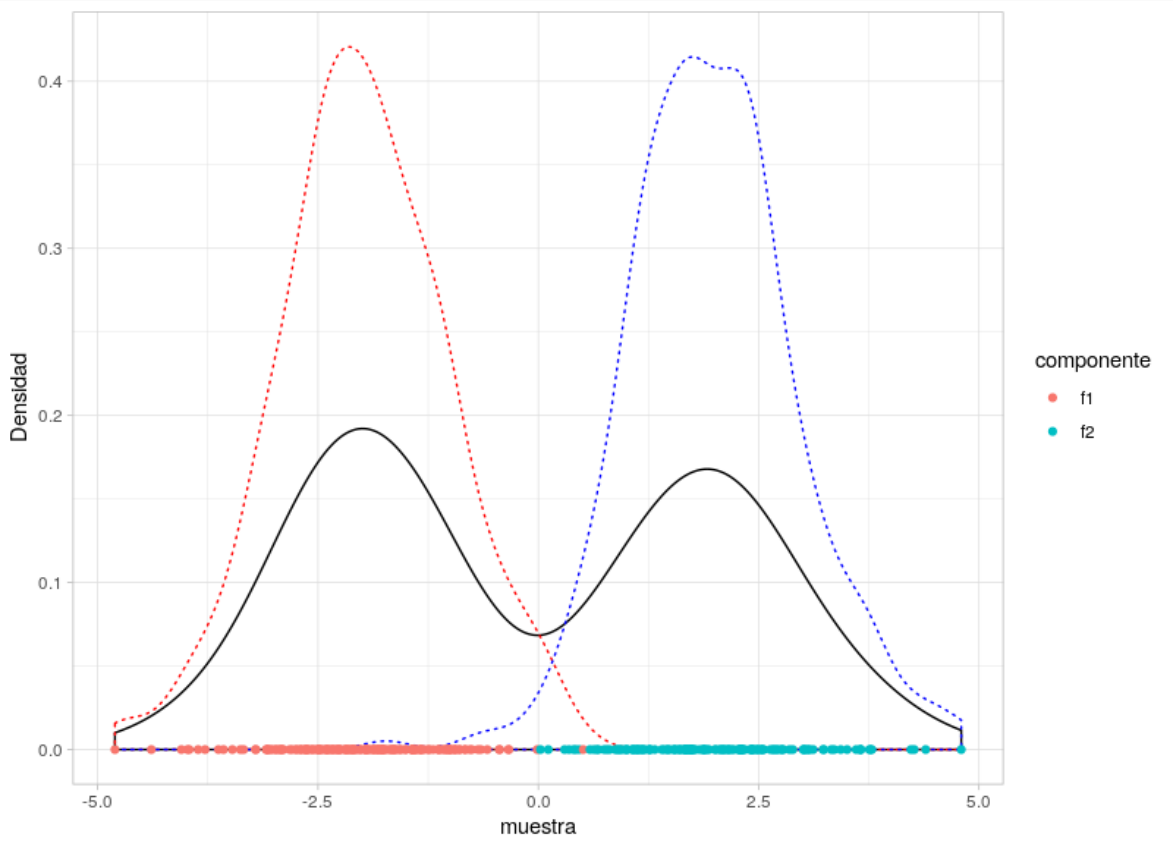

Figura 1: $alpha = 0{,}5$

Dibujito Mixtura Normales Univariadas

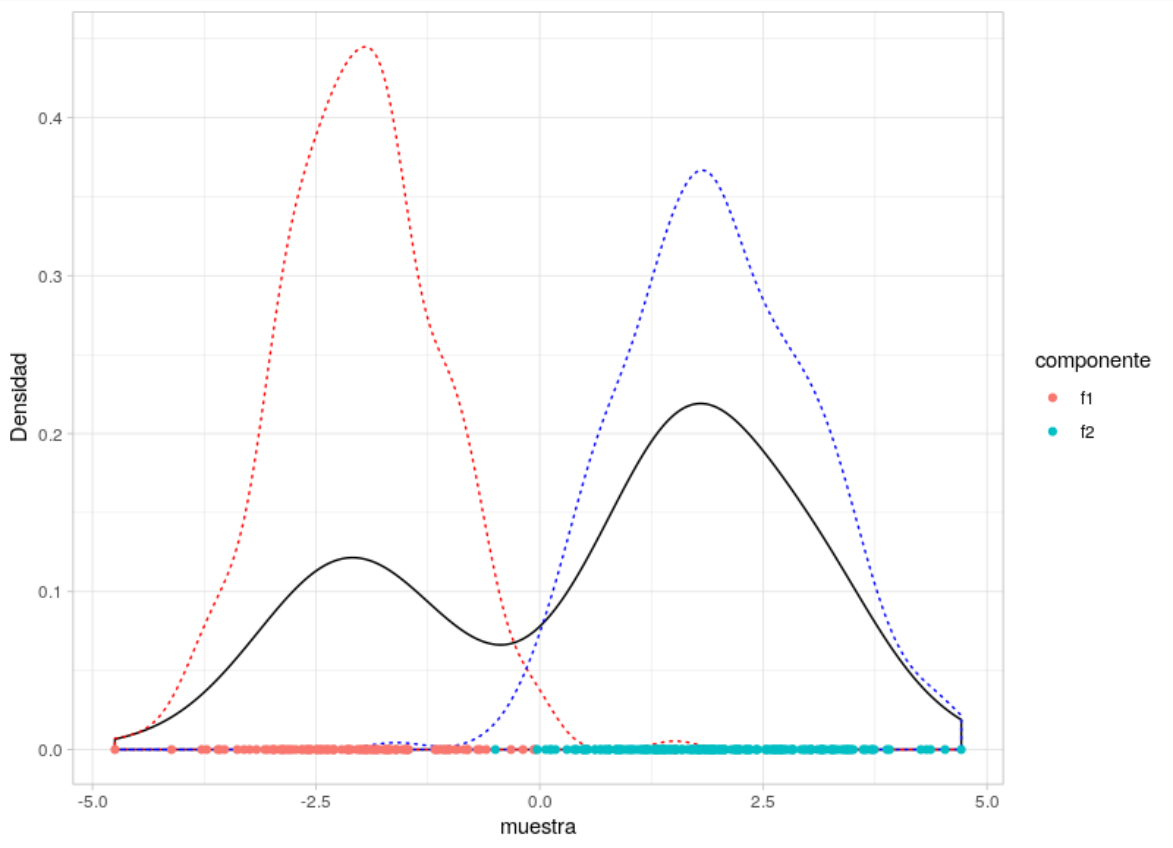

Figura 2: $alpha = 0{,}3$

Alpha es la proporcion

Dibujito Mixtura Normales Univariadas

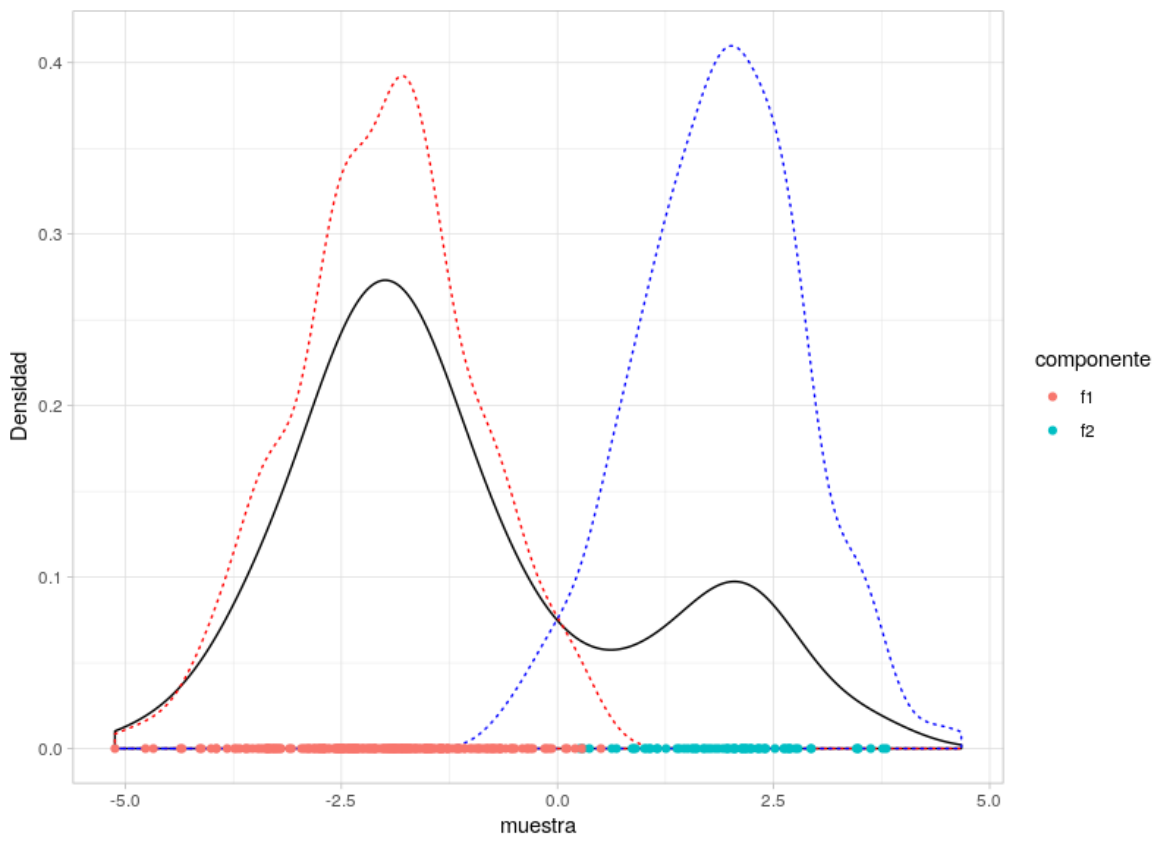

Figura 3: $alpha = 0{,}7$

Dibujito Mixtura Normales Univariadas

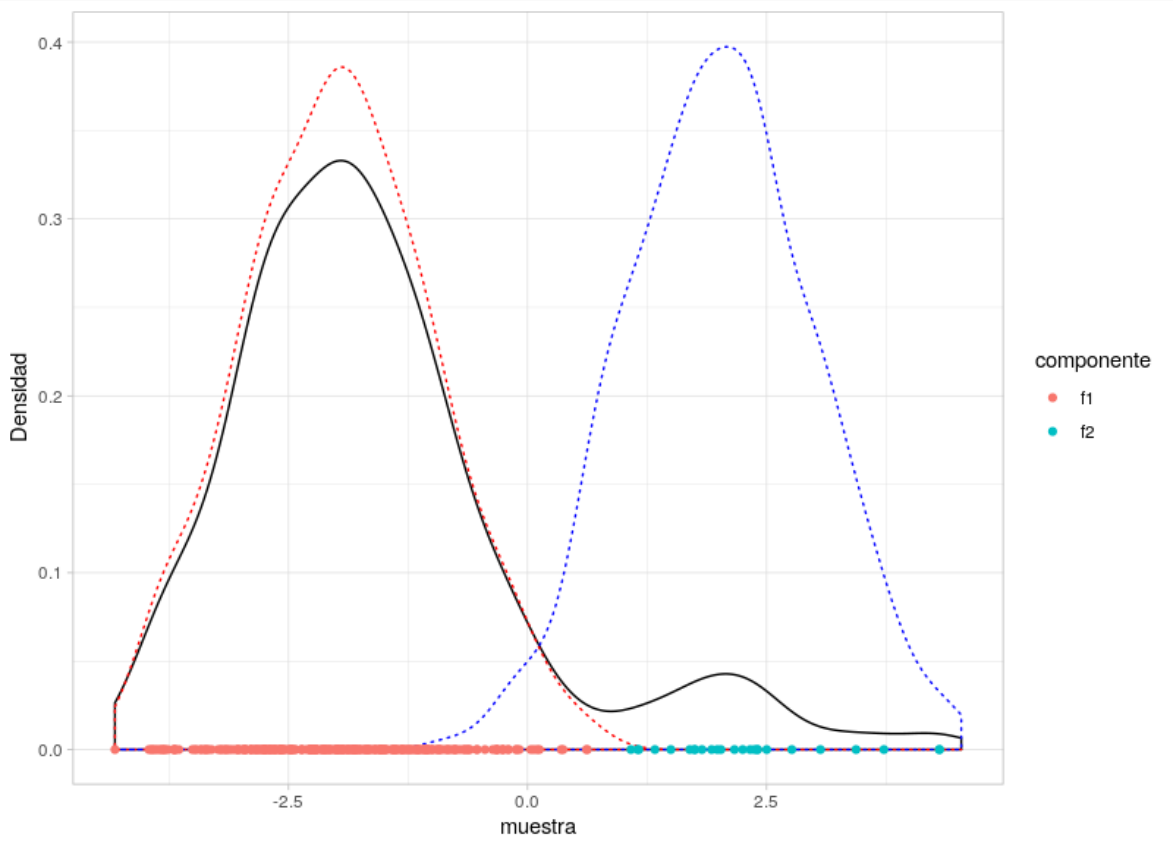

Figura 4: $alpha = 0{,}9$

Ejercicio

Generar una mixtura de 300 muestras provenientes de una mixtura de normales bivariadas con los siguientes parámetros:

- Medias:
  - $\mu_1 = c(0, 0)$
  - $\mu_2 = c(-2, -2)$
  - $\mu_3 = c(2, 2)$.
- Varianzas:
  - $\Sigma_1 = \Sigma_2 = \mathbb{I}$
  - $\Sigma_3 = 2\mathbb{I}$ donde ($\mathbb{I}$ es la matriz identidad).
- $\pi_1 = \pi_2 = \pi_3 = 1/3$.

Si lo resuelve rápido pruebe modificar las π's.

Clusters

Pregunta

¿Cómo podemos relacionar esta estructura estocástica a nuestro problema de agrupar datos en clusters?

Clusters

Pregunta

¿Cómo podemos relacionar esta estructura estocástica a nuestro problema de agrupar datos en clusters?

La conjetura se basa en que las observaciones que son similares provienen de la misma distribución (F_j).

La variable Z representa la etiqueta del cluster.

Clusters

Por el momento no hicimos ninguna suposición sobre las f_k's.

En la práctica reducimos el problema a una mixtura paramétrica donde las f_k pertenecen a la misma familia paramétrica, como por ejemplo, suponer que las f_k son normales con diferentes medias y varianzas.

Llamaremos θ_k al parámetro correspondiente a la componente k-esima, nuestro modelo de mixturas se convierte en

$$F(x) = \sum_{k=1}^{K} \pi_k F(x; \theta_k) \tag{4}$$

Clusters

Ahora que tenemos un modelo paramétrico, conocemos la forma de la distribución que genera nuestros datos. Por lo tanto, nuestro problema de asignar clusters comienza tratando de estimar los parámetros del modelo,

$$\theta = (\pi_1, \pi_2, \ldots, \pi_K, \theta_1, \theta_2, \ldots, \theta_K).$$

Podemos intentar un método conocido: el estimador de máxima verosimilitud.

Clusters

Ahora que tenemos un modelo paramétrico, conocemos la forma de la distribución que genera nuestros datos. Por lo tanto, nuestro problema de asignar clusters comienza tratando de estimar los parámetros del modelo,

$$\theta = (\pi_1, \pi_2, \ldots, \pi_K, \theta_1, \theta_2, \ldots, \theta_K).$$

Podemos intentar un método conocido: el estimador de máxima verosimilitud.

Disclaimer: MV podría no funcionar, el docente no se hace cargo por futuras pérdidas.

Estimador de Máxima Verosimilitud

Sea $X_1, \ldots, X_n$ una muestra aleatoria proveniente de nuestra mixtura F que tiene la forma de (4) que suponemos tiene probabilidad puntual o densidad f.

Recordemos,

$$Lik(\theta) = \prod_{i=1}^{N} f(x_i; \theta).$$

$$Loglik(\theta) = \sum_{i=1}^{N} log\left(f(x_i; \theta)\right) = \sum_{i=1}^{N} log\left(\sum_{k=1}^{K} \pi_k f_k(x_i, \theta_k)\right). \quad (5)$$

Estimador de Máxima Verosimilitud

Para obtener el estimador de máxima verosimilitud de θ necesitamos los valores que maximizan la *Loglik* (5).

Derivemos respecto de cada coordenada de parámetros,

$$\partial_{\theta_j} Loglik(\theta) = \partial_{\theta_j} \sum_{i=1}^{N} log \left(\sum_{k=1}^{K} \pi_k f_k(x_i, \theta_k) \right) \tag{6}$$

$$= \sum_{i=1}^{N} \partial_{\theta_j} log \left(\sum_{k=1}^{K} \pi_k f_k(x_i, \theta_k) \right) \tag{7}$$

Estimador de Máxima Verosimilitud

Derivemos primero cada término (recordar regla de la cadena),

$$\begin{aligned}
\partial_{\theta_j} log\left(\sum_{k=1}^{K} \pi_k f_k(x_i, \theta_k)\right) &= \frac{1}{\sum_{k=1}^{K} \pi_k f(x_i;\theta_k)} \partial_{\theta_j}\left(\pi_j f(x_j;\theta_k)\right) \\
&= \frac{\pi_j}{\sum_{k=1}^{K} \pi_k f(x_i;\theta_k)} \partial_{\theta_j} f(x_j;\theta_k) \\
&= \frac{\pi_j f(x_i;\theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i;\theta_k)} \boxed{\frac{1}{f(x_i;\theta_j)} \partial_{\theta_j} f(x_j;\theta_k)} \\
&= \frac{\pi_j f(x_i;\theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i;\theta_k)} \partial_{\theta_j} log\left(f(x_i;\theta_k)\right).
\end{aligned}$$

Estimador de Máxima Verosimilitud

Llamemos,

$$w_{ij} = \frac{\pi_j f(x_i; \theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i; \theta_k)}, \tag{8}$$

por lo tanto,

$$\partial_{\theta_j} Loglik(\theta) = \sum_{i=1}^{n} w_{ij} \partial_{\theta_j} log\left(f(x_i; \theta_j)\right).$$

Estimador de Máxima Verosimilitud

Observar:

Maximizar la función de verosimilitud para un modelo mixto es como maximizar una función de verosimilitud con pesos (8).

Estimador de Máxima Verosimilitud

Observar:

Maximizar la función de verosimilitud para un modelo mixto es como maximizar una función de verosimilitud con pesos (8).

GRAN PROBLEMA: los pesos dependen de los parámetros que queremos estimar!!

Es decir,

$$w_{ij} = w_{ij}(\theta).$$

EM-Algoritmo

EM-Algoritmo

Lo vamos a utilizar para resolver nuestro problema particular, pero es método muy versatil. Es una forma general de maximizar la verosimilitud cuando tenemos variables que no pueden ser observadas (Z en nuestro caso).

EM-Algoritmo

Lo vamos a utilizar para resolver nuestro problema particular, pero es método muy versatil. Es una forma general de maximizar la verosimilitud cuando tenemos variables que no pueden ser observadas (Z en nuestro caso).

El algoritmo tiene dos pasos y de ahí su nombre:

- M-step viene de maximización.
- E-step viene de esperanza.

EM-Algoritmo

Lo vamos a utilizar para resolver nuestro problema particular, pero es método muy versatil. Es una forma general de maximizar la verosimilitud cuando tenemos variables que no pueden ser observadas (Z en nuestro caso).

El algoritmo tiene dos pasos y de ahí su nombre:

- M-step viene de maximización.
- E-step viene de esperanza.

Veamos de forma general cómo funciona...

EM-Algoritmo

Primero algunas reducciones en la notación para no morir en el intento:

EM-Algoritmo

Primero algunas reducciones en la notación para no morir en el intento:

Recordemos que tenemos nuestras observaciones $x_1, \ldots, x_n$ que provienen de variables aleatorias $X_1, \ldots, X_n$. Llamemos d al vector que contiene las x's, es decir lo que podemos observar, nuestros datos.

Además tenemos nuestras observaciones ocultas $z_1, \ldots, z_n$ que provienen de variables aleatorias $Z_1, \ldots, Z_n$ discretas. Llamemos h al vector que contiene lo que no podemos observar, osea las variables ocultas.

EM-Algoritmo

Nuestro objetivo es lograr maximizar,

$$Loglik(\theta) = log(p(d, \theta)) \stackrel{*}{=} log\left(\sum_h p(d, h; \theta)\right).$$

donde $p(d, \theta)$ es la densidad o probabilidad puntual de nuestros datos observables.

Observar que $(*)$ se deduce inmediatamente por el teorema de probabilida total.

Desigualdad de Jensen

Sean $t_1, \ldots, t_r$ numeros reales y sean $w_1, \ldots, w_r$ tales que $w_i \in [0, 1]$ con $\sum_i w_i = 1$. Entonces,

$$\sum_{i=1}^{r} w_i log(t_i) \leq log\left(\sum_{i=1}^{r} w_i t_i\right).$$

Sea q una distribución cualquiera sobre h,

$$\begin{aligned} Loglik(\theta) &= log\left(\sum_h p(d,h;\theta)\right) = log\left(\sum_h \frac{q(h)}{q(h)} p(d,h;\theta)\right) \\ &= log\left(\sum_h q(h)\frac{p(d,h;\theta)}{q(h)}\right). \end{aligned}$$

Sea q una distribución cualquiera sobre h,

$$\begin{aligned} Loglik(\theta) &= log\left(\sum_h p(d,h;\theta)\right) = log\left(\sum_h \frac{q(h)}{q(h)} p(d,h;\theta)\right) \\ &= log\left(\sum_h q(h)\frac{p(d,h;\theta)}{q(h)}\right). \end{aligned}$$

Aplicando al desigualdad de Jensen,

$$log\left(\sum_h q(h)\frac{p(d,h;\theta)}{q(h)}\right) \geq \sum_h q(h)\,\mathsf{log}\left(\frac{p(d,h;\theta)}{q(h)}\right) = \mathbb{J}(q,\theta). \tag{9}$$

Hemos probado que para cualquier distribución q sobre h y parámetros θ,

$$Loglik(\theta) \geq \mathbb{J}(q, \theta).$$

Hemos probado que para cualquier distribución q sobre h y parámetros θ,

$$Loglik(\theta) \geq \mathbb{J}(q, \theta).$$

Ahora, consideremos $q(h) = p(h|d;\theta)$, observemos

$$\frac{p(d,h;\theta)}{q(h)} = \frac{p(d,h;\theta)}{p(h|d;\theta)} = \frac{p(d,h;\theta)}{p(d,h;\theta)/p(d;\theta)} = p(d;\theta).$$

Entonces evaluando $\mathbb{J}$ en $p(h|d;\theta)$,

$$\begin{aligned}\mathbb{J}\left(p(h|d;\theta)\right) &= \sum_h p(h|d;\theta)\log\left(p(d;\theta)\right)\\ &= log\left(p(d;\theta)\right)\sum_h p(h|d;\theta)\\ &= log\left(p(d;\theta)\right) = Loglik(\theta).\end{aligned}$$

EM-Algoritmo

EM

1. Comenzar con valores inciales $\theta^{(0)}$ y para los pesos de la mixtura $\pi^{(0)}$.
2. Hasta que no haya cambios repetir:
   - E-step: $q^{(t)} = \arg\max_q \mathbb{J}(q, \theta^{(t)})$.
   - M-step: $\theta^{(t+1)} = \arg\max_\theta \mathbb{J}(q^{(t)}, \theta)$.
3. Devolver los valores finales de θ y q.

El gran truco es lograr maximizar la cota inferior y no directamente la función de verosimilitud. Lo que hace $\mathbb{J}$ es ir empujando la *Loglik* para llegar a un máximo local.

EM-Algoritmo

- E-step $\rightarrow \mathbb{J}(q^{(t)}, \theta^{(t)}) \geq \mathbb{J}(q^{(t-1)}, \theta^{(t)})$.
- E-step $\rightarrow \mathbb{J}(q^{(t)}, \theta^{(t)}) = Loglik(\theta^{(t)})$.
- M-step $\rightarrow \mathbb{J}(q^{(t)}, \theta^{(t+1)}) \geq \mathbb{J}(q^{(t)}, \theta^{(t)})$.

$$
\begin{aligned}
Loglik(\theta^{(t+1)}) &= \mathbb{J}(q^{(t+1)}, \theta^{(t+1)}) \overset{E-step}{\geq} \\
&\geq \mathbb{J}(q^{(t)}, \theta^{(t+1)}) \overset{M-step}{\geq} \mathbb{J}(q^{(t)}, \theta^{(t)}) = \\
&= Loglik(\theta^{(t)}).
\end{aligned}
$$

EM-Algoritmo

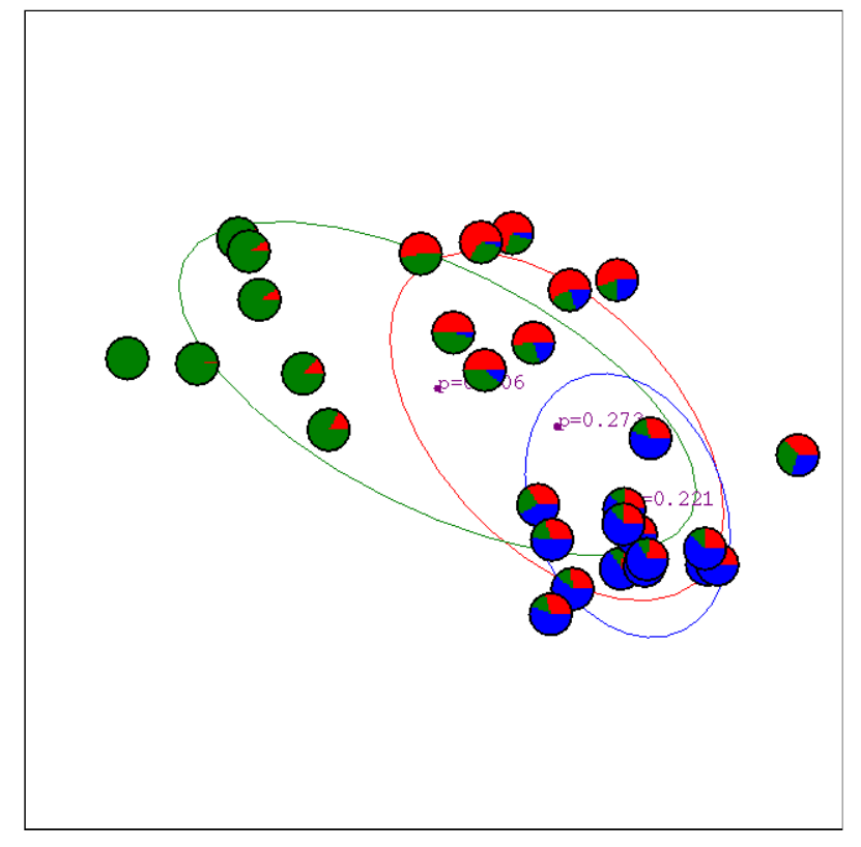

Figura 5: Inicio aleatorio

EM-Algoritmo

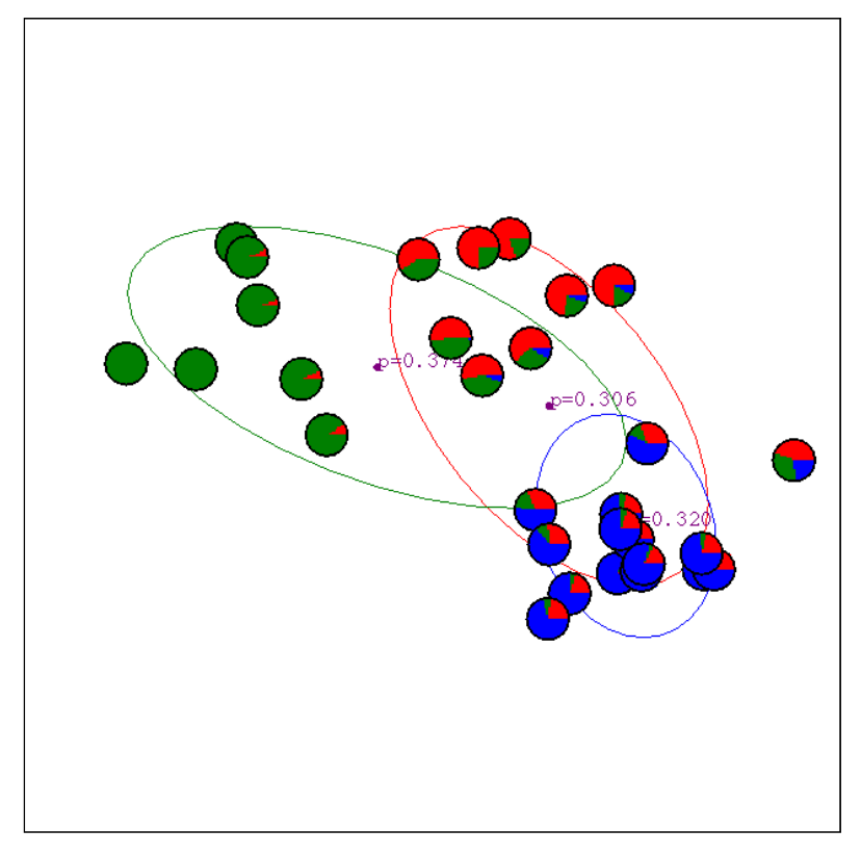

Figura 6: Iteración

EM-Algoritmo

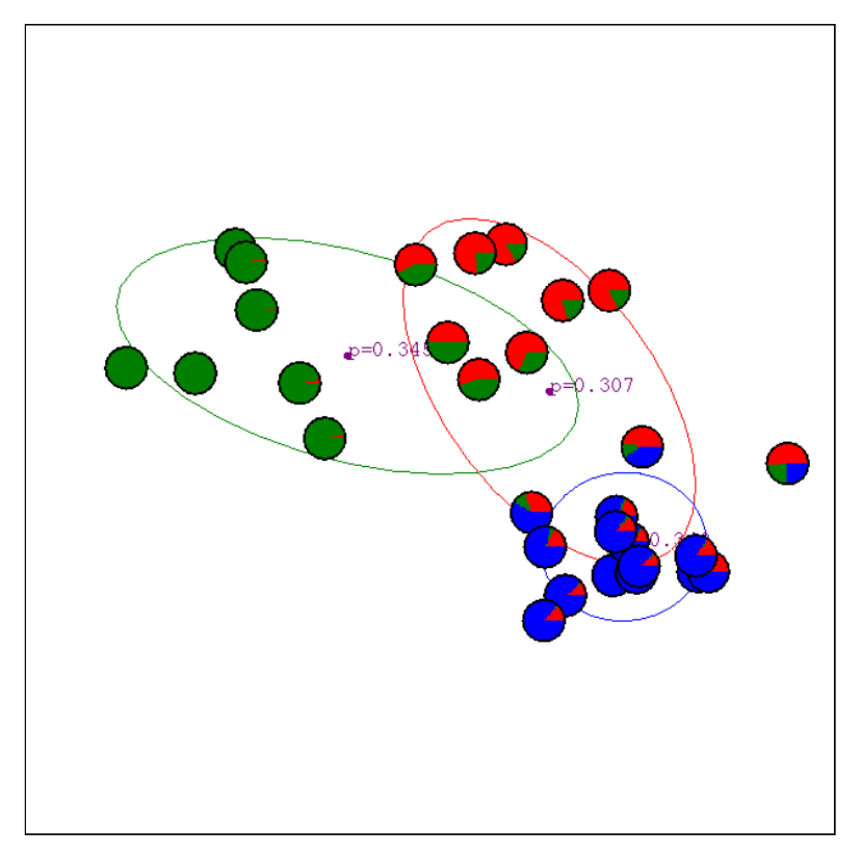

Figura 7: Iteración

EM-Algoritmo

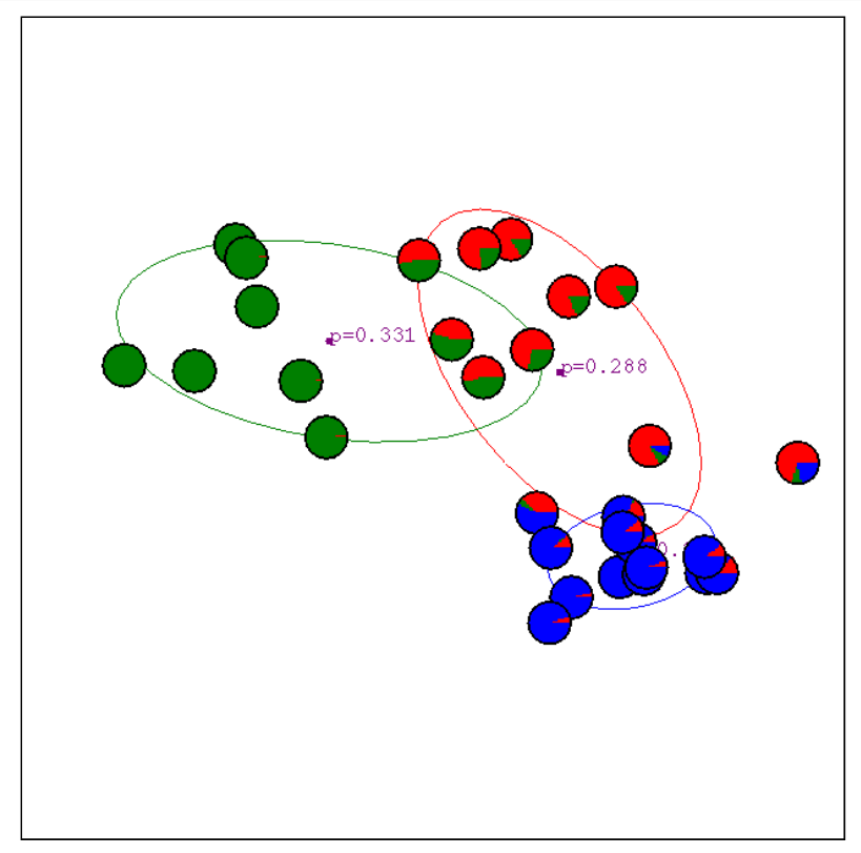

Figura 8: Iteración

EM-Algoritmo

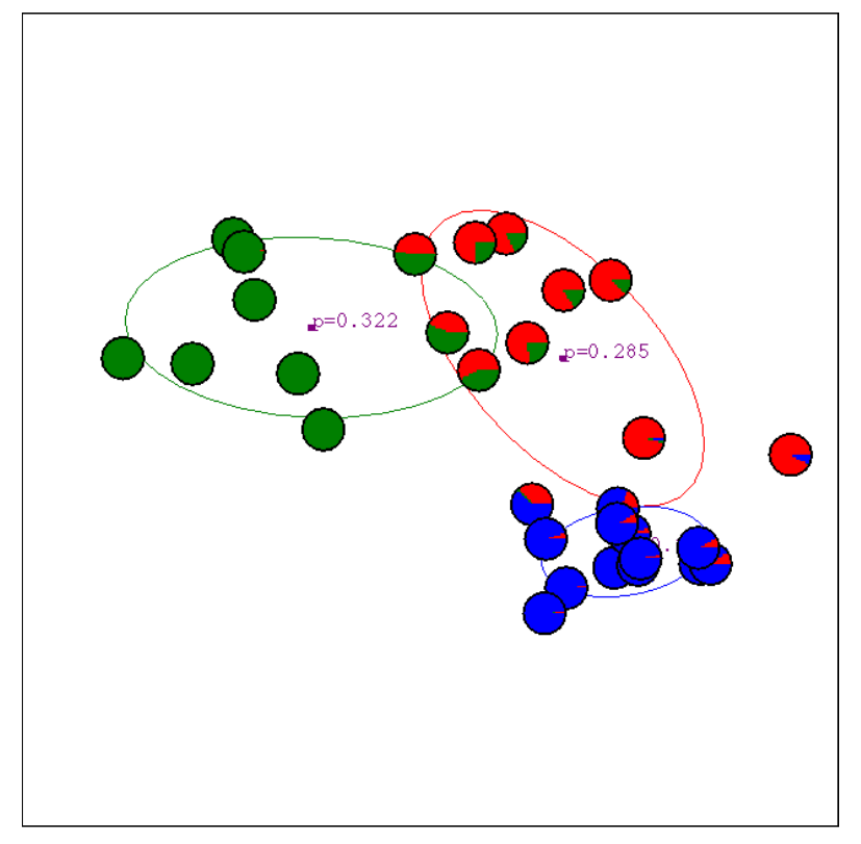

Figura 9: Iteración

EM-Algoritmo

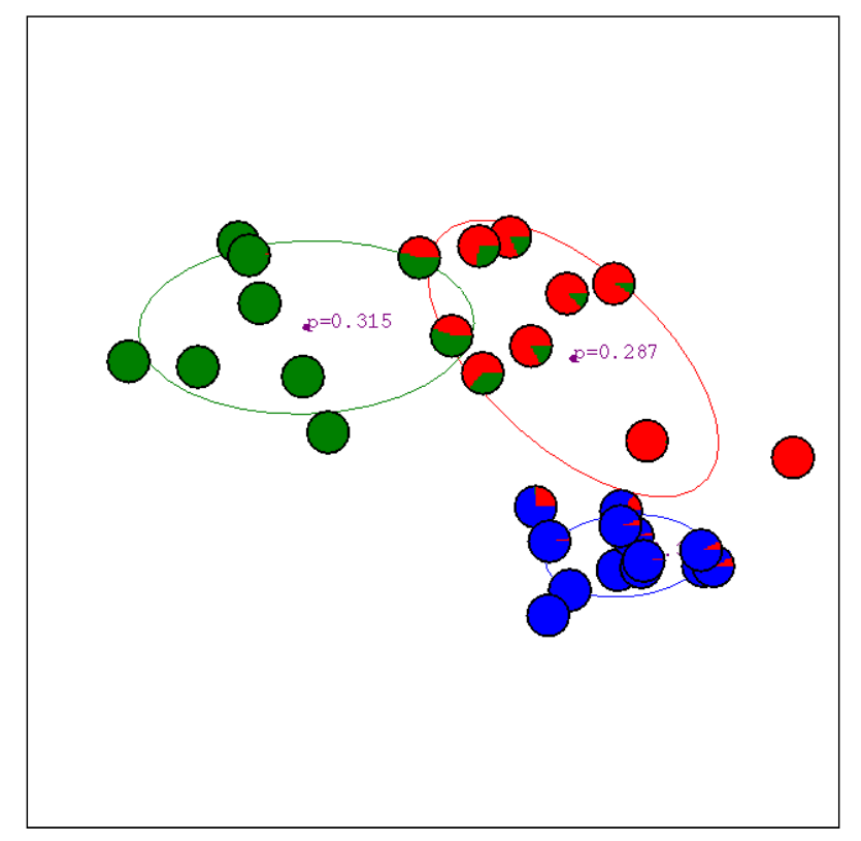

Figura 10: Iteración

EM-Algoritmo

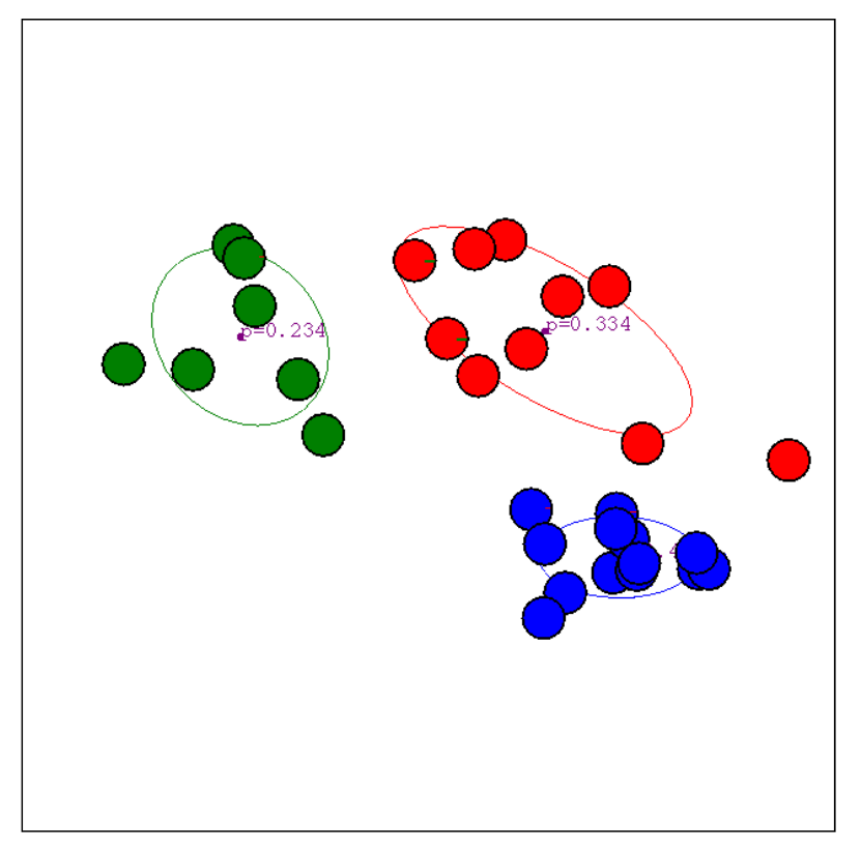

Figura 11: Ya no se modifican los parámetros

COFFEE BREAK