Nota: 10 / 10

Reconocimiento Estadístico de Patrones
Tarea 5

Randy Osbaldo Ibarra Cayo

B. Preguntas Cortas

Pregunta 1 **Nota: 2.5 / 2.5**

Tomamos el primer ejemplo del algoritmo MM, ver `Recpat12n.pdf`, verifica que $h_i(\theta|\theta_m)$ mayoriza $|y_i - \theta|$. Limítate al caso donde $y_i \leq \theta \leq \theta_m$.

Solución

debe ser para y_i <= \theta <= \theta^m

Basta probar que $|y_i - \theta| \leq h_i(\theta|\theta^m)$ para cualquier $\theta \in [\theta^m, y_i]$. Notemos que para cualesquiera $a, b \in \mathbb{R}$, con $0 \leq a$ y $0 < b$, se tiene

$$\begin{aligned} 0 &\leq (a-b)^2 \\ &\leq a^2 - 2ab + b^2 \end{aligned}$$

entonces

$$2ab \leq a^2 + b^2$$

y dado que $b > 0$ se sigue que

$$\begin{aligned} a &\leq \frac{a^2+b^2}{2b} \\ &\leq \frac{1}{2}\left(\frac{a^2}{b} + b\right) \end{aligned}$$ Ok

Tomando $a = |y_i - \theta|$ y $b = |y_i - \theta^m|$, con $\theta \in [\theta^m, y_i]$ y $\theta^m \neq y_i$, se cumple que $0 \leq a$ y $b > 0$, por lo que se obtiene la siguiente desigualdad

$$|y_i - \theta| \leq \frac{1}{2}\left(\frac{|y_i - \theta|^2}{|y_i - \theta^m|} + |y_i - \theta^m|\right)$$ Ok

es decir

$$|y_i - \theta| \leq h_i(\theta|\theta^m)$$

y notemos que la igualdad se cumple para $\theta = \theta^m$

$$\begin{aligned} \frac{1}{2}\left(\frac{|y_i - \theta^m|^2}{|y_i - \theta^m|} + |y_i - \theta^m|\right) &= \frac{1}{2}\left(|y_i - \theta^m| + |y_i - \theta^m|\right) \\ &= \frac{1}{2}\left(2|y_i - \theta^m|\right) \\ &= |y_i - \theta^m| \end{aligned}$$ Ok

Por lo tanto, se concluye que $h_i(\theta|\theta^m)$ mayoriza $|y_i - \theta|$ para cualquier $\theta \in [\theta^m, y_i]$. Bien!

Pregunta 2 **Nota: 2.5 / 2.5**

Usa la desigualdad de Jensen para demostrar que la distancia de KullBack-Leibler es no negativa.

Solución

Sean $P^{(1)}$ y $P^{(2)}$ distribuciones de probabilidad. Dado que $P_i^{(1)}$ y $P_i^{(2)}$ están en el intervalo $(0,1)$ se tiene que

$$0 < \frac{P_i^{(2)}}{P_i^{(1)}} \qquad \text{Ok}$$

por lo que $\frac{P_i^{(2)}}{P_i^{(1)}}$ está en el dominio de la función $-ln$. Por otro lado, notemos que $-ln$ es convexa y $P_i^{(1)} > 0$ para cada i, entonces por la desigualdad de Jensen se tiene que

$$-ln\left[\frac{\sum_i P_i^{(1)}\left(\frac{P_i^{(2)}}{P_i^{(1)}}\right)}{\sum_i P_i^{(1)}}\right] \leq \frac{\sum_i P_i^{(1)}\left[-ln\left(\frac{\sum_i P_i^{(2)}}{\sum_i P_i^{(1)}}\right)\right]}{\sum_i P_i^{(1)}} \qquad \text{Ok}$$

Para el lado izquierdo de la desigualdad se tiene

$$\begin{aligned} -ln\left[\frac{\sum_i P_i^{(1)}\left(\frac{P_i^{(2)}}{P_i^{(1)}}\right)}{\sum_i P_i^{(1)}}\right] &= -ln\left(\frac{\sum_i P_i^{(2)}}{1}\right) \\ &= -ln\left(\sum_i P_i^{(2)}\right) \\ &= -ln(1) \qquad \text{Ok} \\ &= 0 \end{aligned}$$

por otro lado, para el lado derecho de la desigualdad

$$\begin{aligned} \frac{\sum_i P_i^{(1)}\left[-ln\left(\frac{\sum_i P_i^{(2)}}{\sum_i P_i^{(1)}}\right)\right]}{\sum_i P_i^{(1)}} &= \frac{\sum_i P_i^{(1)} ln\left[\left(\frac{\sum_i P_i^{(2)}}{\sum_i P_i^{(1)}}\right)^{-1}\right]}{1} \\ &= \sum_i P_i^{(1)} ln\left(\frac{\sum_i P_i^{(1)}}{\sum_i P_i^{(2)}}\right) \\ &= d(P^{(1)}, P^{(2)}) \qquad \text{Bien!} \end{aligned}$$

Entonces $0 \leq d(P^{(1)}, P^{(2)})$, es decir, la distancia Kullback-Leibler es no negativa.

B. Análisis de Datos

Ejercicio 1 **Nota: 5 / 5**

Implementa el algoritmo **EM** para encontrar grupos en un conjunto de datos en el plano $\{x_i\}$ con una mezcla de K distribuciones Gausianas. Puedes tomar $K = 2$. Lo debes hacer en R pero desde cero sin usar `me()` o `em()`. Elige (construye) algunos conjunto de datos para probar/ilustrar tu algoritmo.

Solución

Conjunto de Datos

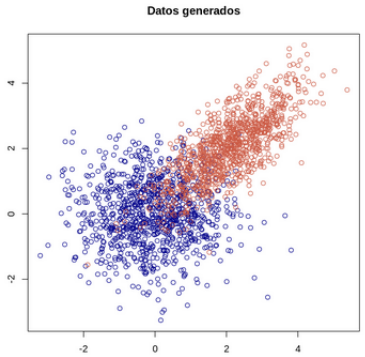
Nos limitamos al caso $d = 2$ y $K = 2$ para simplificar visualización, y suponemos que $p_k \sim \mathcal{N}(\mu_k, \Sigma_k)$ es decir, generamos datos de Gaussianas Bivariadas. Generamos dos muestras de tamaño $N = 100$ con medias

$$\mu_1 = (0, 0), \mu_2 = (2, 2)$$

con matrices de covarianza

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$ Ok

Generamos datos en `R` con ayuda de la función `rmvnorm` del paquete `mvtnorm`, en la figura 1 se presenta una gráfica de la muestra



Ok

Figure 1: Datos Generados a partir de dos Gaussianas

Suponemos entonces que los datos tienen distribución

$$p(x|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(x|z_k, \mu_k, \Sigma_k)$$

donde

- Los $p_k(x|z_k, \mu_k, \Sigma_k)$ son componentes de mezcla, $1 \leq k \leq K$. Cada una es una gaussiana con parámetros μ_k, Σ_k.
- $z = (z_1, ..., z_K)$ es un vector de K variables indicadoras binarias donde una y solo una de las z_k es igual a 1, y las otras son 0. z es una variable aleatoria que representa la identidad del componente de la mezcla que generó x.
- Los $\alpha_k = p(z_k)$ son los pesos de la mezcla, que representan la probabilidad de que el componente k genere una x seleccionada al azar, donde

$$\sum_{k=1}^{K} \alpha_k = 1$$ Ok

EM (Expectation-Maximization)

Definimos el algoritmo **EM (Expectation-Maximization) para mezclas Gaussianas** de la siguiente manera. El algoritmo es un algoritmo iterativo que comienza a partir de una estimación inicial $\Theta = (\{\alpha_k\}, \{\mu_k\}, \{\Sigma_k\})$, y luego procede a actualizar de forma iterativa Θ hasta que la convergencia. Cada iteración consta de un paso E y un paso M.

- E-Step: En este paso se calculan pesos w_{ik}, la probabilidad de que la observación x_i pertenezca a la distribución k, usando los parámetros actuales

$$w_{ik} = \frac{p_k(x_i|z_k, \mu_k, \Sigma_k) \cdot \alpha_k}{\sum_{m=1}^{K} p_m(x_i|z_m, \mu_m, \Sigma_m) \cdot \alpha_m} \qquad \text{Ok}$$

  donde $z = (z_1, ..., z_k)$ es una variable aleatoria que representa la identidad del componente de la mezcla que generó x.

- M-Step: A partir de las probabilidades calculadas en E-Step ajustamos los parámetros mediante

$$\alpha_k^{\text{new}} = \frac{N_k}{N}$$
$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{N} w_{ik} x_i$$
$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{N} w_{ik} (x_i - \mu_k^{\text{new}})(x_i - \mu_k^{\text{new}})^T \qquad \text{Ok}$$

Criterios de convergencia

El algoritmo EM es iterativo, por lo que podemos usar como criterios de convergencia

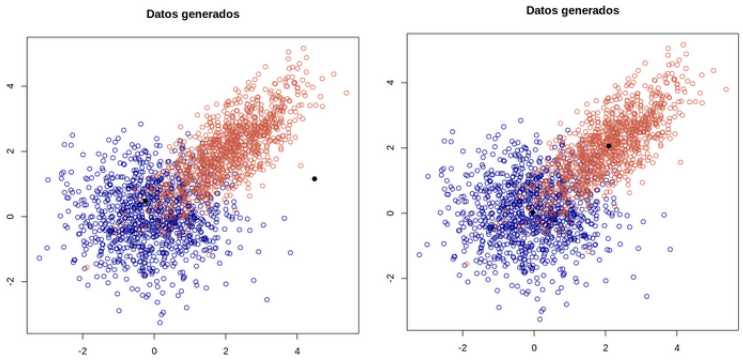
- Funcion de log-Verosimilitud: Cuando el cambio en esta funcion es suficientemente pequeño (menor que un $\epsilon > 0$) detenemos el algoritmo. En este caso se definió $\epsilon = 0.01$.
- Máximo número de iteraciones: Se puede definir un máximo número de iteraciones para evitar caer en un bucle infinito en caso de que el algoritmo se estanque. En este caso se definió un máximo de 100.

Inicialización Aleatoria

Primero tomamos parámetros cualesquiera

- Inicializamos α_k, suponemos la probabilidad de que el dato provenga de la componente k es igual para toda k en $[1:K]$.
- Inicializamos μ_k's a partir de una normal centrada en $(0,0)$ con covarianza $5 \cdot I_{d\times d}$.
- Generamos las matrices de covarianza Σ_k, por simplicidad suponemos cada una como $I_{d\times d}$.

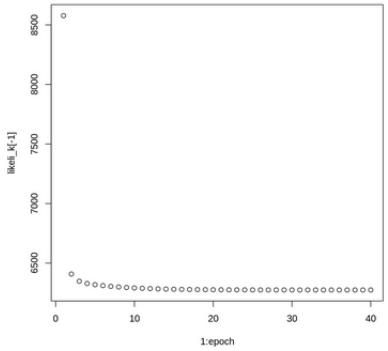
En la figura 2 podemos observar (puntos negros) la estimación inicial de cada media y las estimaciones finales.



Ok

Figure 2: Izquierda Medias iniciales, Derecha Medias Finales

En la figura 3 se muestra la grafica del valor de -logVerosimilitud en cada iteración



Bien!, haber incluido un análisis de convergencia.

Figure 3: -logVerosimilitud en cada iteración

Podemos observar que converge rápidamente, con un total de 40 iteraciones, además los parámetros encontrados se parecen a los de las distribuciones originales, como podemos notar en la figura 4 donde se aprecian

$$\mu_1 = [-0.038, 0.024], \mu_2 = [2.093, 2.059]$$
$$\Sigma_1 = \begin{bmatrix} 1.144 & 0.009 \\ 0.009 & 1.005 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.976 & 0.73 \\ 0.730 & 0.98 \end{bmatrix}$$

```
[1] "Mu_k Calculadas"
             [,1]       [,2]
[1,] -0.03834089 0.02409641
[2,]  2.09376323 2.05907230
[1] "Sigma_k Calculadas"
             [,1]       [,2]
[1,] 1.144435550 0.009396527
[2,] 0.009396527 1.005794513
[3,] 0.976221071 0.730373418
[4,] 0.730373418 0.982062529
```

Figure 4: Parametros encontrados con EM inicializando de manera aleatoria

Inicialización con $K-$Medias

Tomamos los parámetros que nos devuelve $K-$Medias: Las medias y a partir de la clusterización del método estimamos matrices de covarianza iniciales. En la figura 5 podemos observar (puntos negros) la estimación inicial de cada media y las estimaciones finales. Casi no hay diferencia visual

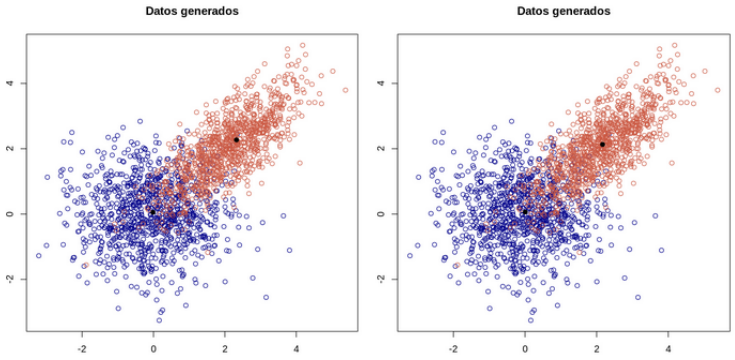
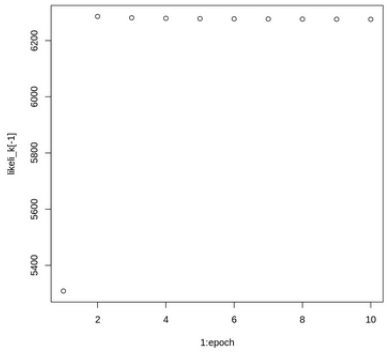


Figure 5: Izquierda Medias iniciales, Derecha Medias Finales

En la figura 6 se muestra la grafica del valor de -logVerosimilitud en cada iteración



Ok

Figure 6: -logVerosimilitud en cada iteración

Podemos observar que converge aún más rápidamente, con un total de 10 iteraciones, y los parámetros encontrados se parecen a los de las distribuciones originales, como podemos notar en la figura 7 donde se aprecian

$$\mu_1 = [0.004, 0.062], \mu_2 = [2.165, 2.131]$$

$$\Sigma_1 = \begin{bmatrix} 1.152 & 0.053 \\ 0.053 & 1.013 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.901 & 0.661 \\ 0.661 & 0.908 \end{bmatrix}$$

```
[1] "Mu_k Calculadas"
             [,1]       [,2]
[1,] 0.004900674 0.06259929
[2,] 2.165841743 2.13103238
[1] "Sigma_k Calculadas"
             [,1]       [,2]
[1,] 1.15290490 0.05333096
[2,] 0.05333096 1.01328295
[3,] 0.90133280 0.66131337
[4,] 0.66131337 0.90840697
```

Figure 7: Parametros encontrados con EM inicializando de manera aleatoria

Concluisiones

Podemos notar que el método logra aproximar bien las distribuciones en ambos casos:

- Inicialización aleatória de parámetros.
- Inicialización de parámetros con $K-$Medias.

Comparando ambas inicializaciones podemos notar que la inicialización con $K-$Medias alcanza la convergencia en un número menor de iteraciones. Podemos notar que el método inicializado de manera aleatória logra ligeramente una mejor aproximación de los parámetros.

Por otro lado, buscamos minimizar la funcion -logVerosimilitud y en los experimentos observamos que la inicialización con las medias μ_k de $K-$Medias con $\Sigma_k = I_{d\times d}$ logró un valor de -logVerosimilitud menor que el valor obtenido con los parámetros encontrados por el método EM.

Referencias

- Notas del curso
- https://www.ics.uci.edu/ smyth/courses/cs274/notes/EMnotes.pdf

Hubiese sido interesante mostrar una comparación en un caso más complejo, por ejemplo, donde las distribuciones están más cercanas o forman alguna estructura particular.

También se puede completar el análisis comparativo entre EM y K-means calculando métricas de desempeño para clustering: Por ejemplo, se podría hacer un análisis de siluetas.