Modelos Mixtos

Nicolás Kossacoff

Octubre 2024

1. Introducción

La idea detrás de los **Modelos Mixtos** es que nuestro conjunto de datos, D, proviene de una variedad de distribuciones, cada una representando un cluster.

A diferencia de aprendizaje supervisado, en donde contábamos con la probabilidad de $X|K$ y en donde podíamos fácilmente calcular las probabilidades de cada clase para luego estimar la probabilidad de $K|X$, en no supervisado no contamos ni con las probabilidades de las clases ni con las probabilidades de $X|K$ (justamente porque no conocemos K).

Por lo tanto, debemos primero estimar estas probabilidades para luego poder estimar las probabilidades de $K|X$. Este modelo nos propone utilizar una distribución de mixtura para lograr esto.

2. Mixturas

Decimos que una distribución f es una **mixtura** de K distribuciones si la podemos definir como la suma ponderada de las distribuciones:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{1}$$

donde π_k son los **pesos** de la mixtura, con $\pi_k > 0$ y $\sum_k \pi_k = 1$. Si hacemos un paralelismo con los modelos de aprendizaje supervisado, π_k se puede interpretar como la probabilidad de que una observación X pertenezca a la clase K.

Definimos el vector de parámetros de la densidad f_k como θ_k, tal que:

$$f(x, \theta) = \sum_{k=1}^{K} \pi_k f_k(x, \theta_k) \tag{2}$$

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

La Ecuación (1) define un modelo estocástico completo. Empíricamente este modelo se puede representar de la siguiente manera:

$$Z \sim Cat(\pi_1, \ldots, \pi_K) \tag{3}$$
$$X|k \sim f_k \tag{4}$$

donde la variable Z es una variable categórica que determina de que densidad f_k proviene la observación X.

Las funciones de densidad, f_k, pueden ser elegidas de manera arbitraria sin afectar a nuestro modelo. Sin embargo, en la práctica es común utilizar **modelos paramétricos**, en donde estas densidades provienen de una misma familia (e.g., son todas densidades Gaussianas con distintas medias y varianzas).

2.1. Estimación

Para estimar la Ecuación (2) utilizamos el método de **máxima verosimilitud**. Este método estima los parámetros en θ tal que hagan más probable observar nuestro conjunto de datos, D.

La función de máxima verosimilitud, asumiendo independencia entre las observaciones, es:

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{5}$$

Aplicando logaritmos en ambos lados, y reemplazando por la Ecuación (2), obtenemos la función de log-verosimilitud:

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{i=1}^{n} \log\left(f(x_i, \theta)\right) \\ &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \log\left(\pi_k f(x_i, \theta_k)\right)\right)\end{aligned}$$

Si derivamos con respecto a un parámetro θ_j obtenemos:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_j} &= \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)} \pi_j \frac{\partial f(x_i, \theta_j)}{\partial \theta_j} \\ &= \sum_{i=1}^{n} \frac{\pi_j f(x_i, \theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)} \frac{1}{f(x_i, \theta_j)} \frac{\partial f(x_i, \theta_j)}{\partial \theta_j} \\ &= \sum_{i=1}^{n} \frac{\pi_j f(x_i, \theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)} \frac{\partial \log\left(f(x_i, \theta_j)\right)}{\partial \theta_j}\end{aligned}$$

Entonces, maximizar la verosimilitud para un modelo mixto es equivalente a maximizar una **función de verosimilitud ponderada**, donde el peso para una observación

x_i es:

$$w_{i,j} = \frac{\pi_j f(x_i, \theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)} \tag{6}$$

Si miramos más en detalle los pesos, vamos a encontrar que en el numerador tenemos la probabilidad conjunta de $X = x_i \cap Z = z_j$, y en el denominador tenemos la probabilidad de obtener $X = x_i$ (i.e., equivalente a cuando utilizamos la ley de probabilidad total). Entonces, los pesos de la función de verosimilitud son iguales a la probabilidad condicional de $Z = z_j | X = x_i$:

$$w_{i,j} = \frac{\pi_j f(x_i, \theta_j)}{\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)} = P(Z = z_j | X = x_i, \theta) \tag{7}$$

El problema es que estos pesos dependen de los mismos parámetros que estamos intentando estimar.

3. Algoritmo EM

El algoritmo EM es un método que nos permite maximizar la función de verosimilitud en contextos en los que tenemos variables no observables, como es nuestro caso con la variable Z.

Hacemos una reducción en la notación para simplificar las cuentas. Las observaciones $x_1, \ldots, x_n$ las vamos a llamar d mientras que a las observaciones ocultas (i.e., etiquetas) $z_1, \ldots, z_n$ las llamamos h. También presentamos la **desigualdad de Jensen**, la cual vamos a utilizar más adelante:

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

Ahora, lo que nosotros queremos es maximizar la función de log-verosimilitud:

$$\mathcal{L}(\theta) = \log\left(\sum_{h} f(d, h, \theta)\right) \tag{8}$$

Definimos una distribución arbitraria para las etiquetas, $q(h)$, y nos queda:

$$\begin{aligned}
\mathcal{L}(\theta) &= \log\left(\sum_{h} f(d, h, \theta)\right) \\
&= \log\left(\sum_{h} \frac{q(h)}{q(h)} f(d, h, \theta)\right) \\
&= \log\left(\sum_{h} q(h) \frac{f(d, h, \theta)}{q(h)}\right)
\end{aligned}$$

Luego, si aplicamos la desigualdad de Jensen, nos queda:

$$\mathcal{L}(\theta) = \log\left(\sum_h q(h)\frac{f(d,h,\theta)}{q(h)}\right) \geq \sum_h q(h)\log\left(\frac{f(d,h,\theta)}{q(h)}\right) \equiv J(q,\theta)$$

Ahora, si elegimos $q(h) = f(h|d,\theta)$, entonces se cumple que:

$$\frac{f(d,h,\theta)}{q(h)} = \frac{f(d,h,\theta)}{f(h|d,\theta)} = \frac{f(d,h,\theta)}{f(d,h,\theta)/f(d,\theta)} = f(d,\theta)$$

lo cual implica que $J(q,\theta) = \mathcal{L}$.

Entonces, la idea del algoritmo es maximizar la cota inferior, $J(q,\theta)$, para de esa manera poder maximizar la función de verosimilitud y obtener una estimación para la probabilidad de $Z|X$.

A continuación se encuentra un resumen del algoritmo:

1. Comenzamos asignando valores iniciales a los parámetros de la distribución y a los pesos, $\theta^{(0)}$.
2. En cada iteración $t = \{1,\ldots,T\}$ hay dos pasos a seguir:
   (a) Dado los valores de los parámetros, $\theta^{(t)}$, buscamos la probabilidad $q^{(t)}$ que maximiza $J(q,\theta^{(t)})$. Es decir:

   $$q^{(t)} = \arg\max_q J(q,\theta^{(t)})$$

   En este caso, la probabilidad $q^{(t)}$ que maximiza $J(q,\theta^{(t)})$ es igual a la probabilidad condicional $f(h|d,\theta)$, así que en este paso sabemos que $J(q,\theta) = \mathcal{L}$.

   (b) Dado $q^{(t)}$, buscamos los parámetros $\theta^{(t+1)}$ que maximizan $J(q^{(t)},\theta)$:

   $$\theta^{(t+1)} = \arg\max_\theta J(q^{(t)},\theta)$$

   Después de este paso nos queda que $J(q^{(t)},\theta^{(t+1)}) \geq J(q^{(t)},\theta^{(t)}) = \mathcal{L}(\theta^{(t)})$, pero todavía es menor a $\mathcal{L}(\theta^{(t+1)})$, porque para eso necesitamos encontrar la probabilidad condicional dado los parámetros $\theta^{(t+1)}$, es decir, tenemos que encontrar $q^{(t+1)}$, por lo que debemos volver al paso anterior.

   Este procedimiento se repite hasta que no hayan cambios significativos en las estimaciones.
3. Devuelve los valores estimados para θ y para la probabilidad condicional q.

La solución que alcancemos depende del punto de inicio que tomamos, $\theta^{(0)}$, por lo que es recomendable iterar este procedimiento con diferentes valores iniciales de los parámetros y quedarse con los mejores resultados.