EST-297

Enfoques estadísticos de Clustering

Juan Zamora O.

Junio, 2024.

INSTITUTO DE ESTADÍSTICA
IES PUCV

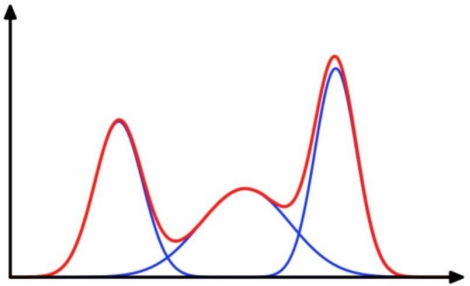
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Estructura de la Presentación

1. Modelo de Mezcla de Gaussianas

Modelo de Mezcla de Gaussianas

- Si bien la distribución Normal tiene importantes propiedades analíticas, tiene también serias limitaciones para modelar fenómenos reales complejos.
- Una manera abordar estas limitaciones consiste en utilizar una combinación de distintas distribuciones normales o gausianas.



Definición del problema

Se tiene un conjunto de puntos $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ consistente de N observaciones de una variable aleatoria d-dimensional **x**. La variable aleatoria **x** se asume distribuida de acuerdo a una mezcla de K componentes. Es decir,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k p(\mathbf{x}|\theta_k)$$

Cada densidad $\mathcal{N}(\mu_k, \Sigma_k)$ se denomina componente de la mezcla y tiene sus propios parámetros. Los coeficientes de la mezcla α_i satisfacen: - $0 \leq \alpha_i \leq 1$ - $\sum_i^K \alpha_i = 1$

Definición alternativa

Una manera alternativa de plantear este problema es mediante una variable aleatoria categórica z_n (asociada a un punto cualquiera $\mathbf{x}_n$) que toma valores sobre $1 \dots K$ con probabilidades $p(z_n = k) = \alpha_k$. Esto último también puede ser expresado como $p(z_{nk} = 1) = \alpha_k$ para un punto cualquiera $\mathbf{x}_n$.

En lugar de usar una sola variable categórica z_n, podemos introducir el vector binario aleatorio K-dimensional $\mathbf{z}_n$ para anotar la etiqueta del componente para $\mathbf{x}_n$. De esta forma, el vector $\mathbf{z}_n$ solo tendrá un 1 en la k-ésima posición asociada al componente que da origen a $\mathbf{x}_n$ y un 0 en todos los demás.

Ejemplo

Por ejemplo, para $K = 3$ clusters, una observación $\mathbf{x}_n$ que corresponda al cluster donde $z_{n2} = 1$, entonces $\mathbf{z}_n$ será representado por el vector columna $\mathbf{z}_n = (0, 1, 0)$. Luego, la distribución marginal sobre $\mathbf{z}_n$ es:

$$p(\mathbf{z}_n) = \alpha_1^{z_{n1}} \alpha_2^{z_{n2}} ... \alpha_K^{z_{nK}} = \prod_{k=1}^{K} \alpha_k^{z_{nk}}$$

Similarmente, la distribución condicional de $\mathbf{x}_n$ dado $\mathbf{z}_n$ puede ser expresada de la forma

$$p(\mathbf{x}_n|\mathbf{z}_n) = \prod_{k=1}^{K} p(\mathbf{x}_n|\theta_k)^{z_{nk}}$$

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \alpha_k p(\mathbf{x}_n|\theta_k)$$

Los parámetros que deben ser inferidos son $\{\alpha_1, \alpha_2, \dots, \alpha_K, \theta_1, \dots, \theta_K\}$. Si suponemos que los puntos son generados independientemente a partir de la distribución, entonces la verosimsilitud queda dada por:

$$p(\mathbf{X}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \alpha_k p(\mathbf{x}_n|\theta_k)$$

Generalmente, se utiliza una forma logaritmica de la verosimilitud para deshacerse de la productoria.

El modelo de mezclas más conocido es el de gausianas **(GMM)**. En este, cada componente corresponde a una distribución gausiana, es decir:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k p(\mathbf{x}|\theta_k) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

Usando una cantidad suficiente de gausianas, ajustando sus medias y covarianzas así también como los coeficientes de la combinación lineal, se puede aproximar cualquier densidad continua.

La distribución resultante está governada por los parámetros α, μ y Σ. Una manera de encontrar los valores para estos parámetros es mediante máxima verosimilitud:

$$\log p(X|\alpha, \mu, \Sigma) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right)$$

El algoritmo EM

Una manera elegante para encontrar soluciones de máxima verosimilitud para modelos con variables latentes es el algoritmo de Maximización de la esperanza o **EM**.

Las Probabilidades a posteriori (*responsabilities*) indican que tanto explica el componente k a la observación x. Se expresan mediante:

$$p(z_k = 1|\mathbf{x}) = \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} = \frac{\alpha_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}$$

- Al derivar la función de verosimilitud respecto de las medias μ_k e igualar a 0 se puede obtener

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n} p(z_{ik} = 1|x) x_i$$

donde $n_k = \sum_{i=1}^{n} p(z_{ik} = 1|x)$.

- Al hacer lo mismo pero respecto de Σ_k se obtiene

$$\Sigma_k = \frac{1}{n_k} \sum_{i=1}^{n} p(z_{ik} = 1|x)(x_i - \mu_k)(x_i - \mu_k)^T$$

- Por último, podemos realizar el mismo procedimiento pero respecto de α_k (este paso requiere incorporar una restricción para la suma convexa de los α_i) y se obtiene

$$\alpha_k = \frac{n_k}{n}$$

Notar que para cada una de estas 3 cantidades existen dependencias entre los mismos parámetros a través de $p(z_{ik} = 1|x)$

Luego, se utiliza un procedimiento iterativo en dos pasos para actualizar cada parámetro: El paso de esperanza (E) y el de maximización (M).

Se escogen valores iniciales para las medias, covarianzas y coeficientes de mezcla. Se calculan la probabilidades a posteriori y luego, se usan estas probabilidades para re-estimar las medias, covarianzas y coeficientes de mezcla.

El algoritmo EM toma bastantes más iteraciones que K-means. Una alternativa es usar este último para encontrar una solución inicial, luego calcular las matrices de covarianza para cada grupo y finalmente, usar esta información para inicializar la mezcla de gausianas.