EST-297

Enfoques estadísticos de Clustering

Juan Zamora O.

Junio, 2024.

INSTITUTO DE ESTADÍSTICA
IES PUCV

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

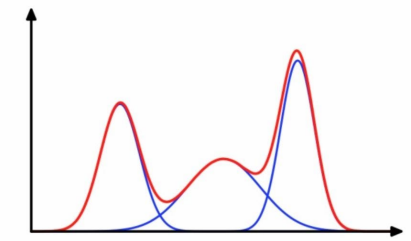
Estructura de la Presentación

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.

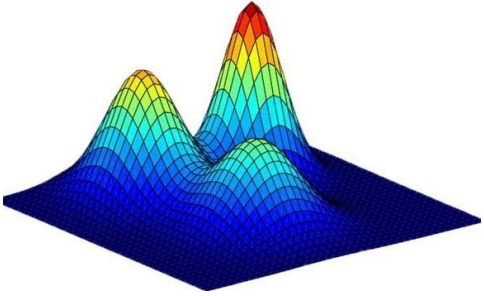
$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x; \mu_k, \sigma_k^2), \text{ con } \alpha_i \geq 0, \sum_k \alpha_k = 1$$



Modelo multivariado

- Dado un conjunto de puntos $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ con $\mathbf{x}_i \in \mathbf{R}^d$
- La variable aleatoria se supone distribuida de acuerdo a una mezcla de K componentes normales multivariados. Es decir,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x; \underline{\mu}_k, \Sigma_k), \text{ con } \sum_k \alpha_k = 1$$



Definición del problema

- Dada la representación mediante K componentes normales, el objetivo es estimar los valores de los parámetros $\underline{\mu}$, Σ y α

Si suponemos que los puntos son i.i.d 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.

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

El algoritmo EM

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

El algoritmo comienza con una inicialización aleatoria de los parámetros del modelo. Luego, itera entre los siguientes dos pasos hasta converger:

1. **Calculo de la esperanza (E)**: Cálculo de log-verosimilitud esperada del modelo con respecto a la distribución de las variables latentes, dados los datos observados y la estimación actual de los parámetros.
2. **Maximización de log-verosimilitud (M)**: Actualización de los parámetros del modelo para maximizar la log-verosimilitud de los datos observados, dadas las variables latentes observadas a partir del paso anterior.

Paso E

La variable latente a calcular corresponde a la pertenencia de cada uno de los n puntos en cada uno de los K componentes gausianos.
Sea $Z_i = k$ la variable que indica que x_i pertenece al componente k. Luego,

$$p(Z_i = k|x_i;\theta) = \frac{p(x_i|Z_i = k;\theta)p(Z_i = k;\theta)}{p(x_i;\theta)} = \frac{\alpha_k \cdot p(x_i|Z_i = k;\theta)}{\sum_{j=1}^{K} \alpha_j p(x_i|Z_i = j;\theta)}$$

Considerando que $p(x_i|Z_i = k;\theta) = \mathcal{N}(x_i;\underline{\mu}_k,\Sigma_k)$,

$$p(Z_i = k|x_i;\theta) = \frac{}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(x_i;\underline{\mu}_k,\Sigma_k)} = \gamma(z_{ik})$$

$\gamma(z_{ik})$ se denomina responsabilidad del componente k por la observación i

La log-verosimilitud esperada con respecto a la distribución de las variables latentes se escribe como una suma ponderada de las log-verosimilitudes de todos los puntos bajo cada uno de los componentes:

$$Q(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma(z_{ik}) \log \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

Esta función Q representa la log-verosimilitud esperada sobre los datos observados y las distribuciones estimadas de variables latentes

Paso M

- θ corresponde a los vectores de medias, matrices de covarianza y pesos de mezcla.
- En este paso se actualizan los parámetros en θ de manera que se maximice la log-verosimilitud esperada $Q(\theta)$
- Las **medias** de cada componente se actualizan mediante

$$\underline{\mu}_k^* = \frac{\sum_{i=1}^n \gamma(z_{ik})\mathbf{x}_i}{\sum_{i=1}^n \gamma(z_{ik})}$$

  - El vector de medias del k-ésimo componente corresponde a una promedio ponderado de todos los puntos, usando las probabilidades de pertenencia de cada punto
  - Esta ecuación proviene de la maximización de la log-verosimilitud esperada (Q) con respecto al vector de medias $\rightarrow \frac{\partial Q}{\partial \underline{\mu}_k} = 0$

- Las **covarianzas** para componente se actualizan mediante

$$\Sigma_k^* = \frac{\sum_{i=1}^n \gamma(z_{ik})(\mathbf{x}_i - \underline{\mu}_k^*)(\mathbf{x}_i - \underline{\mu}_k^*)^T}{\sum_{i=1}^n \gamma(z_{ik})}$$

  - La nueva matriz de covarianza corresponde a un promedio ponderado de las desviaciones de cada punto desde la media del componente.
- Los pesos de la mezcla se actualizan mediante

$$\alpha_k^* = \frac{\sum_{i=1}^n \gamma(z_{ik})}{n}$$

  - El nuevo peso del componente k-ésimo corresponde a la probabilidad total de los puntos que pertenecen a este componente, normalizado por la cantidad de puntos

Inicialización y convergencia

- Este procedimiento converte a un máximo local
- Se repite hasta que la función de verosimilitud sufra cambios muy pequeños o permanezca constante.
- 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.