Clustering mezclado y basado en densidades

por Jose Naranjo

Clustering

- Definición general: particionar objetos de información (patrones, entidades, instancias, observaciones, unidades) en un número determinado de clústers (grupos, subconjuntos, categorías).
- Definiciones operativas:
  1. Crear un conjunto de entidades similares entre ellas y distintas con otras.
  2. Agregar puntos en un espacio de manera que la distancia entre dos puntos en el clúster sea menor que la distancia entre cualquier punto en el clúster y cualquier otro punto fuera de este.
  3. Crear regiones conjuntas en un espacio de características con un número relativamente alto de puntos de densidad, separado de otras regiones con un número relativamente bajo de puntos de densidad.

Clustering

- Definición matemática:
  - Dado un conjunto de patrones de entrada $X = \{x_1, \dots, x_j, \dots, x_N\}$, donde $x_j = (x_{j1}, x_{j2}, \dots, x_{jd}) \in R^d$ y cada medida x_{ji} siendo una característica:
  - Se intenta encontrar una K-partición de X llamada $C = \{C_1, \dots, C_K\}$ $(K \leq N)$ de manera que:
    - $C \neq \emptyset, i = 1, \dots, K$
    - $\cup_{i=1}^{K} C_i = X$
    - $C_i \cap C_j = \emptyset, i, j = 1, \dots, K$
    - $i \neq j$

Clustering particional

- Consiste en asignarle los puntos de un conjunto de datos a un conjunto de clústers sin alguna estructura jerárquica.
- En la organización de los clústers, se busca maximizar o minimizar una función criterio.
- En principio, la partición óptima está basada en la función criterio y en la enumeración de todos los posibles clústers.
- Es una especie de método de fuerza bruta que en la práctica no es factible porque requiere de mucha computación (i.e. enumerar todos los clústers posibles es muchas veces imposible).
- Por ejemplo, agrupar 30 objetos en 3 clústers implica un número de particiones de 2×10^{14}.

Criterios de clustering

- Para resolver estos problemas se utilizan algoritmos heurísticos que retornan soluciones aproximadas.
- La idea es agrupar datos que son homogéneos en sus clústers y bien separados de los demás clústers. La homogeneidad y separación se evalúa a través de las funciones criterio.
- Una función criterio muy utilizada es el criterio del error de la suma de cuadrados que se define como:
  - $J_s(\Gamma, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} \Upsilon_{ij} \left\| x_j - m_i \right\|^2$
  - Donde Γ es la matriz de particiones, M la matriz de centroides y m es el promedio de la muestra.

El algoritmo K-means

- Uno de los más conocidos y populares algoritmos de clustering que busca particionar la data de manera óptima minimizando el criterio del error de la suma de cuadrados de manera iterativa, hill-climbing.
- El procedimiento se resume en 5 pasos:
  1. Inicializar una K-partición aleatoriamente o basada en información previa (previamente definiendo el valor de K)
  2. Calcular la matriz de centroides
  3. Asignar cada objeto del conjunto de datos al clúster más cercano
  4. Recalcular la matriz de centroides basándose en la partición actual
  5. Repetir los pasos 3 y 4 hasta que no haya cambio sustancial en los clústers (e.g. hallar el "codo")

Problemas con K-means

- Como el algoritmo se basa en promedios para hallar el error, es sensible a valores extremos y ruido. El cálculo de los promedios incorpora toda la data del clúster.
- La definición de promedio limita la aplicación de K-means a variables numéricas y excluye a las categóricas. Para resolver esto, se han propuesto conversiones binarias para representar datos categóricos (e.g. dummifying). Para estas modificaciones al algoritmo, la distancia entre un par de puntos categóricos se mide en términos del número de faltas de coincidencias de las características.
- K-means se ha ido modificando para poder recibir distintos tipos de información (e.g. K-modes, K-medoids, Huang-Gupta)

Clustering mezclado y basado en densidades

- Desde un punto de vista probabilístico, cada objeto de data se asume que se genera de una distribución latente de probabilidad.
- Estas fuentes de probabilidad pueden tener distintas formas funcionales como Gaussianos multivariados, distribuciones T, u otras familias con distintos parámetros.
- En teoría, el clustering mezclado involucra estimar los parámetros de los modelos latentes para hallar los clústers.
- Es una metodología de inferencia, que busca una solución única a los parámetros que generan la distribución de datos.

Clustering mezclado y basado en densidades

- Consideremos el siguiente proceso para generar datos:
  - Se asume que se conoce el número de clústers K
  - Se genera un objeto x a través de una densidad de probabilidad condicional $p(x|C_i, \theta_i)$ donde C es el clúster y θ el vector de parámetros desconocidos
  - La probabilidad de que x venga de C depende de la probabilidad $P(C)$, también llamado el parámetro de mezclado.
- Así, la densidad de probabilidad para el conjunto entero se puede representar a través de:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{K} p(\mathbf{x}|C_i, \boldsymbol{\theta}_i) P(C_i),$$

Clustering mezclado y basado en densidades

- El modelo utiliza la fórmula de Bayes para hallar la probabilidad posterior:

$$P\left(C_i \middle| \mathbf{x}, \hat{\boldsymbol{\theta}}\right) = \frac{P(C_i)\, p\left(\mathbf{x} \middle| C_i, \hat{\boldsymbol{\theta}}_i\right)}{p\left(\mathbf{x} \middle| \hat{\boldsymbol{\theta}}\right)},$$

- Y la estimación de Maximum Likelihood para estimar el parámetro desconocido θ:

$$\frac{\partial l\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\boldsymbol{\theta}}_i} = 0, \quad i = 1, \cdots, K.$$