APRENDIZAJE NO SUPERVISADO
MÁSTER EN INTELIGENCIA ARTIFICIAL
UNIVERSIDAD INTERNACIONAL DE VALENCIA

TÉCNICAS DE CLUSTERING

Componentes del grupo:

Coordinadora: Carmen Raposo Jiménez
Secretario: Víctor Dot Mariano
Revisor: Ricardo Sánchez Pastor

7 Abril de 2019

Índice general

Apartado 1

Conjuntos de Datos

Se han escogido dos conjuntos de datos distintos para las dos casuísticas planteadas: uno conociendo la clasificación real de los datos y otro sin saberlo.

1.1. Agrupamiento Real Conocido

Como dataset para el problema cuyo agrupamiento real es conocido se ha seleccionado el ya cargado en las librerías de sklearn, iris, el cual categoriza distintos tipos de iris en tres especies. Este conjunto de datos tiene las siguiente características:

Número de instancias	150
Número de atributos	4
Descripción de atributos	sepal-length. Numérico. Longitud del sépalo. sepal-width. Numérico. Anchura del sépalo. petal-length. Numérico. Longitud del pétalo. petal-width. Numérico. Anchura del pétalo.
Tipo de datos de la clase	Categórico: iris-setosa, iris-versicolor, iris-virginica
Descripción de la clase	Nombre: class. Descripción: Especie de la planta.
Valores ausentes	0

De los cuatro atributos que posee este dataset, nuestro problema se centrará en los tres primeros, por lo que trabajaremos en un espacio de tres dimensiones para nuestros algoritmos.

1.2. Agrupamiento Real No Conocido

Apartado 2

Métricas de evaluación

Teniendo en cuenta la naturaleza del trabajo, necesitaremos definir dos conjuntos de métricas, unas para el problema con datos de agrupación real conocidos y otras para los no conocidos. De esta forma, las métricas quedan divididas en: **extrínsecas** e **intrínsecas**.

2.1. Medidas de evaluación extrínsecas

Las medidas escogidas para evaluar los algoritmos de agrupación de datos reales conocidos han sido tres:

- **Error** : Estudia el número media de casos de errores que se han cometido, representándose por la siguiente fórmula:

$$E = 1 - \frac{1}{n} \max_{\sigma} \sum_{l=1}^{K'} n_{\sigma(l)l}$$

  donde la función sigma asigna a cada clúster original un clúster predicho.

- **Pureza** : Se entiende como la precisión promedio, donde se calcula la precisión media de los clústeres predichos. Para ello, presentemos la siguiente fórmula:

$$Pu = \sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \max_{l \in \{1,\dots,K'\}} P_{kl}$$

  donde P_{kl} representa la precisión, es decir, la proporción de casos sobre un clúster predicho que pertenecen a la mismo clúster que al predicho.

- **F1** : A diferencia de las medidas anteriores esta representa el promedio ponderado de la media armónica de la precisión y el recall, siendo esta última medida equivalente a la precisión pero sobre un clúster real. Su expresión es la siguiente:

$$F1 = \sum_{l=1}^{K'} \frac{n_{\cdot l}}{n} \max_{k \in \{1,\ldots,K\}} \left(\frac{2P_{kl}R_{lk}}{P_{kl} + R_{lk}} \right)$$

2.2. Medidas de evaluación intrínsecas

Para este caso se han escogido otras tres métricas, a diferencia de que estas pueden ser aplicadas para ambos problemas. Son las siguientes:

- **RMSSTD** : Es la raíz del cuadrado de la media de la desviación típica, de ahí las siglas en inglés RMSSTD, y su expresión matemática es:

$$RMSSTD = \sqrt{\frac{\sum_{k=1}^{K} \sum_{x_i \in C_k} x_i - c_{k^2}}{v \cdot \sum_{k=1}^{K} (|C_k| - 1)}}$$

  Estudia la homogeneidad de los clústeres del agrupamiento sin tener en cuenta la distancia interclúster.

- **Ancho de silueta** : Esta métrica supone una diferencia normalizada de la distancia interclúster menos la distancia intraclúster, siendo su fórmula:

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x_i \in C_k} \frac{b_k(x_i) - a_k(x_i)}{\max \{a_k(x_i), b_k(x_i)\}}$$

  donde

$$a_k(x_i) = \frac{1}{|C_k| - 1} \sum_{x_{i'} \in C_k : x_{i'} \neq x_i} d(x_i, x_{i'}) \qquad b_k(x_i) = \min_{h \neq k} \frac{1}{|C_h|} \sum_{x_{i'} \in C_h} d(x_i, x_{i'})$$

- **Índice Calinski-Harabasz** : Evalúa la bondad de un agrupamiento basado en la suma promedio de distancias interclúster e intraclúster al cuadrado, más concretamente de la siguiente forma:

$$iCH = \frac{(n - K) \sum_{k=1}^{K} |C_k| \cdot d(c_k, c)^2}{(K - 1) \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, c_k)^2}$$

Apartado 3

Algoritmos de clustering

En el ámbito del clustering, existen infinidad de técnicas y algoritmos de agrupamiento, pudiéndose hacer una amplia clasificación de los mismos en cinco grupos:

- Algoritmos de agrupamiento basados en particiones
- Agrupamiento jerárquico
- Agrupamiento espectral
- Agrupamiento basado en densidad
- Agrupamiento basado en modelos probabilísticos

En este trabajo se aplican cinco algoritmos de agrupación en total, perteneciendo cada uno a un grupo de los antes mencionados.

3.1. KNN

3.2. Aproximación aglomerativa

3.3. Espectral

3.4. DBSCAN

3.5. Modelos de mixturas

Apartado 4

Pruebas realizadas