Trabajo Práctico 3: Clustering

Introducción

Pondremos en práctica los métodos de agrupado o clustering que aprendimos, utilizándolos para buscar o recuperar clases en algunos datasets. Además se implementarán algunos métodos para estimar el número de clusters y se pondrán a prueba. Se espera que el alumno intente dar una explicación, sin demostrar, del por qué de los resultados obtenidos.

Se pide entregar un notebook con código funcionando y con comentarios y análisis correspondientes a cada resultado obtenido.

Ejercicio 1

a) Analice el dataset “Crabs” que se provee usando los métodos de clustering k-means y clustering jerárquico.
El dataset posee dos columnas (0 y 1) con especie y género de los cangrejos, las cuales son las clases a predecir, y luego tiene 5 columnas (2 a 6) con diversas mediciones realizadas a los cangrejos.

   El objetivo es utilizar los métodos de forma de encontrar las clases de las columnas 0 y 1 con los métodos de clustering mencionados.
   Se sugiere utilizar una transformación logarítmica de los datos en primer lugar, y a partir de allí usar los datos con distintos escalados (e.g. usando *sklearn.preprocessing.scale* o sklearn.decomposition.PCA escalando los datos previamente, o usar PCA y luego escalar los datos, una vez girados).

b) Analice también el dataset “Lampone” que se provee, de la misma forma.
El mismo posee una clase en la primer columna que refiere al año en que fue tomada la medición, y una clase en la columna 142, que refiere a la especie de arándano que le corresponde.

   Nuevamente hay que ver si se pueden recuperar dichas clases con métodos de clustering divisivo o jerárquico, utilizando distintas escalas.
   Para visualizar los datos se recomienda usar PCA ya que poseen muchas dimensiones.

Se proveen algunos ejemplos de carga de los dataset y dos métodos distintos para comparar los resultados de dos soluciones de clustering (o una contra las clases originales): mediante tablas simples y mediante matching óptimo.

Trabajo Práctico 3: Clustering

Ejercicio 2

Prepare código en Python3 para los siguientes métodos:

a) Gap Statistic
b) Estabilidad

Se provee código de ejemplo sobre cómo calcular el score de estabilidad de dos soluciones diferentes de clustering.

Ejercicio 3

Aplique los métodos mencionados en el ejercicio anterior a los problemas de las 4 Gausianas, el dataset “Iris” y el dataset “Lampone”. Comente los resultados obtenidos.

Ejercicio 4 (Opcional, 2 puntos)

Busque un dataset que considere interesante, aplique alguno de los métodos de clustering discutidos y alguno de los métodos que determina la cantidad de clusters presentes. Analice y explique los resultados obtenidos.