Tarea 4 EL4106 - Semestre Otoño 2019

Profesor: Javier Ruiz del Solar
Auxiliar: Patricio Loncomilla
Ayudantes: Gabriel Azócar, Nicolás Cruz, Francisco Leiva, Giovanni Pais

Fecha enunciado: 29 de mayo de 2019.
Plazo entrega tarea: 12 de junio de 2019.

El objetivo de esta tarea es utilizar distintos algoritmos de clustering y analizar su desempeño. Se utilizará la base de datos Anuran Calls MFCCs, que es una base de datos tomadas del UC Irvine Machine Learning Repository. Contiene llamadas de 10 especies distintas de rana, las cuales son representadas mediante 22 carácterísticas. Dichas características son coeficientes cepstrales en frecuencias de mel. La escala mel es una transformación de frecuencias (en Hertz) a un espacio nuevo (en Mel) en la cual las alturas de los sonidos están perceptualmente equiespaciadas. La base de datos se puede bajar desde la siguiente dirección:

https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29

Para poder realizar el trabajo, se usará el siguiente código base (basado en scikit-learn):

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

Los datos de la DB tienen el siguiente formato:

MFCCs_1	MFCCs_2	…	MFCCs_22	Family	Genus	Species	RecordID
0.188	-0.075	…	0.014	Texto	Texto	Texto	1

Se pide utilizar tres algoritmos de clustering en este trabajo: k-means, DBSCAN y clustering aglomerativo.

1) Marco teórico:
   a. Explique brevemente en qué consisten los coeficientes cepstrales en frecuencias de Mel
   b. Explique en qué consiste el algoritmo k-means, y la inicialización kmeans++
   c. Explique en qué consiste el algoritmo DBSCAN
   d. Explique en qué consiste el algoritmo clustering aglomerativo
   e. Explique brevemente la métrica completeness
   f. Explique brevemente la métrica homogeneity
   g. Explique brevemente la métrica v-measure
   h. Explique la métrica silhouette

2) Implemente un código que lea la base de datos. Se recomienda usar la biblioteca Pandas. Los datos se deben dividir en características (MFCCs_1 – MFCCs_22) y labels (Species). Los labels deben transformarse a números en el rango 0-9.

3) Modifique la función bench_k_means( ) agregando una columna extra, correspondiente al número de clusters.

4) A partir de la función bench_k_means( ), implemente una función bench_DBSCAN( ). Notar que:
   a. DBSCAN no posee la variable inertia_, esto debe ser manejado por el alumno.
   b. DBSCAN no asigna todas las muestras a un cluster, las muestras no asociadas tienen label -1. Debido a esto, la función bench_DBSCAN( ) debe entregar la opción de: (i) calcular las métricas sólo con los datos asociados, o (ii) calcular las métricas usando todos los datos, asumiendo que los datos no asociados conforman un cluster extra. Dicha opción debe ser implementada usando una variable extra (True/False) que se le entrega a la función.

c. El algoritmo DBSCAN no genera una cantidad predeterminada de clusters.
d. Dados algunos parámetros, es posible que DBSCAN no encuentre clusters, o bien encuentre uno solo, lo cual imposibilita calcular las métricas. Decida una estrategia para manejar estos casos.

5) A partir de la función bench_DBSCAN( ), implemente una función bench_agglomerative_clustering( ).

6) Hacer pruebas con distintas variantes de algoritmos:
   a. K-Means (con inicialización al azar)
   b. K-means++
   c. DBSCAN con épsilon por defecto
   d. DBSCAN con épsilon 0.7
   e. DBSCAN con épsilon 0.2
   f. DBSCAN con épsilon por defecto, agregando outliers a cluster extra
   g. DBSCAN con épsilon 0.7, agregando outliers a cluster extra
   h. DBSCAN con épsilon 0.2, agregando outliers a cluster extra
   i. Clustering aglomerativo
   j. Repetir todas las pruebas anteriores, después de usar PCA sobre los datos para reducirlos a 2 dimensiones.

7) Analice los resultados obtenidos:
   a. Compare los algoritmos aplicados en base a las métricas indicadas en el marco teórico (*completeness*, *homogeneity*, *v-measure*, *silhouette*), indicando: (i) cuál variante es mejor según cada métrica y (ii) cuál variante es peor según cada métrica.
   b. Indique en cuáles casos no fue posible calcular las métricas de DBSCAN.
   c. Analice el número de clusters obtenidos por DBSCAN y su efecto sobre las métricas.
   d. Analice el efecto de considerar o no los outliers de DBSCAN como un cluster extra y su efecto sobre las métricas.

Los informes y códigos deben ser subidos a u-cursos a más tardar el día 12 de junio a las 23:59. Incluir un corto archivo de texto explicando cómo se utiliza su programa. Las tareas atrasadas serán penalizadas con un punto base por cada día de atraso.

Importante: La evaluación de la tarea considerará el correcto funcionamiento del programa, la inclusión de los resultados de los pasos pedidos en el informe, la calidad de los experimentos realizados y de su análisis, la inclusión de las partes importantes del código en el informe, así como la forma, prolijidad y calidad de este. Se subirá a u-cursos una pauta indicando la estructura del informe y los desgloses de los puntajes.

Referencias

1. https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29
2. https://scikit-learn.org/stable/modules/clustering.html