Laboratorio de Datos

Primer Cuatrimestre 2024

Práctica N° 7: Clustering y clasificación.

Clustering

1. Dada la siguiente tabla de datos

x	-1	0	1	8/5	2	3	4
y	2	1	2	2	1	2	1

Utilizar "a mano" el método de k-medias para agrupar los datos en 2 clusters.

(a) Comenzando con $b_1 = (1, 2)$ y $b_2 = (3, 2)$

(b) Comenzando con $b_1 = (0, 1)$ y $b_2 = (3, 2)$

¿Se obtiene la misma clasificación? ¿Alguna de las clasificaciones obtenidas le parece más apropiada?

2. Considere los datasetets `p7-data1.csv` y `p7-data2.csv` de datos artificialmente generados.

(a) Abra cada dataset en Python y genere un diagrama de dispersión (scatter plot) para cada uno.

(b) Analizando los gráficos "a mano" considere cuántos clusters están presentes.

(c) Pruebe ejecutar el comando `KMeans` con la cantidad de clusters que detectó. Analizar el comportamiento del procedimiento en cada caso.

3. Considerar el dataset `p7-iris.txt` (para leer el archivo, observar que los datos están separados por tabulaciones). En este ejercicio trataremos de identificar las distintas subespecies.

(a) Cargue el archivo `p7-iris.txt`.

(b) Grafique en un diagrama de dispersión la longitud del pétalo vs el ancho del pétalo.

(c) Efectúe un *clustering* k-medias con el comando `KMeans` de los datos basados en las cuatro columnas de datos, considere $k = 3$ clusters.

(d) Repita el inciso b) coloreando en función del índice de cluster obtenido.

(e) Evalúe el error de clustering en función de la siguiente fórmula (within-cluster sum of squares, WCSS):

$$WCSS = \sum_{i=}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde C_i representa el cluster i-ésimo y μ_i es el centroide de dicho cluster, definido como

$$\mu_i = \frac{1}{\#C_i} \sum_{x \in C_i} x.$$

Python ofrece una forma de calcular esto de forma directa.
(Mirar el archivo `p7-ejercicioPetalos.ipynb`.)

(f) Repita el ensayo para distintos valores de k, entre 1 y 10, graficando el $WCSS$ para cada valor de k. Analizar el mejor valor de k posible teniendo en cuenta un compromiso entre "complejidad" (es decir, cantidad de clusters) y nivel de error (es decir, el WCSS).

4. Consideremos el dataset de datos artificiales `p7-dataSinEscalar.csv`.

(a) Cargar los datos y grafícarlos.

(b) A priori y mirando el gráfico, determine la cantidad de clusters que puede detectar en los mismos e imagine inicialmente cómo debieran ser esos clusters.

(c) Realizar un clustering k-medias con el valor de k antes determinado.

(d) ¿Considera satisfactorio el clustering obtenido? ¿Representa lo que usted esperaba?

(e) Uno de los problemas que tenemos es que el método de k-medias es muy sensible a las diferencias de escala entre las dimensiones. Una forma de corregir eso es re-escalando las variables de forma tal que todas se muevan en el mismo rango. Por ejemplo, podemos conseguir eso efectuando una normalización como sigue:

$$X_{ij} = \frac{X_{ij} - min(X_{.j})}{max(X_{.j}) - min(X_{.j})}.$$

De esta manera, logramos que los datos de cada columna caigan entre 0 y 1. Normalice los datos siguiendo este criterio.

(Mirar en Python el comando `MinMaxScaler`)

(f) Vuelva a correr el procedimiento de clustering, tome las etiquetas de clustering obtenidos y grafique los datos originales con un color que dependa del clustering obtenido con los datos escalados.

5. (opcional) Implementar el algoritmo DBSCAN para analizar los sets de datos anteriores. Comparar los resultados con los obtenidos usando k-medias.

Clasificación

6. Implementar un clasificador de k-NN que prediga el sexo de los pingüinos utilizando como variables a el largo del pico y el largo de la aleta. Hacerlo para diferentes valores de k (impares) y evaluar el error de predicción en cada caso. ¿Cómo elegiría el valor de k óptimo?

7. Implementar un clasificador de k-NN que prediga la especie de los pingüinos. Pueden elegir las variables. ¿Con cuáles variables obtienen mejores resultados?

Reportar los resultados con visualizaciones adecuadas.