Matemática 2 (B)
Segundo Cuatrimestre - 2019

Práctica 6: Análisis de datos

1. Implementar un programa que reciba como input un archivo de texto *archivo.txt* y devuelva los coeficientes de la regresión lineal de las componentes principales de *archivo.txt*.

2. Implementar un programa que reciba como input un archivo de texto *archivo.txt* y un número p_acum y devuelva la mínima cantidad de componentes principales que deben considerarse para que el porcentaje de varianza acumulada sea mayor o igual que p_acum.

3. Considerar el dataset *iris.txt*, que representa información del largo y ancho del pétalo y del sépalo de diversas muestras de flores de la especie Iris, la cual se puede distinguir en varias subespecies. Aplicar el programa del ejercicio anterior para determinar la menor cantidad de componentes principales necesarias para alcanzar un 90% de variabilidad. Graficar los datos transformados y la recta de regresión lineal que se obtiene luego de reducir variables.

4. Considere los datasetets *data1.csv*, *data2.csv* y *data3.csv*, de datos artificialmente generados.

   (a) Abra cada dataset en Python y genere un diagrama de dispersión 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.

5. Considerar el dataset *iris.txt*. En este ejercicio trataremos de identificar las distintas subespecies.

   (a) Cargue el archivo *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

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

(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).

6. Consideremos el dataset de datos artificiales *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.

(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.

7. Implementar el algoritmo DBSCAN para analizar los sets de datos anteriores. Comparar los resultados con los obtenidos usando K-medias

8. Considerando los datos de la tabla

(a) Cargar los siguientes datos en Python y graficarlos mediante un diagrama de dispersión.

índice	1	2	3	4	5	6	7	8	9	10
x	5	10	15	24	30	85	71	60	70	80
y	3	15	12	10	30	70	80	78	55	91

(b) "A mano" e informalmente, identificar agrupaciones jerárquicas en el gráfico.

(c) Mediante un clustering jerárquico, basado en la distancia euclídea y usando como criterio de agregación el esquema "single", graficar un dendrograma de los datos.

(d) Marcar los puntos que tomarían en caso de querer considerar dos clusters.

(e) ídem inciso anterior pero ahora con 4 clusters.

9. Utilice clustering jerárquico para poder obtener un clustering del dataset **data3.csv** con dos clusters, compararlo con aquel obtenido usando k-medias.