Implementado Machine Learning

Práctica 4 – Clustering

Material de lectura:

- Sebastian Raschka, Vahid Mirjalili. Python Machine Learning. Capítulo 11.
- Stephen Marsland. Machine Learning: An Algorithmic Perspective. Capítulos 7 y 14.
- Andreas C. Müller, Sarah Guido. Introduction to Machine Learning with Python. Capítulo 3.
- What Is Clustering and Common Clustering Algorithms? https://medium.com/swlh/what-is-clustering-and-common-clustering-algorithms-94d2b289df06

Para todos los ejercicios:

- Para K-means: Utilice las curvas Elbow y Silhouette para interpretar qué ocurre con distintos valores de k. Coeficiente Silhouette: cuanto más alto mejor.
- Para GMM: Utilice la curva Elbow (con GMM=1) y curva BIC para interpretar cómo se adaptan diferentes números de componentes. Coeficiente BIC: cuanto más bajo mejor.
- Para DBSCAN: compare diferentes parámetros de eps y min_samples analizando cantidad de clusters generados y cantidad de outliers.

1. **K-means 2D**. Analice los dataset calabazas.csv, calabazas2.csv, simple_blobs.csv y simple_blobs_2.csv. Trate de determinar el valor óptimo de k. Grafique con un scatter plot los puntos y los centros. El archivo k-means_2D.py posee un código ejemplo.

2. **GMM – 2D**. Entrene un modelo de mixturas gaussianas para los dataset calabazas.csv, calabazas2.csv, simple_blobs.csv y simple_blobs_2.csv. Trate de determinar el valor óptimo de k. Grafique con un scatter plot los puntos y los centros. El archivo GMM_2D.py posee un código ejemplo.

3. **DBSCAN – 2D**. Analice con DBSCAN los datasets 2D_circulos.csv y 2D_moons.csv graficando los resultados. El archivo DBSCAN_2D.py posee un código ejemplo.

4. **Clustering – Países**. Realice un agrupamiento para el dataset who_life_expectancy.csv. Quite los valores nulos y las columnas Country, Year y Status.
    a. Analice diferentes valores de K.
    b. ¿Qué ocurre si no normalizamos los datos?
    c. Intente interpretar los centros en base a las características originales.
    d. Compare los centros obtenidos con K-means y con GMM.

5. **Clustering – Zoo**. Realice un agrupamiento para el dataset zoo.xls. (utilizar pd.read_excel)
    a. Quite las columnas animal y Clase.
    b. ¿Como puede interpretar los centroides de variables que eran binarias?
    c. Ver como se corresponden los clusters encontrados con la columna Clase.

6. **K-means – Cuantización**. El archivo k-means_colorQuant.py posee un código que realiza cuantización de color sobre imágenes RGB.
    a. Analice los resultados con diferentes imágenes y diferentes valores de k.
    b. ¿Puede calcular aproximadamente la superficie de los lagos en la imagen “huechulafquen.jpg”, suponiendo que cada pixel representa 31x31 metros? Para esto, utilice solo la etiqueta del color más oscuro (que representa el color del agua en los lagos).

7. **GMM – Anomalías**. El archivo GMM_anomalias.py posee un código que detecta datos anómalos (con baja probabilidad) para el dataset motores.csv.
    a. Hallar los datos anómalos realizando el mismo procedimiento para el dataset who_life_expectancy.csv.
    b. Compare los datos anómalos encontrados por GMM contra los encontrados por DBSCAN.