DIAGRAMA DE VORONÓI Y CLUSTERING

1. Introducción

Clustering y diagramas de Voronoi son dos herramientas fundamentales en el análisis y visualización de datos. Clustering es una técnica que agrupa conjuntos de objetos o puntos de datos en subconjuntos o grupos homogéneos, mientras que los diagramas de Voronoi son una herramienta matemática que se utiliza para dividir un espacio en regiones basadas en la ubicación de puntos específicos en ese espacio. Estas herramientas son utilizadas en conjunto para visualizar y analizar conjuntos de datos, dividiendo el espacio en regiones y agrupando los puntos de datos dentro de esas regiones en grupos homogéneos para identificar patrones y segmentar los datos en grupos que puedan ser analizados de manera más efectiva.

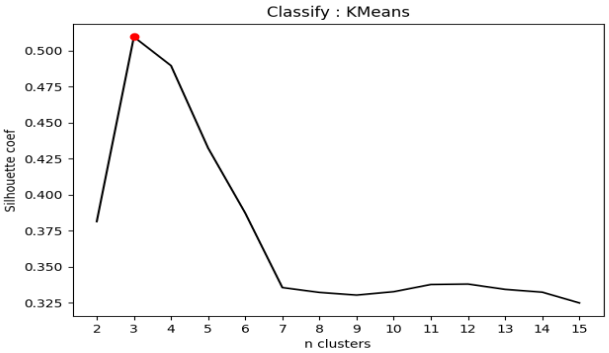
2. Material usado

Como lenguaje de programación, se ha usado python, para realizar todo el código, predicciones y gráficas. Por otro lado como funete de datos se han utilizado los archivos de texto *Grados_en_la_facultad_matematicas.txt* y *Personas_en_la_facultad_matematicas.txt*. En ellos podemos encontrar datos recopilados de alumnos de distintos grados de la facultad de matemáticas, estos son el nivel de estrés y la afición al rock. Según estos datos, sin previamente saber a qué grado pertenece cada alumno, debemos encontrar el número óptimo de grados en los que clasificar dicho grupo de personas. Para ello procederemos a hacer el estudio con los algoritmo K Means y DBSCAN.

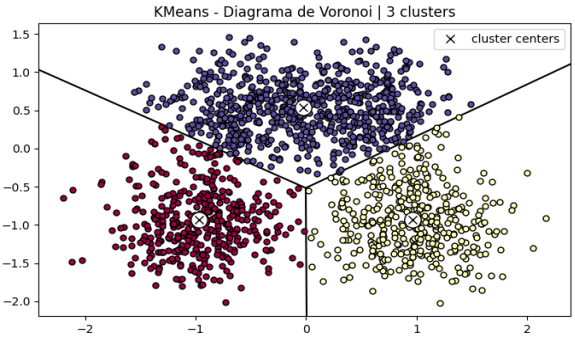
3. Resultados y conclusiones

3.1. Pregunta 1

Podemos observar en la figura (1.a) que para valores de $k \in \{2, 3, \ldots, 15\}$, hay un máximo en $k = 3$ y a continuación según aumenta k disminuye el valor de Silouette, $\bar{s}$. Por otro lado en la figura (1.b) se puede observar una visualización del conjunto de puntos dividido en tres secciones, según el algoritmo de K-Means.



(a) Valor de Silouette según el número de clusters



(b) Conjunto de puntos separado en clusters - K Means

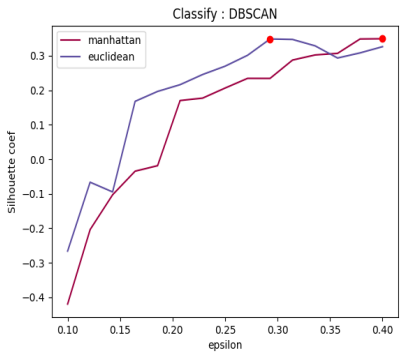
3.2. Pregunta 2

Ahora obtenemos los valores de Silouette, $\bar{s}$, con el algoritmo de DBSCAN. Para ello comparamos con diversos valores de epsilon, $\varepsilon \in [0'1, 0'4]$ y dos métricas distintas, la euclediana y manhatan. Tras evaluar todos los casos observamos como el máximo se obtiene con $\varepsilon = 0{,}4$ y la métrica manhatan. Estos resultados se pueden observar en la figura (2.a).

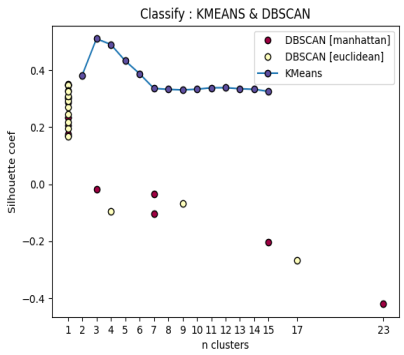
Por otro lado, podemos comparar los resultados del apartado anterior con los actuales en la figura (2.b). Como bien se puede apreciar, los valores de Silouette son bastante mejores en el caso del algoritmo de K-Means. Esto se puede deber a diversos motivos, como por ejemplo el número n_0 escogido para el DBSCAN. Se ha escogido $n_0 = 10$ pero teniendo en cuenta la gran cantidad de puntos a lo mejor no era muy apropiado, además también se observa en la figura (2.a) como

se alcanza el máximo en el extremo del intervalo en vez de en un máximo local, por lo que el intervalo en el que se ha buscado el épsilon también podría mejorar.

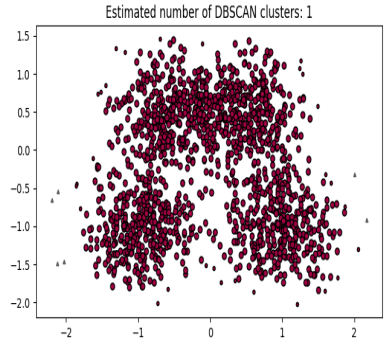
Por último en la figura (2.c) se puede observar la nube de puntos separados en $k = 1$ clusters (es decir, todos pertenecen al mismo grupo sin crear ninguna separación). Donde $k = 1$ es el valor óptimo encontrado por el algoritmo DBSCAN.



(a) Gráfica del valor de épsilon frente al valor de Silouette

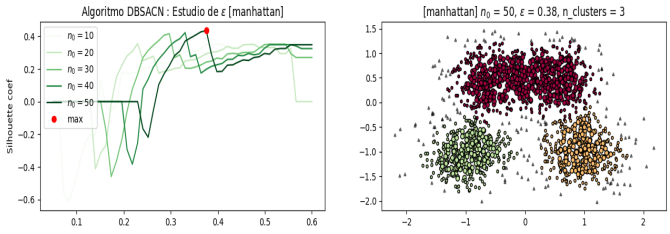


(b) Comparación entre DBSCAN y K-Means

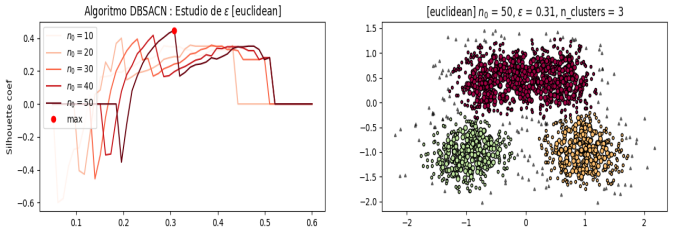


(c) Conjunto de puntos separado en clusters - DBSCAN

A raíz de los resultados del DBSCAN comentados anteriormente, he realizado una ampliación en el estudio, aumentando los intervalos de busqueda de los parámetros óptimos: $\varepsilon \in [0'1, 0'6]$ y $n_0 \in \{10, 20, 30, 40, 50\}$. Como se puede observar, tanto para la métrica eucledianа como para la manhatan obtenemos como valor óptimo el mismo número de clusters que en el K-Means.



(a) DBSCAN - Estudio extra con manhantan



(b) DBSCAN - Estudio extra con euclidia

Los resultados óptimos se dan para $\varepsilon \in [0'1, 0'4]$, es decir, el intervalo de estudio inicial de ε era bueno; sin embargo, el n_0 óptimo se da en ambos casos para $n_0 = 50$ (este está el extremo del intervalo de estudio por lo podría ser mejorable). Por lo tanto, nuestra hipótesis de que $n_0 = 10$ podría ser muy pequeño parece ser correcta.

3.3. Pregunta 3

Como se puede observar en la figura (4) el $(0, 0)$ diríamos que pertenece al grupo azul, mientras que el $(0, -1)$ a simple vista no es fácil distinguir entre el grupo rojo y blanco. Con la función predict de K-Means observamos como efectivamente el $(0, 0)$ lo clasifica en el grupo azul, mientras que el $(0, -1)$ lo clasifica en el blanco.

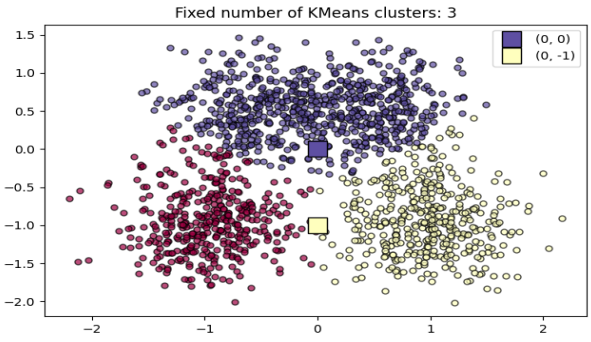


Figura 4: Predicción de 2 nuevos puntos - K Means