Norwegian University of Science and Technology

TDT4300 — **Assignment 3**

CLUSTERING

tdt4300-undass@idi.ntnu.no

Spring 2022

1 k-Means Clustering

1.1 Assignment

This is a programming part of the assignment. Your task is to implement the k-means clustering algorithm and assess the quality of the outputs by calculating Silhouette Coefficient. You are given a Jupyter Notebook[1] (formerly known as the IPython Notebook) file `k_means_clustering.ipynb` in which you have to implement only two functions: `kmeans()` and `silhouette_score()`. Everything else has already been prepared for you. As you have maybe already guessed, the programming language of our choice is Python.

Before you start, you need to install Jupyter Notebook and Python 3. Having that done, open your terminal, navigate to a folder with the `k_means_clustering.ipynb` file, and execute the command `jupyter notebook`. A window of your Internet browser should pop-up with the Jupyter Notebook interface. Open the `k_means_clustering.ipynb` notebook, read it carefully through, and execute it line by line. If this is new to you, get yourself familiar with the Jupyter Notebook and Python.

The assignment is very easy as you do not have to worry about anything else except the core k-means algorithm and Silhouette Coefficient. If you consider yourself a good programmer but without knowledge of Python, you should not have struggles, and you can add a new programming language to your portfolio. If you consider yourself a rather unexperienced programmer and without knowledge of Python, it is a good chance to learn new beginner friendly programming language, and gain more practice in programming. If programming scares you, seek help from other students. Use Piazza to find help if you do not know anyone. We do not have to remind you that plagiarism is not tolerable.

[1] https://jupyter.org/

2 Hierarchical Agglomerative Clustering (HAC)

(a) Explain the Hierarchical Agglomerative Clustering (HAC) and the difference between MIN-link and MAX-link.

(b) You are given a two-dimensional dataset shown in Table 1. Perform HAC (for both MIN-link and MAX-link) and present the results in the form of dendrogram. Use the Euclidean distance. **Describe thoroughly the process and the outcome of each step.**

(c) Verify your results using the KNIME data analytics platform or some Python code. For clarification, MIN-link and MAX-link is in KNIME referred as *SINGLE* and *COMPLETE* linkage methods. We provide you the file *points_hac.csv* containing the very same data. **Present a picture of your workflow and the dendrograms.**

ID	x	y
A	5	7
B	4	3
C	9	8
D	5	6
E	11	3

Table 1: Dataset for HAC.

3 DBSCAN Clustering

You are given following points: $P_0 = (14{,}1)$, $P_1 = (1{,}8)$, $P_2 = (3{,}12)$, $P_3 = (5{,}1)$, $P_4 = (13{,}11)$, $P_5 = (12{,}6)$, $P_6 = (4{,}12)$, $P_7 = (1{,}8)$, $P_8 = (8{,}3)$, $P_9 = (5{,}1)$, $P_{10} = (14{,}12)$, $P_{11} = (12{,}9)$, $P_{12} = (4{,}5)$, $P_{13} = (8{,}4)$, $P_{14} = (2{,}3)$.

(a) Your task is to perform DBSCAN clustering given the parameters $Eps = 4$ (Euclidean metric) and $MinPts = 3$ (including the analyzed point). Identify core, border and noise points. Identify clusters. **Describe thoroughly the process and the outcome of each step.**

(b) Verify your results using the KNIME data analytics platform or some Python code. We provide you the file *points_dbscan.csv* containing the very same data. **Present a picture of your workflow and the scatter plot with marked clusters and outliers.**

Describe thoroughly the process and the outcome of each step.

Submission Requirements

In this assignment we expect you to submit following artifacts:

- A PDF file reporting the results of the sections 2 and 3.
  - Text must not be handwritten.
  - Make sure that the document follows the usual conventions (**names**, assignment/task number, etc.).
- Two files with the programming part in the formats **.ipynb* and **.html*.

All assignment artifacts are to be delivered using ***BlackBoard***. You are allowed to **work in pairs**, however, the identical artifacts must be delivered individually.