Middle East Technical University

Department of Computer Engineering

CENG 414 Introduction Data Mining

Spring 2023 - Assignment 3

Due: 5 June 2023 23:55
Submission: **via ODTUClass**

1 Overview

In this assignment, you are going to implement the K-Means and DBSCAN clustering algorithms and get familiar with Jupyter Notebook, a very famous data science tool. The aim is to make you familiar with certain machine learning algorithms in Python.

2 Tasks

You will use the ”data.csv” file that is provided for you as an attachment in Odtuclass. The dataset has ”x0”, ”x1” and ”y” fields. You need to cluster the instances using the ”x0” and ”x1” fields. Note that ”y” is a categorical variable and it consists of 6 categories. The tasks are explained in the following sections. Note that you can either write your own implementation of K-Means and DBSCAN algorithms or use the scikit-learn implementation of these models in your solutions.

2.1 Task1 - 35 Points

Generate K-Means models for k= 2, 3, 4, 5, 6, 7, 8, 9, and 10. Determine the optimal value of k by looking at the WSS graph and silhouette scores.

2.2 Task2 - 35 Points

Generate DBSCAN models for epsilon values between 0.1 (included) and 2.5(included) incrementing it with steps=0.10 and for min_samples=5, 10, 15 and 20. Find the optimal value of k by adjusted rand scores. (While calculating adjusted rand scores, you will need true labels. [1])

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

2.3 Task3 - 30 Points

Plot the clusters for only the optimal k values based on predicted clusters that you obtain Task1 and Task2. Plot the original data based on the true labels. Did you obtain the same optimal k values in Task1 and Task2? Comment on the quality of the clusters that you obtained in Task1 and Task2. Can we predict the true label of the data by using clustering? Which algorithm gave a better result? What could be the reason one of them outperforms in different cases?

3 Submission

- You are expected to submit a single "ipynb" file named: metu-username_HW3.ipynb (e.g., "e123456_HW3.ipynb").
- You are supposed to be able to interpret your findings. Hence, you shouldn't just find a number or result and just leave it. You need to comment on your findings by giving as much as necessary details.
- You must write your comments as a markdown cells in the IPython file.

4 Tutorials

- Pandas
- Jupyter Notebook
- scikit-learn
- Matplotlib

5 Regulations

- Late submission is not allowed.
- We have zero tolerance policy for cheating. People involved in cheating will be punished according to the university regulations.