Middle East Technical University | Department of Computer Engineering

# CENG 414

## Introduction to Data Mining

Spring 2021-2021

## Homework 3

Due date: 05 06 2022, Sunday, 23:55

# Introduction

In this assignment you will implement the k-means clustering algorithm and get familiar with Jupyter Notebook, a very famous data science tool.

# Tasks

**Task 1.**

In this part of the assignment, you will implement the well-known k-means algorithm yourself in Python. Here are some details:

- Assume that the number of columns in the dataset will always be same to the one given in the homework. However, number of observations might change during testing.
- You will implement 3 different initialization techniques:
  - Manually initialize with an array of shape (n_clusters, n_features).
  - Randomly select from observations.
  - Divide the dataset into k sequential groups. For each group calculate the average of observations, use the result as a centroid. An example is provided in the homework files.
- Finally calculate Dunn index for the clustering result. You have to implement the Dunn index yourself, no libraries allowed.
- More details regarding output format, implementation and submission are given at the end of this file.

**Task 2.**

For this assignment you will use the dataset provided in the homework files. Refer to numpy docs on how to load a .npy file.

- The dataset has some missing values in it. Fill the missing values with the mean value of that attribute. You are not allowed to use any libraries for this task.

- Run your algorithm with k values between [2,10]. Draw a "k versus Dunn index" plot.
- Plot the whole dataset using t-SNE. Color the datapoints according the best clustering you have found in the previous step. You can refer to this link on how to use t-SNE and matplotlib for visualization
- Discuss the results quantitatively, indicating the best k value, and also qualitatively, discussing the reason.

**Task 3.**

- Now, apply "z-score normalization" to dataset and repeat the steps from the previous task. Again, you should implement this part yourself, no libraries allowed.
- Finally compare and discuss the results of pre and post normalization clusterings. Did normalizing improved clustering? Why/why not? Discuss the results quantitatively, indicating the best results and the amount of improvement, and also qualitatively, discussing the reason.

# Submission Details

- Send your k-means implementation in a file named "myKmeans.py". In this file implement a function called "kmeans". You can assume that the input data file will be in the same folder, under the same name with the one provided to you. The function will take 3 arguments.
  - k value.
  - Initialization method. ("manual","random","average")
  - List of initial cluster centers, only to be used when "manual" initialization is selected.
- When run, your program should output the list of designated cluster numbers of each observation (which cluster each observation belongs to), list of the coordinates of cluster center and finally the Dunn index, all in a new line.
- For your further work, create a Jupyter Notebook file called "analyses.ipynb". Create a new cell for your every step. For example, filling the missing values should occupy its own cell, so does the loop you iterate over different k values. This way, I can see your results after every step and grade accordingly.
- Since Jupyter Notebooks support markdown, you can write your discussions inside the notebook file after the relevant cell. After adding your discussions, please export your notebook to a pdf or an html file since the algorithms used are stochastic and they produce different results when re-run. This file will be considered as your report.
- Zip your "myKmeans.py", "analyses.ipynb", and "report.pdf" then submit it through ODTUCLASS.