CENG313 Introduction to Data Science
Fall 2021-2022
Lecturer: Dr. Duygu Sarıkaya
Teaching Assistant: Dr. Begüm Mutlu Bilge
Gazi University, Department of Computer Engineering
Assignment 2 due on 22nd of December 2021, Wednesday 23:59

Assignment 2: K-means Clustering of Iris Flowers

K-means is a clustering algorithm which divides data samples into k distinct clusters. In this assignment you will be working on implementing the K-means clustering algorithm to divide Iris flower instances into K different **clusters based on two features only: Sepal Width and Sepal Length.** For this assignment, we will use all 150 flowers.

Important Note: For this assignment **you will be implementing the K-means clustering algorithm from scratch using Python. You are not allowed to use a K-means method from a library** (such as skicit-learn) but instead you will be implementing this algorithm and defining a method that uses this algorithm yourself. You can use scikit-learn library for other aspects of the assignment (such as loading the dataset etc.). **Your method should work for an arbitrary number of K** (for K=1, K=3, K=5 etc. where $0< K \leq$ total number of training samples). Please check the steps of K-means algorithm on the course slides (Week 6 – Clustering, available on guzem).

Iris Dataset:

The Iris flower data set or Fisher's Iris data set is a multivariate data set. The data is collected to quantify the morphologic variation of Iris flowers of three related species. The dataset consists of **50 samples from each of three species of Iris** (Iris setosa, Iris virginica and Iris versicolor, **where total number of samples is 150**). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. **You can load the Iris dataset using scikit-learn (sklearn.datasets. load_iris)**

For this assignment, you will:

1. Plot (scatter plot) the Iris flower data instances based on the two features of each flower: Sepal Width and Sepal Length. You will have a 2D plot showing where each data instance (=flower) falls on the axes relating to Sepal Width and Sepal Length.
2. Using the **labels** given for each flower in the dataset**, color code** the plot you have at Step 1. You should plot each data instance (=flower) with the color set for its label.
3. In your k-means method, **plot the initial random cluster centers for k=3** on the scatter plot that you have at Step 1.
4. Plot a similar plot to Step 3, this time **with the new cluster centers for each iteration of your k-means algorithm until it converges.**
5. Show the plot you got at Step 2 and the plot you got at the last iteration of your k-means algorithm at Step 4 (after it converges) together **side by side** in the same plot. Please discuss what you see in these plots, and how the real labels are different from your clusters. Write this discussion in comments.

You will submit a jupyter notebook (ipynb file) with executable Python script with comments that explain the code. You can zip your file (or rar) when you submit on guzem as it may not allow ipynb extensions.

You should import all the libraries you will use at the top of your notebook. Please refer to course slides, tutorials and practicals to set up a running Python environment, Jupyter notebook and to import these libraries. You can check the documentation of each library (available online) to get more information about the functions you will use. **Important Note: Please submit your file name with this format: Studentno_StudentName_StudentSurname**

Tips:

You can use Euclidian distance to find the distance between any two data points.

Important Note:

You will receive points only if your script executes, the k-means is written from scratch and works for an arbitrary number of K, if you have covered each point mentioned (plotted all the plots mentioned above), and written comments that explain each main step.

This is an individual assignment, meaning that you will be working on it alone (please check the Class Rules and Expectations below, also available in the syllabus)

Submission:

You will submit a jupyter notebook (ipynb file) with executable Python script and comments (explanations). The file will be uploaded on lms (guzem). You can upload a zip file that contains the jupyter notebook (ipynb file).

Grading:

The total is 100 points. You will receive points only if your script executes, the k-means is written from scratch, works for an arbitrary number of K, and works correctly, if you have covered each point mentioned (plotted all the plots mentioned above), and written comments that explain each main step.

Course Rules and Expectations

All work on programming assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, however, everything that is turned in for each assignment must be your own work. In particular, it is not acceptable to: submit another person's assignment as your own work (in part or in its entirety), get someone else to do all or a part of the work for you, submit a previous work that was done for another course in its entirety (self- plagiarism), submit material found on the web as is etc. **Important Note: Material found online and used as is will lead to your code being similar to many others.** These acts are in violation of academic integrity (plagiarism), and these incidents will not be tolerated. Homeworks, programming assignments, exams and projects are subject to Turnitin and Moss (Measure of Software Similarity) checks. Use sources to learn from only, and write your own code from scratch.