كلية علوم وهندسة الحاسب
College of Computer Science & Engineering

جامعة جدة
University of Jeddah
(059)

CCAI 323 Machine Learning- Tri-Semester 2- 2022/2023

Project

Due date is 20 Feb 2023

Marks S1 Total 2 Obtained ____
Marks S2 Total 2 Obtained ____
Marks C2 Total 6 Obtained ____
Total ____

Introduction

Most of the traditional clustering techniques, such as k-means and hierarchical clustering, can be used to group data without supervision. In this project, we will implement two clustering algorithms from scratch and try it on a customized dataset. Finally, we will check the quality of each clustering using external index to compare between algorithms.

Objective

The objective of this project is to demonstrate:

1- The ability to apply Machine Learning algorithms to analyze the different scenarios.
2- An appreciation of how tools from Probability, Statistics, Calculus and Linear Algebra are combined to build learning algorithms.
3- The ability to translate theory in to practice by using open-source or commercial software tools to build machine learning solutions to practical problems.
4- Applying unsupervised learning techniques to identify clusters.
5- Analyzing the output for a better understanding and decision making.

Project Learning Outcomes (LLO)

By completion of the lab, the students should be able to:

1. **Design** a suitable machine learning solution for a given problem (S1)
2. **Implement** a suitable machine learning algorithm for a given problem using python (S2)
3. **Demonstrate** comprehension of professional, ethical, and social responsibilities associated with artificial intelligence (C2)

Deliverables

Submit a presentation (ppt file) and the code (pdf file) as a team of three students (Max). Each student must upload those two files in his account on the black board.

Project Assessment

1. Each student will be judged separately on their presentation (All students must present). Each student will be asked questions from the code which they will have to answer (All students must know the code and be able to explain any section) 7 Marks

2. Combined marks for the completion of the project and its results as presented by the students (Combined marks i.e. all students of the group will get the same marks) 3 Marks

Project Description

(Option1) with the Full grade of the project 10 out of 10.

The task is to implement from scratch the K-mean algorithm and one of your choice from (Hierarchical clustering, DBSCAN, or GMM). Create datasets as described in Dataset section. Then, show a comparison between the results of the two models.

(Option2) with the partial grade of the project 6 out of 10.

The task is to implement from scratch the K-mean algorithm. Also, apply Hierarchical clustering, DBSCAN, and GMM by using the available machine learning ready packages such as sklearn. Create datasets as described in Dataset section. Then, show a comparison between the results of all models.

Tasks:

1- Implement K-Mean algorithm from scratch and
   a. (option1 as 100% of the project grade) implement your second choice of clustering algorithm (Hierarchical clustering, DBSCAN, or GMM) from scratch.
   b. (option2 as 60% of the project grade) use any ready machine learning package to apply all three rest algorithms (Hierarchical clustering, DBSCAN, and GMM).
2- Generate dataset1.
3- Apply created models on dataset1 and compare the results.
4- Generate dataset2.
5- Apply created models on dataset1 and compare the results.
6- Generate dataset3.
7- Apply created models on dataset2 and compare the results.
8- Generate dataset4.
9- Apply created models on dataset3 and compare the results.
10- Show the accuracy of each model using the following index:
   a. F-measures

b. Normalized mutual information
c. Rand Statistic

11- Summarize the measures values for all algorithms and compare the results (use also visualization).

This project is not yes/no task. Please be creative and show me your best.

Datasets:

You need to generate 3 blobs datasets, but first you (as a group) will specify the following attributes:

1- n_samples = 300.
2- random_state is calculated by the addition of all numbers in each student id. For example

Student1 id	Student2 id	Student3 id
1950234	2040258	2041159
1+9+5+0+2+3+4	2+0+4+0+2+5+8	2+0+4+1+1+5+9
24	21	22
24+21+22=67		
random_state=67		

Then you can generate the three datasets that you will use in your code by the following lines:

Dataset1: blobs dataset

X,y = datasets.make_blobs(n_samples=n_samples,random_state=random_state)

Dataset2: Anisotropicly distributed dataset

X, _ = datasets.make_blobs(n_samples=n_samples,random_state=random_state)

transformation = [[0.6, -0.6], [-0.4, 0.8]]

X = np.dot(X, transformation)

Dataset3: noisy moons dataset

X, y = datasets.make_moons(n_samples=n_samples, noise=0.1,random_state=random_state)

Dataset4: noisy circles dataset

X,y = datasets.make_circles(n_samples=n_samples, factor=.5, noise=.05, ,random_state=random_state)