Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 3
I recommend you explore the recommendation systems!

Homeworks Guidelines and Policies

- It is expected that the students submit an assignment report (HW3_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW3_[student_id].zip). Please combine all your Persian reports just into a single .pdf file by problems order. Code without report has an exact zero point.
  Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- In each homework, 4 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- By default, we assume you implement your codes in Python. If you are using MATLAB or R, you have to use the equivalent functions when it is asked to use specific Python functions.
- Your codes must be separated for each question.
- Make sure you have access to Courses, because that is where all assignments as well as course announcements are posted. Homework submissions are only made through Courses.

- Please submit your work **before the end of June 12nd**.
- There will be a 10% penalty for every late day.
- You are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- If there is any question, please do not hesitate to contact us through the Telegram group chat or following email addresses: **m.ebadpour@aut.ac.ir**, **faramarz.aghajani@aut.ac.ir**, and **danial.alizadeh@aut.ac.ir**.

Sec. 1) Clustering DataStream (50 Pts.)

This section consists of two descriptive and coding parts, for the first part, you must write the required explanations in full, and for the second part, you must state the desired outputs along with the explanations. Avoid bringing the code in the report as much as possible.

1. Answer the following questions.

a) State CVFDT solutions to adapt concept drift (three items).

b) State the main difference between CVFDT and VFDT.

c) State and explain the steps of CVFDT algorithm.

2. There is a lot of data around us; Like profile pictures, tweets, sensor apps, credit card transactions, emails, news, etc., data is everywhere and generated at an incredible speed. Despite these seemingly unlimited streams of data, one of the key challenges is to create lightweight models that are always ready to predict and adapt to changes in the data distribution. The limitations of traditional machine learning methods in these problems have led to the development of online learning methods (also called incremental learning).

a) Explain incremental learning and if necessary, you can explain it by mentioning an example.

b) Study the Adaptive Random Forest classifier and explain how it works and explain how to call it in the scikit-multiflow library in your report.

c) The dataset provided to you was prepared using SEAGenerator in the scikit-multiflow library. Read it as a data stream and incrementally train the model on it.

d) d) Model the dataset with the HoeffdingTree and AdaptiveRandomForestClassifier and finally report the Accuracy and Kappa values using EvaluatePrequential in the scikit-multiflow library.

Sec. 2) Recommendation Systems (50 Pts.)

1. Consider the following utility matrix:

users / ***items***	**U1**	**U2**	**U3**	**U4**	**U5**
A	4	4	4	1	1
B	3	1		4	
C	4	2		2	3
D		2	3		1
E			1	4	3
F	1	1			2

a) Assume that we use the Pearson correlation coefficient as the similarity measure and that we predict a rating by averaging the two most similar neighbours. Which two users do we use to predict the rating of item “D” by user “U4”?

b) What is this predicted rating?

2. Name some of the advantages and disadvantages of the Collaborative Filtering?

3. In this exercise, we want to build a recommender system using the collaborative filtering method. The "rating.csv" file has been provided to you as data, which contains user ID, movie ID and user rating for that movie in each line.

Recommend 6 movies with the most similarity to the user with ID 126 by each of the methods requested in parts b, c, d, and e.

a) Report the ten most similar users to the user with ID 126 (use Pearson correlation coefficient and cosine similarity).

b) user-user collaborative method.

c) item-item collaborative method.

d) Use baseline estimate in item-item collaborative method.

e) Combination of methods b and c.