GE461: **Introduction to Data Science**
Assignment for Data Stream Mining
May 9, 2020 Final Version
ay 7, 2020 Draft Version
Due date: May 20, 2020; 11:59 pm

Notes: This assignment is about data stream mining. Classifying data streams is a challenging problem due to time and memory limitations as well as variation in data distribution. Our aim is to effectively classify the data as they continuously keep entering to the system. In your work you will self learn and use scikit-multiflow[1], which is a data mining framework for the Python programming language.

Teaching Assistant: Sepehr Bakhshi, sepehr.bakhshi@bilkent.edu.tr

TA Zoom Office Hours: Between 1 pm and 3 pm on May 13 Wednesday, May 17 Sunday. Sepehr will send you the necessary zoom meeting information by email a few minutes before the office hour meeting.

TA Office Hour Administration: Please send your question to Sepehr before his office hours and do it as early as possible for efficiency.

A. What to Submit

Your submission has two components.

1. **Code**. It must contain proper comments. You must also include your name at the top as a signature that confirms that you are the programmer. Please remember that MOSS is in our plans for plagiarism check.
2. **Report**. Your report must be in pdf form and must cover all "Work to be Done" sections of the assignment. For each section of your work briefly explain the purpose and what has been done and achieved in that section. Provide a comparison of results that contains tables and plots as appropriate. Make sure that you follow the principles of scientific writing. Use simple past or simple present tense in your report. If you plan to propose future work then in that case you may use future tense.

   Your report must have proper title like a scientific paper, reflecting its true content. It must have your name and address etc. If you like, for experience and fun, you may use the ACM conference paper format[2]. You must use latex or Microsoft Word or their equivalent.

   Optional: As an optional part, at the beginning of your report, you may have a related works section that covers data stream mining briefly with proper references.

   Optional: Another optional part is comparison of the effectiveness of the methods using statistical tests. The design and administration of these tests should be decided by you by looking at the available papers in literature.

   See Justin Zobel's book *Writing for Computer Science* for further hints on the style of CS related scientific paper writing.

B. Submitting Your Work

You will submit your work by uploading it to Moodle in a zipped file. Its name must be streamMiningYourFirstNameYourLastName. For a student with the name "Ali Can Ok" it is streamMiningAliOk.

C. Work to be Done

Your work has six components.

1. **Dataset Generation**
   a. **RBF Dataset**
      Generate a dataset with 10,000 instance using Random RBF Generator and write it into a file called “RBF

[1] https://scikit-multiflow.github.io/
[2] https://www.acm.org/binaries/content/assets/publications/taps/acm_layout_submission_template.pdf

Dataset". Your dataset should have 10 features and 2 class labels, the other options unchanged. It should be something like below data instances.

0.6987, 0.2568, 0.570, 0.949, 0.1970, 0.3285, 0.4474, 0.3355, 0.585, 0.5411, 0
0.0679, 0.0819, 0.6529, 0.9023, 0.314, 0.788, 0.3094, 0.3311, 0.4241, 0.342, 1

b. **RBF Dataset10**
Generate a dataset with 10,000 instance; again 10 features and 2 class labels and the other options unchanged; but this time with "Random RBF Generator Drift" with drift speed of 10 and write it into a file called "RBF Dataset 10".

c. **RBF Dataset70**
Generate a dataset with 10,000 instance; again 10 features and 2 class labels and the other options unchanged; but this time with "Random RBF Generator Drift" with drift speed of 70 and write it into a file called "RBF Dataset 70".

2. Data Stream Classification with Three Separate Online Single Classifiers: HT, NB, MLP

Write a script in Python that constructs and trains the following online classifiers using the three RBF Datasets generated in step 1.

a. HoeffdingTree as **HT** online learner,
b. Naïve Bayes as **NB** online learner,
c. Multilayer Perceptron **MLP** composed of 4 hidden layers of 200 neurons.

3. Data Stream Classification with Two Online Ensemble Classifiers: MV, WMV

Write a script in Python that constructs and trains the following ensemble classifiers that combines HT, NB, and MLP for the three RBF Datasets generated in step 1.

a. Majority voting rule **MV,**
b. Weighted majority voting rule **WMV.**

4. Batch Classification with Three Separate Batch Single Classifiers: HT, NB, MLP

Write a script in Python that constructs and trains the HT, NB, and MLP as batch classifiers using the three RBF Datasets generated in step 1.

5. Batch Classification with Two Batch Ensemble Classifiers: MV, WMV

Write a script in Python that constructs and trains the ensemble classifiers MV and WMV using the three RBF Datasets generated in step 1. For each dataset combine the three batch learners defined in step 4.

6. Report/Paper: Comparison of Models

a. In your comparison of models consider the following and additional questions as needed by considering the results you obtained for three datasets.
b. Compare temporal accuracies of online classifiers, the ones generated in the steps 2, 3 using Interleaved-Test-Then-Train approach using instances of the datasets you generated in the first step.
c. In the comparison of online classifiers try different window sizes and discuss if window sizes are influential in understanding the performance of the methods.
d. Are ensemble methods better than individual models?
e. Compare all online and batch models (single and ensemble) in terms of their overall accuracies. How do they compare with each other in terms of overall accuracy and why?
f. How can you improve the prediction accuracy of the online classifiers (single and ensemble)? Try to find a method and incorporate it to your test cases. Try to show it at least with one classifier.

Is there any difference among the models in terms of their efficiency: provide quantitative data about this?

In your comparisons use plots and tables when appropriate. Number all plots and tables and provide proper subtitles for them. Make sure that you refer to each of them in teh text of your report. Help your reader by providing a simple and easy to follow presentation.