Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3
Topic: Online Data Stream Classification
November 5, 2019
Due date: ~~December 3~~ December 6, 2019

Notes: As you can see information retrieval and data mining subjects are merging with each other. In this assignment you will do independent learning on online data stream classification. It is an important tool in information filtering. You have ample time for the assignment; however, note that it overlaps with other class activities and I may give another assignment that may overlap with this one.

Description: You are required to write a short paper in ACM conference format (about 5 pages) based on your results and research. In the paper introduce the data stream problem, data stream classification, MOA, concept drift concept and other items that you find interesting. If it has some relationship with your thesis provide some reflections from that topic. Perform a literature review using Google Scholar on this problem. Find papers in the journal such as *IEEE TKDE*, *IEEE TNNLS*, *ACM Computing Surveys* and conferences such as ACM CIKM and ACM SIGIR (you are not limited by this list). In the related work part of your paper try to provide a short survey possibly with a table that compares methods on data stream classification. (Yes I know that it is just an assignment.)

Submission: Bring your paper to the class in printed form on the due date and please also email me a pdf copy of it with the subject line containing: CS533: HW3 paper. To your email please also attach the program(s) you wrote for this assignment.

1. MOA Download: Download the package of MOA from below link and unzip it so that you can access the jar library.
http://moa.cms.waikato.ac.nz/downloads/

2. Synthetic Dataset Generation: Open a project in your IDE and add the MOA jar libraries to it. Firstly generate a dataset with 10,000 instance using RandomRBFGenerator library and write it into a file as our dataset called "RBFdataset". Your dataset should have 10 features and 2 class labels. It should be something like the following data instances. You can left other options unchanged.
.
0.698711,0.256831,0.570704,0.94898,0.197067,0.328425,0.447455,0.335523,0.5852,0.541185 ,class1
0.067955,0.08193,0.652916,0.902311,0.311488,0.788677,0.30943,0.331129,0.424125,0.342803 ,class2
.

3. Testing: Write a script in Java that reads RBFdataset and constructs and trains one HoeffdingTree as HT and one NaïveBayes NB online learner using instances of the dataset. You should compare these two learners by their accuracy of prediction. For calculating accuracy use Interleaved Test-Then-Train approach.

Interleaved Test-Then-Train: Each individual instance can be used to test the model before it is used for training, and from this the accuracy can be incrementally updated. When intentionally performed in this order, the model is always being tested on examples it has not seen. This scheme has the advantage that no holdout set is needed for testing, making maximum use ofthe available data. It also ensures a smooth plot of accuracy over time, as each individual example will become increasingly less significant to the overall average.

~~**4. Constructing and Ensemble**: Combine these two learners votes based on **majority voting** rule and calculate the accuracy again for the dataset. How does it change when we combine these two methods.~~ (skip this part).

5. **Effect of Number of Features**: Repeat the above experiment with different number of features. Everything is the same as defined above in step 3 just change the number of features. Do this multiple number of times and try to obtain some observations on the effect of number of features to the prediction performance.

6. **Effect of Number of Classes**: Repeat the initial experiment with different number of classes. Everything is the same as defied above in step 3 just change the number of classes. Do this multiple number of times and try to obtain some observations on the effect of number of class labels to the prediction performance

7. Paper Content: Discuss the results on your paper. Use graphs/tables for comparison. So you compare HT, NB, ensembles under different conditions. Make sure that you show prequential results, the change in performance as you process the data items, in your plots.

Reference: Albert Bifet, Bernhard Pfahringer, Ricard Gavaldà, Geoff Holmes. *Machine Learning for Data Streams: With Practical Examples in MOA.* MIT Press, 2018.