**ECLT5810/SEEM5750 E-Commerce Data Mining**

2024-2025 First Term

Assignment 2

Due: November 15 (Fri) 17:00

**Problem Statement**

This assignment involves using Weka to predict client subscription to term deposits in a bank. The prediction process consists of three phases: 1) data preprocessing, and 2) modeling and model assessment, (3) utilization of the learned model for prediction for new clients. This assignment involves all three phases. In fact, the process is somewhat similar to Assignment 1 with some modifications and enrichment. The details are given below.

**This is an individual assignment, and you must follow the requirements listed below.**

1. Use the "Bank Marketing (with social/economic context)" dataset, specifically the "**bank-additional.csv**" file, to build the model.
2. The target variable is *"y"* which indicates whether the client subscribed to a term deposit or not.
3. Similar to the spirit of Assignment 1, you will conduct data pre-processing. However, you need not follow the precise data preprocessing as in Assignment 1. Instead, you can perform any kind of data pre-processing.
4. Split the whole dataset into two subsets, namely, a training set and a testing set.
5. Save the training and testing sets as two separate files for subsequent processing.
6. Similar to the spirit of Assignment 1, you use the training set to train a model and use the testing set to measure the performance of the model.
7. Repeat the above steps as many times as you like. However, you can only use decision tree model, or neural network model, or logistic regression model. You can try any parameters in these models.
8. After some trials, you should decide and save the most desirable model for delivering to the bank for operation.
9. Similar to Assignment 1, the delivered model will be used for operation and the model will predict the term deposit subscription of new clients. To simulate this operational setting for a bunch of clients, we provide a new client data set, namely, **bank-new-client.csv**. Note that the true class labels in this data set are hidden. Your aim is to make prediction for each of the new clients. Since the data in this new client data set was not available during the learning stage, this new data has not been pre-processed. Therefore, in order to make it suitable for use, you should perform appropriate data pre-processing.
10. Load the pre-processed model obtained in Step 8. Then use it to conduct prediction on the pre-processed file in Step 9. Output the prediction results to a CSV file similar to the format specified in Assignment 1.

You need to write a report describing the process of developing the delivered model, including data pre-processing, selected model and model parameters. You will submit the report in PDF format. You should write clearly the process. In addition, for the delivered model, you should also provide Weka screenshots demonstrating the data pre-processing, the adoption of the data mining model and its parameters, the accuracy of the testing data, and the prediction results of new clients.

**Assessment Criteria**

This assignment will be assessed based on a range of assessment components including quality of new client prediction, rationale and quality of the whole process, quality of report presentation, etc. The component related to the quality of new client prediction will attain a high weighting. For this assessment component, the TA will measure the prediction accuracy and rank the accuracy among students. The higher the rank, the higher of this assessment component score. Note that there will be late penalty for late submission.

**Assignment Submission**

1. There are four files for this assignment to be submitted:
   a) One model object file representing the delivered model;
   b) One CSV file for the prediction results;
   c) One PDF file for the report;
   d) One ARFF file for the **pre-processed** new client data set.

2. Compress these four files into a single ZIP file. Ensure all the file names follow the "id-name" format. For example, if your name is Chan Tai Ming and your student ID is 00123456, the files should be named as follows:
   a) “00123456-CHANTaiMing-Assignment2-model” for the model object file;
   b) “00123456-CHANTaiMing-Assignment2-predict.csv” for the prediction results;
   c) “00123456-CHANTaiMing-Assignment2-report.pdf” for the assignment report;
   d) “00123456-CHANTaiMing-Assignment2-input.arff” for the **pre-processed** new client data set;
   e) “00123456-CHANTaiMing-Assignment2.zip” for the ZIP file.
3. Upload the ZIP file to Blackboard for submission.
4. You may resubmit your assignment before the deadline, with only the latest submission being assessed.
5. For any inquiries, please email clzhou@link.cuhk.edu.hk

**Optional Task**

Optionally, you can assess the quality of new client prediction by yourself, i.e., the prediction quality on **bank-new-client.csv**. Note that this assessment is just for your own interest. Therefore, for this assignment, it is just an optional step. There is no need to submit the result of this assessment. To conduct such assessment, you can upload your CSV file for the prediction results on our course website under the section “Evaluation”. Then you will get a score indicating the accuracy of your prediction of new clients.
Also note that the above online assessment tool will only be available just 1 day before the deadline.