Faculty of Science

Course:	CSCI 4030U Big Data Analytics
Lab:	#7-8
Topic:	Data Mining with Weka
Due Date:	Apr. 2 at 11:59pm

Objective

The objective of this project is to gain hands-on experience in using data mining software to build models from real world data sets and to conduct evaluations of different data mining algorithms.

Software

The software to be used in this assignment is Weka [1]. You can download and install it on your own machine.

[1] http://www.cs.waikato.ac.nz/ml/weka/

You can use "java -Xmx256m -jar weka.jar" to modify the heap size when you invoke weka. You can increase the value of 256m if it is not enough.

Read attached documentation about WEKA

All Weka datasets can be found here:
https://drive.google.com/drive/folders/1UH-cuOnsepo_o1xkM3UyRr01rvvw8FIv?usp=sharing

A guide on how to use Weka's Explorer can be found here:
https://drive.google.com/file/d/1XTgnNRoR23AfuACGZ4iFIJ4LsVzOBgtO/view?usp=sharing

A page documenting the ARFF data format used by Weka can be found here:
https://drive.google.com/file/d/1sROl-8hdB1qBHjyjGj_FmN8hIc8NGLx1/view?usp=sharing

Lab 7

You are given a data set on credit card application approval (credit-a.arff) in the ARFF format used by Weka. The data set describes 690 customers with 16 attributes. The last attribute is the class attribute describing whether the customer's application was approved. The data set contains both symbolic and continuous attributes. Several of the condition attributes contain missing values (which are marked by "?"). All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.

Randomly split the data set into a training set (70%) and a test set (30%), which can be done by using the "percentage split" test option in Weka's "Classify" section. Apply each of the following algorithms to learn a classification model from the training set and classify the examples in the test set.

- C4.5 (weka.classifier.trees.J48)

- Naive Bayes Classifier (weka.classifiers.bayes.NaiveBayes)
- Neural Networks (weka.classifiers.functions.MultilayerPerceptron)

You will be submitting a pdf. Report the classification accuracy of each learning algorithm on the test data set.

Please note that C4.5, naive Bayes and neural networks can handle missing values and continuous attributes automatically.

Note: some algorithms cannot handle these. You need to fill in the missing values and discretize the continuous attributes before using these algorithms. Use the global estimation method to replace missing values and the entropy-based discretization method to discretize all the continuous attributes before using some other algorithms. Both data preprocessing methods are provided by Weka.

Lab 8

Conduct 10-fold cross validation to evaluate the following classification learning algorithms:

- C4.5 (weka.classifiers.trees.J48)
- RIPPER (weka.classifier.rules.JRip)
- Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)
- k-Nearest Neighbor (weka.classifiers.lazy.IBk)
- Neural networks (weka.classifiers.functions.MultilayerPerceptron)

on the following data sets from the UCI repository:

- Ecoli database (database_ecoli.arff and database_ecoli.txt).
- Glass Identification Database (database_glass.arff and database_glass.txt).
- Image segmentation Database (database_image.arff and database_image.txt)

You will be submitting a pdf. Report the misclassification rate and run time of each algorithm on each data set. Discuss the results.

Discuss the results regarding whether there is an overall winner and whether the misclassification rates for the algorithms are significantly different.