Romita Benerjee and Christoph F. Eick

COSC 4335*“Data Mining”* Assignment3 Spring 2018

*Making Sense of Data—Learning and Comparing Classification Models for a Dataset*

First Draft

Individual Project

Due dates: April 12, 2018 at 11p

Last updated: March 22, 8:40a

This course assignment is an opportunity for you to investigate different classification approaches; the idea is to apply different classification techniques to a challenging dataset, to compare the results, to potentially enhance the accuracy of the learnt models via selecting better parameters/preprocessing/using kernels/incorporating background knowledge and to summarize your findings in a report. You must choose one dataset from the below list and compare and interpret the results of using different approaches for the dataset.

List of interesting datasets:

1. Spambase Data Set: <https://archive.ics.uci.edu/ml/datasets/Spambase>
2. Internet Advertisements Data Set: [https://archive.ics.uci.edu/ml/datasets/internet+advertisements](https://archive.ics.uci.edu/ml/datasets/internet%2Badvertisements)
3. Covertype Data Set: <https://archive.ics.uci.edu/ml/datasets/Covertype>

The classification algorithms you must apply are three of the following 4 approaches to obtain classification models for the dataset you chose from the 3 datasets above:

1. Neural Networks
2. Support Vector Machines
3. Random Forest
4. Decision Trees

Other requirements for Assignment3:

* Accuracy of classification algorithms should be measured using 10-fold cross validation.
* Classification models that achieve higher accuracies will get more points.
* In your report after comparing the experimental results, write a paragraph or two trying to explain/speculate why, in your opinion one classification algorithm outperformed the others.
* Include a brief discussion in your report, how you have selected the parameters of particular data mining algorithms.
* In the report also include a brief description of the software you have used in the project.
* Finally, at the end of your report provide a 1-2 paragraphs summary that summarizes the most important findings of Assignment3
* R supports support vector machines, neural networks, and random forests. However, you can use any tool you like for Assignment3; e.g. scikit-learn is another popular tool.

Deliverables:

* Please submit the report using blackboard.
* Name the report as *<last name>\_StudentID\_*P2.docx (or *<last name>*\_*StudentID\_*P2\_.pdf )