Dr. Eick

Final Draft COSC 4335*“Data Mining”* Assignment3 Spring 2016

*Making Sense of Data—Learn Classification/Prediction Models for an Interesting Dataset/Problem*

*Group Project (3 (4) Students per group)*

Due dates: Status Report due: April 4(short) and April 11, 11p; 8-10 page Final Report due: Monday, April 18/19, 11p; project presentations are on Th., April 21, in our class room.

Last updated: April 13, 2014 at 11:30a

This course assignment is an opportunity for you to explore different classification (or prediction) approaches; the idea is to apply different classification techniques (typically, different group members will explore different approaches) to a challenging dataset, to compare the results, to potentially enhance the accuracy of the learnt models via preprocessing/using kernels/incorporating background knowledge and to summarize your findings in a report and to share your project findings with your class mates in a 18-minute presentation on April 21, 2015 (14 minutes presentation and 4 minutes for questions). It is your job to find an interesting dataset and then to apply multiple classification (or prediction) techniques to the dataset and to compare and interpret the results of using different approaches for the dataset. For finding interesting datasets, check out the data mining competitions associated with the KDD conference (called KDD-Cup); moreover, checking out the following links might help to find some interesting datasets:

* <http://select.cs.cmu.edu/class/10701-F09/projects.html>
* <https://docs.google.com/document/d/1Ph-__LSg6I-BftTY3yBk2erh8hp0Kik12fjj9PUOygM/pub>
* <http://www.cs.cmu.edu/~aarti/Class/10601/proj.shtml>
* <http://www.cs.cmu.edu/~epxing/Class/10701/project.html>
* <http://www.cs.cornell.edu/courses/cs6780/2010fa/projects_2010.html>
* <http://www.kddcup2015.com/information.html>

Other requirements for the project:

* Accuracy of classification algorithms should be measured using n-fold cross validation
* In your report after comparing the experimental results, write a paragraph or two trying to explain/speculate why, in your opinion some classification algorithms outperformed other algorithms.
* More details about final report formats will be given by April 11 or earlier.

Please submit the following information in your 1-page draft project description (due April 4, 11p the latest—try submit it as soon as possible!) to both Dr. Eick and Can Cao:

1. Names
2. Project Title
3. Dataset Used
4. Project Idea (use 2-3 paragraphs)
5. Software/Algorithms you will write and/or use are
6. Project Milestones
7. Project Completion Plan

.

Your project description can be sketchy with respect to items 5-7. The progress report, you deliver on April 11, should address items 5-7 in much more details.

*Assignment3 Final Reports are due on Mo., April 18, 11p (5% bonus), but we still accept report until Tu., April 19, 11p. Each report* should have 7-9 pages and should use NIPS format: <http://nips.cc/Conferences/2013/PaperInformation/StyleFiles> The submitted report should follow the following organization:

Abstract (about 1/4 page)

1. Introduction
2. Problem Specification and Methods Used
3. Experimental Evaluation
4. Conclusion
5. References

You can add additional sections to your report, if you feel this is beneficiary! Moreover, feel free to introduce subsections in your report, if helpful; and you have appendices for material that is too space consuming to be included in the report.

Assignment 3 Groups

Group **Yellow**: Avila, Steven Jose, Bolin, Axel Daniel, and Bonilla, Sebastian.

Group **Red:** Gaeth, Lance J, Hampton, Russel Wade, Hornik, Johnathan, Ryan and Flores, Hector A

Group **Blue**: Le, Truc Thi Phuong, Tran, Quan Anh, and Tran, Jenny Nguyen

Group **Green:** Ticer, Michael Craig, Yang Tzu-Wei and van Esch, Mark