Dr. Eick

Fourth Draft COSC 4335*“Data Mining”* Assignment3 Spring 2017

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

*Group Project*

Due dates: Status Report due: April 10, 11p; Final Report is due: Wednesday, April 19 (project scoree will be multiplied by 1.05); however, project reports will still be accepted until Friday, April 21, 11p—project reports received after this deadline will not be graded; project presentations are scheduled for Th., April 20, in 563 PGH, 11:30a-1:20p, *likely in 563 PGH*.

Last updated: April 13, 9:40p

This course assignment is an opportunity for you to investigate different classification approaches[[1]](#footnote-1); 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 10-minute[[2]](#footnote-2) presentation on April 20, 2017 (8 minutes presentation and 2 minutes for questions). It is your job to find an interesting dataset and then to apply multiple classification 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 14 or earlier.

Please submit the following information in your 1-page draft project description (due April 10, 11p the latest—try submit it as soon as possible; e.g. already on April 3!) to both Dr. Eick and Romita:

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

.

*Assignment3 Final Reports are due on April 19 (5% bonus)/21 at 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. Summary and 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 are allowed to have appendices for material that is too space consuming to be included in the report.

Assignment 3 Groups

There will be 9 groups in the order they will give presentations on April 20, 2017:

Group: Yellow: Tang,Daniel H; Tice,Simon Robert Tran,Brian A

Group Grey: Collins,Jacob; Conway,Stephen Michael ; Gruber,Jay David

Group Purple: Pham,Khoi; Samad,Abdul; Troutman,Nicholas D; Vadlamani,Sai Rakesh

Group Green: Capece,Ryan W;, El Awad Ibrahim,Salim Alejandro; Shaheen,Joyce N

Group Turquoise: Gutierrez,Juan; D Hudson,John; L Johnson,Matthew James

Group Red: Lamell,Josef Manfred; Lehmann,Andrew R; Lundgaard,Ethan Eliot

Announcement *COSC 4335 Abalone Data Mining Cup Runner-Up & Winner* (5 minutes)

Group Black: Ansong,Edward; Bremner, Michael; Duka,Taqees S Edmonds,Nicholas Damon

Group Blue: Cooke,Jonathan; Coomes,Cody William; Cupples,Christopher David; Diallo,Nour

Group Orange: Mai,Tony; Nguyen,An Quoc; Perry,Joshua

1. Alternatively, you could choose a dataset and apply sequence mining tools to your chosentdataset. However, if you are interested in performing a sequence mining rather than a classification project, see Dr. Eick during his office hours. [↑](#footnote-ref-1)
2. Groups of 4 give a 10 minute presentation, followed by 2 minutes of questions! [↑](#footnote-ref-2)