Nour Smaoui and Christoph F. Eick

COSC 3337*“Data Science I”* Assignment 2 Fall 2019

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

Third Draft

Group Project

Due date: Th., October 31, 11p (if you submit by Tu., October 29,. 11p you get a 3% bonus)

Last updated: October 4, 2p

Weight: 20-32% of the available Assignment points

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 will also learn how to work in a team and will get some practical experience in comparing and evaluating different classification methods. Datasets to be used in the project include:

1. Image Segmentation Data Set - <https://archive.ics.uci.edu/ml/datasets/Image+Segmentation>
2. Molecular Biology Data Set  
   <https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29>
3. Page Blocks Classification Data Set  
   <https://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification>
4. Vertebral Column Data Set  
   [https://archive.ics.uci.edu/ml/datasets/Vertebral+Column#](https://archive.ics.uci.edu/ml/datasets/Vertebral+Column)

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | 1. Activity Recognition system based on Multisensor data fusion (AReM) Data Set [https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+system+based+on+Multisensor+data+fusion+%28AReM%29#](https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+system+based+on+Multisensor+data+fusion+%28AReM%29) 2. Contraceptive Method Choice Data Set <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>  |  |  | | --- | --- | | |  | | --- | | 1. Activity recognition with healthy older people using a batteryless wearable sensor Data Set <https://archive.ics.uci.edu/ml/datasets/Activity+recognition+with+healthy+older+people+using+a+batteryless+wearable+sensor> | | | |

However, there are restrictions concerning choosing datasets: Each Group can choose one of the following 3 of the above list of datasets[[1]](#footnote-1):

Groups 1-2: Dataset 1 or 2 or 3

Groups 3-4: Dataset 4 or 5 or 6

Groups 5-6: Dataset 1 or 2 or 7

Group 7: Dataset 3 or 5 or 6

Group 8: Dataset 4 or 1 or 7

Groups 9-10: Dataset 2 or 6 or 7

After your group choose one of the 2 dataset assigned to your group, each member of the group will use one of the following approaches:

It is mandatory to use these 2 approaches to obtain classification models for the dataset you chose above:

1. Neural Networks
2. Support Vector Machines

Next select any 2 (1 for groups of 3) from the following 3 approaches to obtain classification models for the dataset you chose above:

1. KNN
2. Random Forest
3. Decision Trees
4. Naïve Bayes

Other requirements for Assignment2:

* Each group will give a 7-8 minute presentation about their project. The Group project presentation have tentatively been scheduled for We., October 30, 2:30-4p!
* 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 your most important findings of Assignment2
* Your report must contain all the results you obtained for the 4 (3 for groups of 3) classification models.
* R supports all the classification techniques mentioned earlier. However, you can use any tool you like for Assignment2; e.g. scikit-learn is another popular tool.

Deliverables:

Create a folder and name it as *G<group number>\_HW2.* HW2 folder should include:

* A *README* file with detailed information on the contribution of each member.
* 4 (3 for groups of 3) directories named with the used technique containing the code specific to every technique. Example: Neural network, SVM …
* The report named as *G<group number>\_P2*.docx (or *G<group number>\_P2*.pdf )
* Include the Slides of your group presentation in the Assignment2 deliverables!

Submit the *G<group number>\_HW2* folder in a zipped file (.zip no .rar , .7z …) through Blackboard.

Remark: Points will be deducted for incomplete submission.

1. To make sure that not every group works on the same dataset with the goal to make the student presentations more interesting as groups will use 3 or more different datasets. [↑](#footnote-ref-1)