Mini-Project

First draft due: November 28, 2017, 11:59 p.m. (5% of 15%)
Second draft due: December 7, 2017, 11:59 p.m. (10% of 15%)
Total marks: 100

Question 1. [100 marks]

Pick any three learning methods and run them on a dataset or problem of your choice. You can pick amongst the algorithms that you implemented from the previous assignments, or even use algorithms from packages. Use your knowledge about model comparison to formally conclude which of the three algorithms is better. This includes proper training-test splits, statistical significance tests, and proper meta-parameter selection techniques (e.g., cross-validation). You can use statistical significance tests built-in to python (or other languages). Provide a precise conclusion of your experiment. Unlike the assignments where you are required to implement the algorithms with optimization packages (other than numpy), you can now use optimization software, such as lbfgs in scipy.

Initial Draft requirements for both Cmput 466 and Cmput 551:

- Should be an almost complete draft
- The only parts that can be missing are the final results, where there can be placeholders
- Must include a description of the data set; a problem specification and question that you are asking; a description of the three approaches that are being tested, and why they are appropriate; and design of experiments, including how data will be split and what statistical significance tests will be used.

Final Mini-Project Writeup for both Cmput 466 and Cmput 551:

- A description of the dataset, indicating the number of samples, number of features and their type (categorical, numerical...), and the target variable.
- A brief description of the importance of the problem (Why the data was collected in first place?)
- A clear description of the parameters that they tunned.
- A detailed description of the methodology followed.
- Presentation of the results in tables/graphs showing confidence intervals/statistical significance analysis
- An analysis stating the "winner" algorithm, and why they think this might be the case.

Comments about marking scheme

Every project is different, so it is complicated to design an evaluation criteria that will fit all the reports. In general, we will be analyzing the following points to determine your grade:

- Selection of a proper metric for measuring performance --i.e. if a classification dataset is highly unbalanced, then accuracy is not a very good metric.

- Report the performance on a separate hold-out set if the dataset is "big enough", or using cross-validation if it is "smaller".
- Proper tuning of parameters of the algorithms selected, and use internal cross validation for this (or divide the data into train/validation/test if the dataset is big enough).
- Check if the differences among learning algorithms are statistically significant.
- Selection of learning algorithms actually make sense.
- That the written report reflects that you did all of this.

Note that the most important part of the project is the methodology chosen for comparing the algorithms and not the performance achieved in a dataset. For example, if you followed an appropriate methodology, but could not get more than 60% accuracy, you can still get full marks.