Assignment 3, Machine Learning (CS7CS4/CS4404); 2017/18

You are a machine-learning engineer working for an IT firm. Your firm wants to develop a new machine learning system. Your task is to carry out some experiments with the datasets provided and with two machine-learning algorithms of your choice, and to write a short report for your boss on the feasibility of developing such a system. The report should be structured as outlined below.

Choose only ONE of the following three tasks.

Task 1: Twitter-User Gender Classification

Your firm wants to develop a system that predicts if a Twitter user is male or female. For your work, use the "Twitter User Gender Classification" dataset provided in the shared Dropbox folder https://www.dropbox.com/sh/euppz607r6gsen2/AACcVFIxekZXYTEM5ZsMSczEa?dl=0.

Task 2: Wine Rating Prediction

Your firm wants to develop a system that predicts ratings for various wines. For your work, use the "Wine Quality Ratings and Chemicals" dataset (use either the red-wine or white-wine dataset) provided in the shared Dropbox folder https://www.dropbox.com/sh/euppz607r6gsen2/AACcVFIxekZXYTEM5ZsMSczEa?dl=0.

Task 3: Movie Recommender System

The system takes a user ID and a movie ID as input and then outputs a prediction of how the user would rate that movie. The training data consists of a set of movies and their ratings for each user, but this is incomplete i.e. for each users ratings are observed for only a small number of movies. For your work, use the 100k MovieLens dataset provided in the shared Dropbox folder https://www.dropbox.com/sh/euppz607r6gsen2/AACcVFIxekZXYTEM5ZsMSczEa?dl=0.

Details for the Report

The word limit for the report is 1,000 words (figures and tables are not included in the word count), plus an appendix that contains the name, individual student IDs, a brief explanation about the individual contributions of the team members and the time spent by each student conducting the work. The report must contain your team id, task number and name.

The structure and marking of the report is as follows.

1. Data & Pre-Processing [10% of the assignment's marks]

Briefly describe the dataset used, including the format of the data, and how the data was processed. Explain how and why you (pre-)processed the data to make it suitable for your analysis.

2. Algorithm & Feature Selection [30% of the assignment's marks]

Describe the machine learning algorithms selected and how you went about selecting appropriate values for the algorithm parameters. Remember to use cross-validation where appropriate. Present plots justifying your choices and discuss your decisions. Given the limited time you have, your boss does not expect a perfectly tuned system. Rather it is the critical discussion here that is important and this should cover the major issues affecting your choices plus the level of uncertainty that your analysis indicates for the parameter choices.

3. Evaluation [30% of the assignment's marks]

Evaluate the predictive performance of the algorithms on test data (which should not be the same as the data used for training). Again, remember to use cross-validation where appropriate to estimate confidence intervals. Your boss doesn't know yet how he would measure the performance of the algorithms. Hence, you have freedom in choosing evaluation metrics. Once again you should critically discuss the results obtained (it is not enough to simply present plots, a reflective discussion of the results is essential and any limitations of your analysis.

4. Conclusion [10% of the assignment's marks]

Based on the results, summarise your conclusions and provide a suggestion as to which algorithms, if any, are worth pursuing.

5. Appendix

Give the names and student IDs of the team members, a brief explanation about the individual contributions of the team members and the time spent by each student conducting the work.

6. Source Code

Also include source code you have developed at the end of the report.

Another 20% of marks are given for the quality of writing and presentation of the report.

Submission

Submit the report as a PDF file by 15th December 2017, 20:00 on Blackboard. Name your report "assignment_3--$task_id--$team_id.pdf" (e.g. assignment_3--task_02—team_73.pdf).