COMP-551: Applied Machine Learning

Mini-project #2: Text Classification

Due on February 16, 6:59pm (predictions), 11:59pm (report and code).

Background:

For this project, you will participate in an in-class Kaggle competition on Text Classification. The goal is to devise a machine learning algorithm to analyze short conversations extracted from the Reddit website, and automatically classify them according to their topics, which include hockey, movies, nba, news, nfl, politics, soccer and worldnews. You will be able to download a training set, including labels, as well as a test set that doesn't include labels.

The competition, including the data, is available here (you must create a Kaggle account using your *@mail.mcgill.ca* account): *https://inclass.kaggle.com/c/comp-551-miniproject-2-reddit-classification*

A description of the dataset is included: *https://inclass.kaggle.com/c/comp-551-miniproject-2-reddit-classification/data*

Performance on the Kaggle Leaderboard will be calculated based on % of instances in the test set that are correctly classified.

The project should be completed in a group of 3. Remember: you must work with different team members on each mini-project. You can use the discussion board on myCourses to find a team.

General instructions:

To participate in the competition, you must submit a list of predicted outputs for the test instances on the Kaggle website (see example file on Kaggle website for the file format). You can submit multiple prediction entries throughout the competition, and track your performance on the Kaggle Leaderboard. The test set is divided into two parts; one set (the public set) is used to update the scoreboard, one set (the private set) is used to calculate the final score of each team.

To solve the problem, you must try the following methods (the 3rd one is optional):

1) At least one baseline linear classification algorithm (from lectures 4-5), such as Naïve Bayes, fully implemented by your team.
2) At least one non-linear classification algorithm (from lectures 7-11), such as decision trees, nearest neighbor, SVM, fully implemented by your team.
3) (Optional) Any other machine learning method of your choice. Existing packages can be used, e.g. *scikit-learn*, if referenced in your report.

Your written report should provide results for all methods considered, including at least 1 from categories 1 & 2 above. For the Kaggle competition, you can submit results from your best performing method, from any of these categories.

Submission requirements (1 submission per team, not per individual):

- You must **submit the code** developed during the project. The code can be in a language of your choice. The code must be well-documented. The code should include a README file containing instructions on how to run the code. Submit the code as an attachment (see below).
- The **prediction file** for this project must be submitted online at the Kaggle website.
- You must **submit a written report** describing your methodology and results. The report should respect the following structure:
    - Project title. (Do not include a cover page.)
    - Name of your team as it appears on Kaggle.
    - List of team members, including their full name, email and student number.
    - Introduction: briefly describe the problem and summarize your approach (1 paragraph).
    - Related work: previous literature related to text classification problem (max. ½ page).
    - Problem representation: data pre-processing methods, feature design/selection methods.
    - Algorithm selection and implementation (for each of the categories above). You do not need to include details that are in the class notes (e.g. SVM derivation, etc.), unless necessary to understand other details. Include any decisions about training/validation split, distribution choice for naïve bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.
    - Testing and validation: detailed analysis of your results, outside of Kaggle.
    - Discussion: pros/cons of your approach & methodology.
    - Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project (e.g. defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.) At the end of the Statement of Contributions, add the following statement: "We hereby state that all the work presented in this report is that of the authors." Make sure this statement is truthful!
    - References (optional). Use appropriate referencing style throughout the report, with the list of references given at the end of the report. References are optional, but should appear if you used any additional data, software, or methods that were not presented in class.
    - Appendix (optional). Here you can include additional results, longer details of the methods, etc.

The main text of the report should not exceed 5 pages. References and appendix can be in excess of the 5 pages. The format should be double-column, 10pt font, min. 1" margins. You are encouraged to use the standard IEEE conference format, e.g. *ewh.ieee.org/soc/dei/ceidp/docs/CEIDPFormat.doc*.

Evaluation criteria:

Marks will be attributed based on: 30% for performance on the private test set in the competition; 60% for the written report. Both these components will be assessed per team, not per individual (including late penalties). The remaining 10% is attributed following participation in the peer-review process. This is assessed individually. The code will not be marked, but may be used to validate other components.

For the competition, the performance grade will be calculated as follows: The top team, according to the score on the private test set, will receive 100%. A Random predictor, entered by the instructor, will score 0%. All other grades will be calculated according to interpolation of the private test set scores between those two extremes.

For the written report, the evaluation criteria include:

- Technical soundness of proposed methodology (feature selection, algorithms, optimization, validation plan)
- Clarity of methodology description, plots, figures (don't forget captions, axes labels, etc.)
- Overall organization and writing (don't forget to spell-check!).

For the peer-review, the instructions and evaluation criteria will be given in class (and included in slides) later; this is not due on February 16.

Submission instructions:

Predictions on the test set must be submitted on Kaggle. Create one team per group:
https://inclass.kaggle.com/c/comp-551-miniproject-2-reddit-classification/submit

The deadline for the predictions is set to Feb.16, 6:59pm Eastern = 11:59pm UTC.

We will continue to use CMT to coordinate submission of project files and peer-reviews:
https://cmt3.research.microsoft.com/COMP551Y2017

The deadline for the report is set to Feb.16, 11:59pm Eastern = 8:59pm Pacific (the default).

You should use the same account as for the previous project on CMT, but submit to a new track, "Project 2". The new report should be submitted as a "New Submission" (one per group), linking other team members as co-authors. The code should be submitted as Supplementary Material, in a single file named *project_code.zip*. Other acceptable file formats for this are *.gz, .tar, .tgz*. Make sure that the code is set up so that we can run it (e.g. include a README file).

If you are submitting the project late (subject to automatic 30% penalty), send all files by email to the course instructor.