Machine Learning & Data Mining
Miniproject 1

Caltech CS/CNS/EE 155
Released January 29^{th}, 2018

---

1 Overview

In miniproject 1, you will use bag-of-words representations of Amazon reviews to predict the sentiments that the reviews express. The bag-of-words representation is constructed from counts of the top 1,000 words appearing in the reviews, excluding a list of such stop words as "and" and "the." These words are also stemmed, so that words such as "want" and "wanted" are collapsed into a single feature. For each review, a label of 0 indicates a 1 or 2-star review, while a label of 1 indicates a 4 or 5-star review. Note that 3-star reviews, i.e. those expressing a neutral sentiment, are not included in this dataset. The file 'training_data.txt' contains 20,000 reviews that you can use to train your model. The first row is a header listing the label heading and 1,000 selected words in the bag-of-words model. Each subsequent row contains a label indicating the sentiment of that review (1-2 stars or 4-5 stars) followed by the count of each word in the given Amazon review. The file 'test_data.txt' contains a header and 10,000 reviews in the same format, but excluding the label.

The miniproject involves a competition on Kaggle, a site for data science competitions. The link to the competition, which includes the datasets and description of the data, can be found here:

https://kaggle.com/c/caltech-cs-155-2018

In order to join the competition, visit the link below:

https://www.kaggle.com/t/5a0dce599fdb4750ad6b64e9bd2925b2

In the competition, you are to use the training data (training_data.txt) to come up with predictions for the test data (test_data.txt). There will be a public leaderboard that will show your performance, but it only consists of half of the test set. The private leaderboard with results from the other half of the test set will be revealed in class on the due date. The competition will end on Thursday, February 8th at 1:59 PM PST.

You will submit prediction files on the competition site, in the format of 'sample_submission.txt', provided on the competition website. Each row in your prediction file should have an ID and a prediction. The first prediction should have an ID of 1, with the ID incrementing in each subsequent row, and IDs following the same order as the samples in 'test_data.txt.' Each prediction will be a label, either 0 or 1.

2 Key Notes

- The competition ends on Thursday, February 8th at 1:59 PM PST.
- The report is due the following Monday, February 12th at 9 PM PST, via Moodle. See below for the report guidelines. The report should explain your process and results in a thorough manner.
- You can work in groups of up to three, but must make submissions from a single account.
- You may collaborate fully within your miniproject team, but no collaboration is allowed between teams.
- You can make up to 5 submissions per day; however, at the end, you need to select the 2 submissions that you think will perform the best on the private test set for the competition.
- If you have questions, please ask on Piazza! As with any Kaggle competition, it's best to get started early since you are only allowed to make 5 submissions a day.

- You can use any open-source tools and Python, using both concepts you learned in class as well as any other techniques you find online, to get the best score that you can.
- **You may not search for additional data related to this task; you may only train your models using the provided training set.**

3 Report Guidelines

- **Due date:** Monday, February 12th at 9 PM PST
- **Format:** The report should be 4-8 pages long in single column format. Only include code if necessary—your code should not be a significant portion of the report. We recommend a link to a GitHub repo. You should use graphs in your report, as visualization is very helpful!

We highly recommend that you use the LaTeX file provided to you and simply fill in the blanks. See our example file for guidelines. The structure is as follows:

1. **Introduction:** This section is purely for the TAs and should be brief.
   - Group members
   - Team name
   - Division of Labour: Your team must ensure that each member has an equal amount of workload during the competition. If there is a noticeable discrepancy in the division of labour, team members may receive differing grades.
2. **Overview:** This section should be a concise summary of your attempts. More detailed explanations should go in the next section.
   - Models and techniques tried: What models did you try? What techniques did you use along with your models? Did you implement anything out of the ordinary?

     Descriptions should be concise, at most 1-2 sentences. Again, more details can be included in the next section. However, this section is meant to be a more general overview.
   - Work timeline: What did your timeline look like for the competition?
3. **Approach:** This section should be a more detailed explanation of how you approached the competition.
   - Data processing and manipulation: Did you manipulate the data or the features in any way? What techniques and libraries did you use to accomplish such manipulation?
   - Details of models and techniques: Why did you try the models and techniques that you used? What was your process of using them? What are the advantages and disadvantages of using such methods?
4. **Model Selection:** This section should outline how you chose the best models.
   - Scoring: How did you score your models? Which models scored the best?

- Validation and Test: Did you use validation techniques? How did you test your models? What were the results of these tests, and what did the results tell you?

5. **Conclusion:** This section should be used to summarize the report, as well as to include any last-minute details.
   - Discoveries: What did you learn from this competition? Did you learn anything new outside of lecture?
   - Challenges: What could you have done differently? What obstacles did you encounter during the process?
   - Concluding remarks: Anything else you'd like to mention?
6. **Appendix (optional):** Use this section for anything else you'd like to include. Don't include this section if the above sections have covered everything you would like to discuss.

4 Grading metrics

You will be evaluated on both your public and private leaderboard performance; the two leaderboards have equal weighting.

The report is worth the majority of your grade. That is, we care more about the process and thoughts behind your results rather than the scores.