Math 514

Project Proposal

April 9, 2019

Project Proposal

Project proposals are due Wednesday, April 10. Project teams will give a short presentation during class on April 24. A set of slide templates using the Xaringan environment will be provided to help with preparation. Please make an effort to share what you have learned about your topic with the class. The final project is due by midnight May 5. The final deliverable should use reproducible research methods and must run successfully.

1. Identify team members. Include name and email

Our team consists of Mary Peng (xp19@georgetown.edu) and Michael Leibert (mll82@georgetown.edu)

2. Provide a problem statement and a description of what you hope to achieve. If you are starting with a project/model found online please provide a link and describe what you will do to add to the work or experiment with.

The problem we are looking at is how to extract a person's sentiment given something they have wrote. Specifically, we will use fine foods reviews to predict user sentiment toward products reviewed on Amazon. The sentiment is measured on a scale of 1 to 5, with 5 being the highest positive sentiment. Our dataset contains both a short review summary, as well as the full length review.

We start by focusing on predicting user sentiment from the short review summaries, then attempt to scale up and use the full reviews for prediction. The dataset also contains other variables, such as user ID, product ID, and the "helpfulness" of the review (number of users that rated the review helpful and the number of users that rated the review), that we may use in constructing the model.

We are particularly interested in different regularization techniques, and will compare the effectiveness of dropout vs. L1 / L2 regularization. Our data processing phase will incorporate some elementary ideas that are used in NLP. Some preliminary techniques we are investigating are tokenization, how embedding layers work, bag of words, `word2vec`, and other pretrained word embeddings. Some approaches could be dropped and others adopted as we move forward.

3. Give a description of the data set

The dataset comes from Stanford's SNAP database (https://snap.stanford.edu/data/web-FineFoods.html). The dataset contains 568,454 reviews by 256,059 between 1999 and 2012. We will split the dataset into 60% training, 20% hyperparameter tuning, and 20% test. Sample data format is as follows:

product/productId: asin, e.g. amazon.com/dp/B001E4KFG0 review/userId: id of the user, e.g. A3SGXH7AUHU8GW review/profileName: name of the user review/helpfulness: fraction of users who found the review helpful review/score: rating of the product review/time: time of the review (unix time) review/summary: review summary review/text: text of the review