Text Classification

COMP 551: Applied Machine Learning
Kaggle Team: Chris-HP

Christopher Glasz
christopher.glasz@mail.mcgill.ca
260720944

Truong Hoai Phuoc
phuoc.truong2@mail.mcgill.ca
123456789

I. Introduction

Blah blah blah

II. Related work

III. Problem Representation

We chose to represent the text data as vectors of TF-IDF scores. We initially implemented our own code to calculate these feature vectors, but eventually turned to the NLTK library to produce them for us, to speed up the calculation time.

At the outset, we had planned on using the 2,000 most common unigrams and the 200 most common bigrams as features, but this proved to be ineffective. We improved our classification accuracy (at the cost of feature preprocessing time) by increasing to 90,000 unigram and 10,000 bigram TF-IDF scores, for a total of 100K features. Along the way we tried 22K and 55K features (both with a 10/1 unigram-bigram ratio), and found that the performance of our models just continued to increase as we added more features, so we settled on a nice even 100K, as that's just about how many unique terms exist in the training corpus.

Our performance was further improved by lemmatizing the input tokens using NLTK's WordNetLemmatizer. This prevented terms carrying the same information (e.g. "authenticate" and "authenticated") from being treated as two separate features. By tagging words by their part of speech before passing them to the lemmatizer, we were able to further improve our features (by reducing terms like "is", "are", and "am" to their common meaning of "be"). This process greatly increased the computation time, but was well worth it, as we only needed to calculate the feature vectors once.

Another decision we made over the course of the project was from where to extract the features. Early on, we were very strict about not using the text in the test corpus to select features, as we did not want to inadvertently influence the training of the model with information from set-aside data. However, as we increased the number of features, the difference between the set of most common words in the training corpus and in the combined training and testing corpora became negligible. Eventually we decided to produce 4 datasets:

1) *X_trn*: 80% of the training examples, with features drawn only from that corpus
2) *X_val*: The remaining 20% of the training examples, with features drawn only from that corpus
3) *X_all*: All examples in the training corpus, with features drawn from the words in both corpora
4) *X_tst*: All examples in the test corpus, with features drawn from words in both corpora

X_trn was used to find optimal hyper-parameters for our models, and used to train before prediction on the validation set. *X_val* was our validation set, and was only used to calculate our validation accuracy. *X_all* was used for final training before predicting labels on the test set, and was also the set over which we performed cross-validation. *X_tst* was, as one would expect, the dataset used for predicting labels on the test set.

All of these sets are saved as compressed sparse row matrices to save space. They can be read by using the `load_sparse_csr` function in `utils.py`.

IV. Algorithm Selection and Implementation

A. Linear

B. Nonlinear

V. Testing and Validation

A. Linear

B. Nonlinear

VI. Discussion

VII. Statement of Contributions

A. Christopher Glasz

B. Truong Hoai Phuoc

VIII. References

IX. References