Natural Language Processing
Assignment 1

Anik Roy (ar899)

November 6, 2019

1 Introduction

This report considers the problem of the positive or negative sentiment classification of reviews. We reimplement a subset of techniques used in Pang et al. (2002)[1], specifically support vector machine (SVM) and Naive Bayes (NB) classifiers. The data used was a set of movie reviews given in the framework of an NLP course

2 Background

A bag of n-grams representation is used to classify documents. Pang et al. also choose some combinations of parameters to investigate the effect on accuracy of the final system, for example using unigrams or bigrams.

2.1 Naive Bayes

The Naive Bayes classifier assigns documents to a class based on the likelihood of it being in that class $c^* = \arg\max_c P(c|d)$. To get $P(c|d)$, we use Bayes Rule to derive:

$$P_{\text{NB}}(c|d) := \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

, making the naive assumption that features are dependent only on class.

2.2 Support Vector Machines

SVMs are a type of classifier which treat each document as a vector of features. Training consists of finding a hyperplane to separate the two classes, and classifying documents by measuring distance to the hyperplane.

We use Joachim's (1999) SVMlight package[1] for training and testing.

3 Method

The systems described were reimplemented in python. Stemming was performed on words before converting to features, using the porter stemmer implemented in the nltk package[2].

We consider different types of features - unigrams, bigrams, and unigrams + bigrams, as well as both presence and frequency. No feature cutoff was implemented.

We use 10-fold stratified cross-validation, and dividing the data using round robin splitting. With 2000 documents (1000 positive, 1000 negative), and training on 9 folds (90% of the documents) there are 446506 bigrams and 52556 unigrams.

[1] `http://svmlight.joachims.org`

[2] `https://www.nltk.org/_modules/nltk/stem/porter.html`

Features	**frequency or presence?**	**NB**	**SVM**
Unigrams	frequency		
Unigrams	presence		
Bigrams	presence		
Unigrams + Bigrams	presence		

Table 1: Accuracies of systems with different feature types, finding the average over 10 fold cross validation

4 Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

5 Conclusions

[wordcount: 500][3]

[3]calculated using detex | wc

References

[1] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.