Natural Language Processing

Assignment 1

Anik Roy (ar899)

November 7, 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. We also compare

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.

3 Method

The systems described in Pang et al. were reimplemented in Python. We use Joachim's (1999) SVMlight package[1] as our SVM implementation.

We consider several different types of features - unigrams, bigrams as well as both unigrams and bigrams. We also perfom stemming on words before converting to n-grams, and consider the effect of stemming on the performance of the classifiers. The porter stemmer implemented in the nltk package[2] was used. No feature cutoff was implemented, and no stoplists were used. We also do not consider the position of words in sentences, nor do we carry out parts-of-speech tagging.

We use 10-fold stratified cross-validation, and dividing the data using round robin

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

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

Features	**Frequency or presence?**	**NB**	**SVM**
Unigrams	frequency	**81.0**	73.3
Unigrams	presence	82.8	**86.3**
Unigrams + stemming	presence	82.3	**85.6**
Bigrams	presence	**85.5**	83.0
Unigrams + Bigrams	presence	85.6	**87.4**

Table 1: Accuracies of NB and SVM systems with different feature types, averaging over 10 fold cross validation, in percent

System A	**System B**	**P value**
NB, Unigrams	SVM, Unigrams	$\mathit{6.20} \times \mathit{10}^{-4}$
SVM, Unigrams + Frequency	SVM, Unigrams + Presence	$\mathit{5.09} \times \mathit{10}^{-9}$
NB, Unigrams + Frequency	NB, Unigrams + Presence	0.421
NB, Unigrams + Presence	SVM, Unigrams + Presence	0.117
NB, Unigrams + Bigrams	SVM, Unigrams + Bigrams	$\mathit{2.33} \times \mathit{10}^{-6}$
NB, Unigrams + Presence	NB, Bigrams + Presence	*0.0418*

Table 2: Systems compared for statistically significant difference (system B outperforming A), using a two-tailed sign test, with p-values under 0.05 in italics

splitting. With 2000 documents (1000 positive, 1000 negative), and training on 90% of the documents, the system is trained on 446506 bigrams and 52556 unigrams - the first 9 folds.

4 Results

The SVM classifiers perform significantly[3] better when frequency is not taken into account. While we observed an improvement in the accuracy of Naive Bayes when using presence over frequency, it was not significant.

We did not observe an improvement in accuracy when using stemming. For both classifiers, the best results were observed when using both unigrams and bigrams. In particular, SVM was significantly better than NB in this case.

5 Conclusions

[wordcount: 500][4]

[3] When using a paired sign test with $\alpha = 0.05$

[4] 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.