**Opinion Mining Using Support Vector Classifier**

**Navjeet Kaur**

University Of Essex, Colchester, United Kingdom

[nk23254@essex.ac.uk](mailto:nk23254@essex.ac.uk)

Abstract

Opinion Mining, also known as Text Classification, is one of the most fundamental tasks in Natural Language Processing. In this study, I employed a Support Vector Machine, a type of machine learning algorithm, to solve this problem. To train and test the classifier, I utilized the IMDB Reviews dataset, which contains two linguistic features. I have tried to find the best-fit kernel type by comparing the accuracy score and the time taken to train the model

1 Introduction

Opinion Mining can help users decide whether to watch a movie or not, based on reviews. Similarly, it can aid in analysing product reviews on e-commerce sites and assist in deciding which product to purchase. Reviews can also be beneficial in determining which places to visit. Text classification can play a vital role in segregating comments on social media platforms and preventing trolling. The possibilities are endless, and this area is ripe for exploration. Unbiased data also plays a crucial role in the performance of the model. Having an uneven number of examples can lead to wrong predictions.

2 Data Exploration

A graph with a line

Description automatically generated

Figure 1

A graph with a line

Description automatically generated

Figure 2

Figure 1 and Figure 2 display the similarity in the Frequency of the words used in negative and positive comments are almost the same.

3 SVM

Support Vector Machines (SVMs) are used for solving different types of problems such as Regression, Classification and Anomaly Detection. Depending on the problem, SVMs are given different names like SVR and SVC. SVMs work by creating a hyperplane and marginal planes based on the support vectors to separate the features and divide them into categories. A hyperplane can be a linear one-dimensional line, a polynomial curve, or a higher dimensional space depending on what's best for separating the features. Support vector machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing h. Support vector machines find the hypothesis h which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H.SVMs are very universal learners. In their basic form, SVMs learn linear threshold functions. Nevertheless, by a simple "plug-in" of an appropriate kernel function, they can be used to learn polynomial classifiers, radial basic function (RBF) networks, and three-layer sigmoid neural nets.

3 BOW

Before model fitting, the training set underwent pre-processing which entailed the removal of stop words, punctuations, and links. As machine learning algorithms cannot operate on raw text, it was necessary to convert the text into numerical vectors, specifically in the form of a CSR matrix. To achieve this, the Bag Of Words Vectorizer was utilized.

The Bag of Words (BoW) model is a fundamental and highly effective method for representing text in numerical format. We can extract valuable insights by transforming a sentence into a bag of word vectors. For instance, let's consider the scenario where we have three distinct types of movie reviews that we analysed previously.

Review 1: What an absolutely stunning movie

Review 2: This movie was frustrating.

Review 3: This movie is spooky and good

We will first build a vocabulary from all the unique words in the above three reviews. The vocabulary consists of these 12 words: ‘This’, ‘movie’, ‘is’, ‘absolutely’, ‘stunning, ‘and’, ‘good, ‘was’,  ‘frustrating, ‘spooky’,’what’, ’an’.

We can now take each of these words and mark their occurrence in the three movie reviews above with 1s and 0s. This will give us 3 vectors for 3 reviews:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **What** | **Stunning** | **Movie** | **frustrating** | **good** | **this** | **an** | **was** | **and** | **is** | **absolutely** | **spooky** | **Length of the review** |
| **Review 1** | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 5 |
| **Review 2** | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 4 |
| **Review 3** | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 6 |

Vector of Review 1: [1 0 1 0 0 0 1 0 0 0 1 0]

Vector of Review 2: [0 0 1 1 0 1 0 1 0 0 0 0]

Vector of Review 3: [0 0 1 0 1 1 0 0 1 1 0 1]

And that’s the core idea behind a Bag of Words (BoW) model.

4 Experiment

The following experiment compares the performance of SVMs using linear, polynomial and RBF kernels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear | Polynomial with degree 3 | RBF | Sigmoid |
| Accuracy% | 85.088 | 71.272 | 88.016 | 57.76 |
| Time(sec) | 7371.87 | 1767.5 | 1528.33 | 529.97 |

Results show that RBF kernel has the best performance in terms of Accuracy and Time.

5 Conclusion

The experiment proves that SVMs a very promising and easy-to-use method for learning text classifiers from examples.

6 References

﻿﻿﻿C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20:273- 297, November 1995.

﻿﻿﻿Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

﻿﻿﻿﻿Y. Yang. An evaluation of statistical approaches to text categorization. Technical Report CMU-CS-97-127, Carnegie Mellon University, April 1997.

﻿﻿﻿﻿Y. Yang and J. Pedersen. A comparative study on feature selection in text categorisation. In International Conference on Machine Learning (ICML)