Machine Learning Engineer Nanodegree

Capstone Proposal

Ahmed Kaboud
September 12th, 2018

Domain Background

During my Masters study, I read a lot about opinion mining and sentiment analysis. Summarized many papers. But I didn't do any experiment by myself. So, it encourages me to go on this topic and do this experiments.

Opinion mining and sentiment analysis is the computational study of people's opinions, attitudes and emotions toward an entity. It is rapidly growing area. This is very much helpful to both the individuals, who are willing to buy that product and the organizations which want to improve and market their products. There are many types of product. The product can be electronic device, song, movie, picture…etc.

Opinion mining is one of the most active research areas in Natural Language Processing. With the growth of Social Media, there is a need to know people's opinions. So, here Opinion mining play an important role by analyzing and clustering the user generated data like reviews, blogs, comments, articles etc. which can help us to understand people attitudes and know if it is a positive feeling or negative feeling.

Problem Statement

I will use sentiment analysis in analyzing movies reviews. There are many movies in the internet or cinemas. A lot of people can't decide which movie to watch. The trivial solution to decide, is to search manually for any website which provides reviews for every movie. Absolutely, with growing up the number of websites which provide these reviews, the mission to search and decide become more complex.

Here machine learning play an important role to facilitate the difficult task to decide which movie you should watch. Sentiment analysis uses machine learning to analyze the reviews of people for many movies. So any one can find an overall positive or negative feedback of the movie. It not only classify the movie reviews not in positive or negative But also can give a stochastic rating like average, good, very good and excellent. So, it reduces the time to decide which movie to watch.

Datasets and Inputs

The datasets which I used here is come from Kaggle competition **Bag of Words Meets Bags of Popcorn.** The competition was based on using machine learning and deep learning in sentiment analysis. They tried to analyze the Movies review through 2 datasets.

The data set consists of 50,000 **IMDB movie reviews**, specially selected for sentiment analysis. The sentiment of reviews is binary and that is meaning that the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does not include any of the same movies as the 25,000 review test set.

The training set exists in tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review. There is another file that is tab-delimited file and has a header row followed by 25,000 rows containing an id and text only. We will predict the sentiment.

- **ID**: integer represent the key of every row in the dataset.
- **Sentiment**: it the label of film reviews. It has Binary value which be 0 for negative or 1 for positive.
- **Text**: it contains the reviews of every movie which we will use to train and build the model.

Solution Statement

The problem is binary classification problem. So I will build a model that tries to classify the reviews into 1/positive or 0/negative. I will try to take the advantage of labeled data and use supervised learning algorithm to predict the sentiment of the reviews. I will try different classifier and build different model. Also, using Deep learning, I can build a convolution neural network to predict the sentiment.

Benchmark Model

After building, different model, I will compare between them to get the model that has the best accuracy. Every model has a binary output 0 or 1. I will compare this value with the label of the row then, I will get accuracy of the model. The first one which I'm going is Bayesian method specially **Naïve Bayes**. Because it deals well with text, super simple and a good binary classification best on statistical model.

$$\underbrace{P(c\mid x)}_{\text{Posterior Probability}}=\frac{\overbrace{P(x\mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Evaluation Metrics

The Kaggle competition use **ROC curve** as an evaluation matric for this project. But I will try different metrics also like mean square error.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). The area under the curve is a measure of accuracy.

Project Design

The expected design for the movies reviews analysis. I will walk through the steps that exist in Kaggle competition.

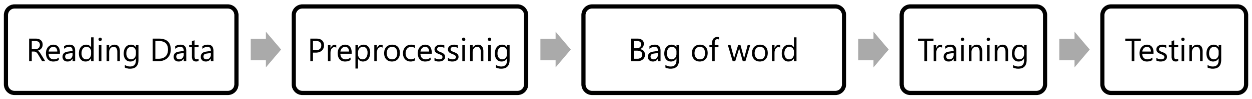
Reading the data: we will remove the delimiters and put the data from the file into arrays. That's for id and sentiment.

Preprocessing step and data cleaning: we will remove HTML tags from the texts (reviews). Also, we will remove punctuation, numbers and stop words using NLTK and regular expressions. Then, we convert the remaining text into lowercase. Then we add this word to array of data which had been created before.

Creating Features from a Bag of Words: we will convert the reviews text to some kind of numeric representation for machine learning by Bag of Words. It models each document by counting the number of times each word appears. Also, I can try to explore pre-trained word vector if it quicken the modelling step.

Training: I will model the problem and training it by the data before. I will try different technique like Bayesian classifier.

Testing: Test every model I had built to get the best one which has the best accuracy.



References

Angela Chapman, "Bag of Words Meets Bags of Popcorn", https://www.kaggle.com/c/word2vec-nlp-tutorial , 2014

Vidisha M. Pradhan, Jay Vala and Prem Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining", International Journal of Computer Applications, 2016.

Wikipedia contributors, "Receiver operating characteristic", Wikipedia, 2018.