Machine Learning Engineer Nanodegree

Capstone Proposal

Jitendra Reddy Muthyala
June 6th, 2017

Proposal

Domain Background

In machine learning, classification is that the problem of distinguishing to that of a group of classes a new observation belongs, on the basis of a training set of data containing observations whose class membership is known. And the field of study that focuses on the interaction between human languages and computers is known as Natural Language Processing (NLP).

NLP is a new way to analyze language things by computers, which allows us to understand and derive the meaning from human language in elegant and stylish way. By using NLP, we can translate one language data into another language, automatic summarization, named entity recognition, speech recognition, sentiment analysis, and topic segmentation. In present Internet world, opinions abound in tweets, blog comments, status updates, reviews and documents on social media and on other platforms has became important to understanding users opinions on some aspects.

Lots of research is done in this domain, few of them are

Citation: KL Santhosh Kumars "Opinion mining and sentiment analysis on online customer review", Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference. This paper is relevant to this project because it provides details on how to study the opinions of users by analyzing their opinions.

Problem Statement

Sentiment analysis is one of the most widely studied and challenging problems to be solved. The agenda in sentiment analysis is classifying the polarity of a given text at the document, sentence or feature level. Here I am trying to find weather the expressed opinion in a movie review is positive or negative.

Datasets and Inputs

In this project we will be working on a large dataset of movie reviews for the Internet Movie Database (IMDb) which has collected by Maas et al. The movie dataset consists of 50,000 movie reviews that are labeled as either positive or negative; positive in the sense that a movie was rated with more than five starts on IMDb, and negative means movies rated below five stars on IMDb, that neutral reviews are not included in the dataset. These reviews are used to study the user opinions by implementing Logistic Regression and Stochastic Gradient Descent models for classification of data.

A compressed archive of dataset (84.1MB) can be downloaded from http://ai.stanford.edu/~amaas/data/sentiment/ as a gzip compressed tarball archive.

Solution Statement

The downloaded data is not ready to use, so I need to assemble the individual text documents from the decompressed downloaded archive into a single CSV file. The individual text files from the pos and neg directories over the train and test subdirectories in the main aclImdb directory to be appended to a DataFrame with a integer class labels as 1 for positive and 0 for negative reviews.

The appended data need is to be cleaned by removing unwanted symbols (Mostly HTML markups) and unnecessary words (is, and, this, has) with help of Pythons regular expression modules and stop-words techniques of NLTK module respectively. By using word stemming techniques, we can transform each word into its root form which maps the words to the same stem. To transform the categorical data into numeric data I would like to use bag-of-words model which allows us to represent text as numerical feature vectors and also creates a vocabulary of unique tokens from the entire set of documents with included counts of how often each word occurs.

I will use logistic regression model to classify the movie reviews into positive and negative reviews. By using GridSearchCV object to find the optimal parameters form the regression model using 5-fold stratified cross-validation. As using GridSerchCV may make my model complex in processing huge data, I would like to try Stochastic Gradient Decent classifiers partial_fit function by streaming documents from our local drive and train a logistic regression model with the help of mini batches of documents.

Benchmark Model

The benchmark model of my algorithm is to predict the opinion in a submitted review with good accuracy. Hence a simple logistic regression model with 60% of classification accuracy will be considered as the Benchmark Model.

Evaluation Metrics

In this project, our aim to classify the movie reviews into positive or negative. Hence, I would like to consider recall, precision and F-Measure as the evaluation metrics used by my proposed benchmark and solution models.

Project Design

The basic idea behind my project is movie review classification. I start by cleaning and preparing my movie data and converting text documents into vector features by using NLP techniques. Then we are going to train the prepared data with a logistic regression model to classify positive and negative movie reviews. Next we will use a GridSearchCV object to find the optimal set of parameters for our model.

Next we will save the current state of a trained machine learning model using SQLite database for data storage. We will develop a web application using Flask Web framework to give a User Interface for our model to submit a model. Finally we will submit a review form the text box provided in web application, then the submitted review is classified as either positive or negative and displayed on the screen with prediction accuracy.