Machine Learning Nanodegree Capstone proposal

Kareem Ahmed Abdel-salam

Sentiment Classification using IMDB Dataset

October 06, 2018

Domain Background

Sentiment Classification is a common task in the field of Natural Language Processing, it is about extracting the sentiment of a person given a text they wrote as input.

Although this task is considered a complex task even for humans due to absence of facial expressions and voice tones in text but it is fairly much easier when given a paragraph of more of text written for the purpose of delivering sentiment such as the case with movie reviews.

The dataset was constructed back in 2011 by Andrew L. Mass et al. in the paper “Learning Word Vectors for Sentiment Analysis”, was a Kaggle challenge in 2016 and considered a Sentiment Analysis benchmark.

Problem Statement

Given the 50k labeled IMDB movie reviews in text format it is required to extract the sentiment and classify each review (from the 25k test set) to be either positive or negative, then compare the predicted results to the real labels and output the classification accuracy.

Accuracy= (number of correct labels) / (total number of labels) * 100%

Datasets and inputs

The IMDB movie review dataset was first proposed by Maas et al.in ”Learning Word Vectors for Sentiment Analysis” as a benchmark for sentiment analysis.

The dataset consists of 100K IMDB movie reviews and each review has several sentences.

The 100K reviews are divided into three datasets: 25K labeled training instances, 25K labeled test instances and 50K unlabeled training instances. Each review has one label representing the sentiment of it: Positive or Negative. These labels are balanced in both the training and the test set.

The dataset can be obtained from : http://ai.stanford.edu/~amaas/data/sentiment/

Solution Statement

The intended solution is using a Supervised Learning model to do the classification task.

Data is split into training and testing sets 25k each and each one is equally split between positive and negative reviews

Text data will be cleaned and preprocessed then proper features will be extracted from text where the classifier will use those features to capture patterns in positive and negative reviews throughout the training phase and then generalize and be tested on the testing set.

The solution will be evaluated by the classification accuracy percentage produced.

Accuracy= (number of correct labels) / (total number of labels) * 100%

Benchmark Model

A good benchmark to see whether a machine-learning solution works or not is random chance.

Given that the data is evenly split and the 25k test data is composed of 12.5 k positive samples and 12.5k negative samples then a random chance classifier will give a 50% accuracy, so my model must outperform this percentage in order to be effective.

Another historical benchmark to compare my solution to is the paper mentioned before and in the Credits section buy Andrew Mass : the highest classification accuracy was 88.89% on this dataset. So an efficient model will give close numbers or outperform it.

Project Design

1. Data preprocessing

   The first step in the solution is preprocessing the text data to remove anything that will make the classification harder and will confuse the classifier, this includes stop-words removal , HTML tags removal and punctuation removal and lemmatization (getting roots of words).

2. Feature Extraction

   The second step would be to modify the shape of data for the classifier and this includes transforming data into vectors using one of word embedding techniques (TF-IDF , word2vec , Doc2vec).

3. Supervised learning

   The third step to be able to classify reviews is training a supervised classifier such as SVM,Decision Trees,Logistic Regression or one of ensemble methods on a subset of data (the training set) and testing performance of learning on another subset (the testing set) to ensure that the model generalizes well and be able to judge the performance.

Credits

InProceedings{maas-EtAl:2011:ACL-HLT2011, author = {Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher},

title = {Learning Word Vectors for Sentiment Analysis},

booktitle = {Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies},

month = {June},

year = {2011},

address = {Portland, Oregon, USA},

publisher = {Association for Computational Linguistics},

pages = {142--150},

url = {http://www.aclweb.org/anthology/P11-1015} }

References

Potts, Christopher. 2011. On the negativity of negation. In Nan Li and David Lutz, eds., Proceedings of Semantics and Linguistic Theory 20, 636-659.