CMPE 255-02 DATA MINING

Programming Assignment Report

TEXT MINING

by

SHIVANI MANGAL (012530362)

Rank at time of submission: 2

F1-score at time of submission: 0.5383

Abstract

The goal of the assignment is to to develop predictive models that can determine, given a particular abstract, which one of 11 classes it belongs to. We will be using F1-score as the scoring function. I have tried the following classifiers - Lasso Regression, Ridge Regression, Gradient Boosting, Support Vector Machines, KNN and Multi-Layer Perceptron. Lasso regression performs the best and results in the highest micro-averaged F1score.

I have split my analysis into two python notebooks, one for data analysis and feature engineering and the second for model training.

Data Understanding and Transformation

The following steps were followed for **scrubbing of data**:

1. Removing NANs (missing values)
2. Removing unicode chars
3. HTML decode
4. Replacing URLs
5. Removing Numbers
6. Replace contractions to their full version like won't to will not etc (Effrosynidis, 2017)
7. Spelling fix for words with non-ascii characters using HunSpell
8. Removal of Punctuations and html encoding

Feature Engineering

The following features were created :

1. Length of the sentence.
2. Binned Length.
3. Number of words in the sentences.
4. Number of capital letters.
5. Number of special chars/punctuations.
6. Number of emoticons.
7. Number of continuous exclamatory marks.
8. Number of continuous stop marks.
9. Number of continuous question marks.
10. Number of positive/negative words in the sentence (keeping negation in mind).
11. Number of misspelled words.
12. Number of slangs (Effrosynidis, 2017)

13. Number of elongated words.
14. Overall positive/negative scores of 5 most significant non-stop words in the sentence.
15. Number of various emotions from the NRC lexicon

From the newly generated features, we observe high negative correlation with most significant sentiment score and number of positive words and slightly positive correlation of number of misspelled words with sentiment.
In the interest of time, instead of selecting features manually, I used lasso regression's feature selection nature.

Classifier Evaluation

I created a **30% data split and using it as hold out**. This split will not be used for any training purposes, and will only be used for calculating the final performance metric of each of the models. The remaining 80% data will be used to train models using **cross validations while performing grid search for hyper parameters.**

All classifiers were trained with three different vectorization, namely default vectorization, vectorization without stopwords and vectorization with stemming.

These are my observations for the various classifiers tried:

1. **Linear Support Vector Classifier:** A weak F1-score of around 0.36 was observed using linear support vector classifier while tuning the cost parameter through various values and using 5-fold cross-validation.

2. **Gradient Boosted Trees**: An f1-score of about 0.33 was observed using gradient boosted trees with tuning parameters like number of estimators, maximum depth, learning rate etc tuned at various values and 5-fold cross validation.

3. **Lasso Regression:** Lasso regression performed the best giving an f1 score of around **0.38** and I trained this model using maximum iteration values ranging from 200 to 300 and regularization strength ranging from 0.5 to 3. The most optimal classifier used a max iteration value of 220 and regularization strength parameter of 1.44.

4. **K-Nearest Neighbor Classifier**: KNN classifier was also trained with various values of K but it performed worst among all the classifiers evaluated giving an f1-score of 0.24

5. **Multi-Layer Perceptron:** I also tried building a neural network with 2 hidden layers of sizes 10 each which gave an f1 score of 0.29. However, again due to lack of resources on my system, the model was not able to converge successfully.

During the model training phase, I observed that models with linear tendencies like linear SVM and Lasso were performing better over tree based non-linear classifiers. Because of this observation, I skipped training non-linear models like decision trees and random forest.

Summary

Based on my observations, we can conclude that a Lasso regression classifier performs the best with f1-score value of 0.38

Resources

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017, September). A comparison of pre-processing techniques for twitter sentiment analysis. In *International Conference on Theory and Practice of Digital Libraries* (pp. 394-406). Springer, Cham.