December 19, 2015

1 Introduction

Our task consist in predicting a binary label that was given to movies reviews. The review is labeled 1 when it is considered positive and 0 when considered negative. To complete this task, we will start with a relatively simple model, and then analyze the misclassified reviews. The size of our training dataset is 25,000. Simple models have an accuracy close to 90 %. Therefore, if we train on 80% of the training dataset and evaluate on the remaining 5,000, we could collect 500 misclassified reviews.
Therefore, we don't want to read the mislabeled movie reviews one by one. Even if a focused human might not be able to classify every movie review correctly (If the commenter write a seemingly negative review but gives it a good rate) - which means that there might be Bayes noise in the dataset -, we have to try to understand what patterns in the data are still not captured by our model.

If the model used yields a probability of belonging to a specific class, we could analyze the one that were misclassified and had the highest probability of belonging to the class that was wrongly predicted.

Then we can read a small subset of the wrongly classified observations and try to verify hypotheses related to the unindentified patterns.

2 First model

Logistic Regression with stochastic gradient descent seems to yield the best results (If you google "state-of-the-art" sentiment classification, you can find an answer by olivier grisel which states that logistic regression is used by twitter engineers to classify tweets with the same binary labels).

What features should we use ?

- TF-IDF
- TW-IDF
- Extra handcrafted features

3 Analysis of the wrongly classified observations

AUC using probability of being wrongly classified ???