IST736 Text Mining
HW7

Compare MNB and SVMs for Kaggle Sentiment Classification

Task 1:

So far we have learned how to use sklearn to build MNB and SVMs models and evaluate them using various test methods and measures. Now consult the sklearn documentation and revise the instructor's given script to output confusion matrix, precision and recall values for the Kaggle Sentiment training data. Remember the sample script used 60% data for training and 40% for testing.

- Build a unigram MNB model and a unigram SVMs model.
- Print the top 10 indicative words for the most positive category and the most negative category from the MNB and SVMs models respectively.
- You can change other parameters to your preference. Report your choice and explain why.
- Report the confusion matrix, precisions, and recalls. Explain whether your models performed equally well on all categories, or some categories turn out to be easier or more difficult for MNB or SVMs.
- Submit your revised script along with your report.

Task 2:

Consult the sklearn website to learn more about the CountVectorizer. Revise the script to build a MNB model and a SVMs model based on both unigram and bigram. For fair comparison, please keep the same 60% for training and the rest 40% for testing. Also keep your other vectorization parameters the same as in Task 1.

- Compare the confusion matrix and other evaluation measures (accuracy, precision, recall). Discuss whether adding bi-grams was helpful for sentiment classification, based on MNB and SVMs respectively.

Task 3:

Now revise the sample script to build your best SVMs model by tuning parameters and using the entire training data set (changing from 60% to 100%). Report what parameters you used to train the model, and its cross validation accuracy.

Then use this model to predict the Kaggle sentiment test data. Submit the prediction result to Kaggle, use screenshot to show your accuracy and ranking.
https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/submit