Machine Learning Programming Assignment

Sentiment Analysis

Reflection questions :

Why it performed well?

1. It performed well as categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed.

Poorly?

1.I hated this movie. the ending sucked and to me it didnt make a whole lot of sense but thats my opinion. my kids liked it.

It is negative but output is positive

Why do you think it has failed?

Unable to understand grammar

What was hard about the target text?

Here it’s that mom didn’t like the movie but her kids like it. I checked it several times and the word “liked” making it strongly positive which might be used many times in the positive text file. So the system just calculates occurrence of words rather than grammar.

2. **This is the first film that Billy Bob Thornton let me down, as did the directors. Perhaps they should have watched some film noir before they tried to ape it. If they had they would have noticed that the main characters do more than just smoke and ponder.**

It Is negative but output is positive

Why do you think it has failed?

The classifier unable to understand the text, it just analyzes the words based on the probability of occurrence rather than meaning of sentence.

What was hard about the target text?

This is just a simple text where some strong negative words like “worst” or “bad” isn’t used. Classifier unable to recognize the meaning of the text.

3. **It's a highly overrated Star Trek movie! After saving Earth from the alien probe in ST:IV, the Enterprise crew is on vacation on Earth, but soon they will find themselves out to rescue alien ambassadors**

It is positive but output is negative

Why do you think it has failed?

It is basically data distribution rather than logic.

What was hard about the target text?

Dependences of existing variables in text files , in English a word which changes whole sentence which my classifier failing to recognize . Hard part is that it just depends on data in existing files.

In my improved version of bayes I'm just looking for all the punctuations.

Precision for positive : 0.948538919059
Precision for negative: 0.944346289753
Recall for positive : 0.988679245283
Recall for negative : 0.781718464351
F1-measure for positive: 0.855371074215
F1- measure for negatives : 0.968193216313

Improved bayes:

Precision for positive : 0.948538919059
Precision for negative: 0.944346289753
Recall for positive : 0.988679245283
Recall for negative : 0.781718464351
F1-measure for positive: 0.855371074215
F1- measure for negatives : 0.968193216313

why you think the systems performed poorly?

Data set is small than generative class will work well. We are just doing a bunch of counts.It will converge quickly. No curse of dimensionality. Data set is large don't use it. Don't use when features are correlated. Independence assumptions do not hold.

Ways to improve:

1. Using other distribution instead of data distribution .
2. Can use Fisher method - i think of Fisher as *normalizing* (more correctly, *standardizing*) the input probabilities. An Naïve bayes uses the feature probabilities to construct a 'whole-document' probability. The Fisher Method calculates the probability of a category for *each* feature of the document then combines these feature probabilities and compares that combined probability with the probability of a random set of features.