CSCI 575: Machine Learning (Course Project 2)

Implementing a Naive Bayes algorithm to classify texts into newsgroups

Name: Gautham Venkatesha Reddy

CWID: 10802128

Method Undertaken:

- We first obtain the path names using Python's glob function which gets us the respective paths of our text data.
- Filter out unnecessary texts from each data and store the preprocessed texts in a vector.
- We then pass this vector into a function called 'Count Vectorizer' which converts the words into a frequency matrix. The 'Count Vectorizer' function also helps remove common words such as 'a', 'the', 'an' etc. which may skew our analysis if included in our algorithm.
- We then split this frequency matrix into training and testing dataset using python's 'train_test_split' function.
- We then create a function that will take in which two/three newsgroup to classify into, counts the number of times this label occurs in the training dataset and calculates the probability of relative occurrence of these labels with each other.
- We then implement conditional probability to determine the probability of a class given each testing dataset. The maximum probability of the class is the prediction.
- We then determine the accuracy.

Results:

For pairwise binary classification, we see that similar classes have slightly lower performance than dissimilar classes. For example, 'rec.motorcycles' and 'rec.autos' have an accuracy of 88% while classes such as 'sci.med' and 'rec.sport.baseball' or 'talk.politics.guns' and 'comp.sys.mac.hardware' have accuracies of 92% and 95% respectively. It is also interesting to note that 'alt.atheism' and 'talks.religion.christian' have an accuracy of 88% although these two groups are highly dissimilar. This is because our method of using bag-of-words discounts grammar and sentiments of sentences which is crucial for determining such classes as these classes use very similar words (such as 'God', 'Believers', 'Faith' etc).

We also see this trend in tri-class classification. Similar groups such as 'talk.politics.mideast', 'talk.politics.misc' and 'talk.politics.guns' have much lower accuracy (73.6%) compared to dissimilar groups such as 'sci.med', 'comp.sys.mac.hardware' and 'talk.politics.guns' which has an accuracy of 92%.

Pairwise Binary Classification:

Class 1	Class 2	Accuracy (%)
alt.atheism	*soc.religion.christian*	*87.92*
talk.politics.guns	*comp.sys.mac.hardware*	*95.85*
rec.autos	*sci.space*	*93.02*
rec.motorcycles	*rec.autos*	*88.30*
sci.med	*rec.sport.baseball*	*92.00*

Tri Class Classification:

Class 1	Class 2	Class 3	Accuracy (%)
comp.graphics	misc.forsale	sci.crypt	88.29
talk.politics.mideast	talk.politics.misc	talk.politics.guns	73.59
rec.sport.baseball	alt.atheism	rec.autos	89.52
rec.sport.baseball	sci.space	rec.sport.hockey	85.33
sci.med	comp.sys.mac.hardware	talk.politics.guns	91.73