Task 2

Subhrajeet K B Ray

May 27,2020

1 Problem Statement

Classify whether a given chunk of text describes a commercial establishment that offers office space for rent

2 Classifier

The Naive Bayes Classifier is used for this problem statement.

3 Description

Naïve Bayes Classifier is a simple probability-based model which makes the naïve assumption that the features of the model are independent of each other. In the current context each word in the chunk is independent of the others. Naïve Bayes classifies by using Bayes equation to calculate probability of a chunk to be renting a space or not renting a space by conditioning on the current available vocabulary.Bayes equation is stated below.

$$p(RentingSpace|word1, word2, ...)\ \alpha\ p(RentingSpace) * \prod_i p(word_i|RentingSpace) \quad (1)$$

4 Procedure

4.1 Data Preparation

1.Remove all punctuation
2.Convert the chunk into lower case
3.Split the chunks to separate word

4.2 Data Split

We have divided the data between the training set and the test set. But ensure that the data split is not imbalanced.The data is split in ratio 7:3 within the training and testing data.A higher ratio for training data will lead to over-fitting and will not generalize.While a lower training set will also lead to bad generalization.

4.3 Calculation

We calculate $p(word_i|RentingSpace)$ based on the given equation

$$p(word_i|RentingSpace) \;=\; \frac{N_{word_i|RentingSpace} + \alpha}{N_{RentingSpace} + \alpha.N_{vocabulary}} \tag{2}$$

$$p(word_i|NotRentingSpace) \;=\; \frac{N_{word_i|NotRentingSpace} + \alpha}{N_{NotRentingSpace} + \alpha.N_{vocabulary}} \tag{3}$$

where
α is co-efficient of cases where the word does not occur in vocabulary
$N_{vocabulary}$ is total number of words in the dataset
$N_{word_i|NotRentingSpace}$ is part of chunk in dataset which are labeled as not renting
$N_{word_i|RentingSpace}$ is part of chunk in dataset which are labeled as renting space
$N_{RentingSpace}$ is total number of words in chunks which are labeled as renting a space
$N_{NotRentingSpace}$ is total number of words in chunks which are labeled as not renting a space

5 Performace

The given Naive Bayes Classifier has a prediction accuracy of 93.33 percentage when the training and test data were split in the ratio of 7:3. Any increase of traing data wil lead to overfitting and will not generalize well for newer dataset.Since the dataset is relatively small and the vocabulary is small, the classifier may not perform well when new dataset with new vocabulary is presented.This problem can be solved by using a bigger training dataset.

6 Conclusion

Thus Naive Bayes Classifier is a simple probabilistic model and work well for text classification such as the above problem statement,spam classifier, classifying news articles as sport,entertainment,politics etc