Applied Machine Learning Laboratory Assignment - 2

6th Feb 2016

Develop a Naive Bayes Classifier for Twitter dataset

You are given a dataset of tweets pertaining to 'Congress'. The word 'Congress' is ambiguous and hence, it contains tweets pertaining to Indian National Congress, US Congress, Mobile World Congress etc. The goal is to build a Naive Bayes Classifier which will disambiguate the tweets i.e, assign a class 1 if the tweet is related to Indian National Congress (INC), 0 otherwise.

1. Take the dataset and preprocess it. That is, remove # symbol, @ symbol, URLs, stopwords etc. (Use nltk.stopwords for stopword removal, nltk.word_tokenize() for tokenizing a sentence)

2. Use 80 % of the dataset for training and 20 % for testing.

3. The Naive Bayes algorithm is given by :

   P(c|d) = P(d|c) * P(c) / P(d)

   P(d) is the same for all the documents and hence can be ignored.
   P(c) = count of documents with c = 0 / size of the dataset.
   Similarly, compute P(c) for c = 1.

   If d = $w_1, w_2, ..., w_n$
   P(d|c) = P($w_1, w_2, ..., w_n$ |c) = P($w_1|c$) * P($w_2|c$) * … * P($w_n|c$)
   (because of the independence assumption)

3. Take a subset of the dataset where c = 0 and c = 1 respectively.
   Compute :
   P($w_i|c$) = count(w_i) / total no of words in that class
   Do for both c = 0 and c = 1.

Now the algorithm is trained.

4. Test the classifier on the test dataset (20 % of the actual dataset)
   Calculate P(c|d) for both the classes c = 0 and c = 1 and the argmax of the two values is the label for that class, i.e
   class predicted = argmax(P(c=0|d), P(c=1|d))

   Calculate the accuracy of the classifier.
   Accuracy = No of tweets correctly tagged / size of the test dataset
   Report as percentage.

Deliverables :

1. Source code of the program by 3:30 pm, 6th Feb 2016 to the Google group
2. Comparative study of Decision tree and Naive Bayes Classifier, which one performed better, possible reasons, difference in accuracy etc, Accuracy of NB Classifier and any other observations by 9:00 pm 7th Feb 2016 (Sunday) to the Facebook group