Machine Learning, Assignment 1,

Chetan There, 1001248919.

This assignment is to classify text documents by Naïve Bayes. We have given data set of 20 classes. Each class containing 1000 docs. Therefore, 20,000 docs in total. Please find methodology that I followed to solve this problem.

Step1: Preprocessing and Training

Read first 500 files from sub folder representing class and add them into mega doc of class. Repeat this for each subfolder. At the end we will get 20 mega docs containing words of input files belong to that class. Removing stop words (less meaningful common words) and Tokenizing words such that they contain alphabets only (I am removing numbers, punctuation marks etc). So each mega doc contains word and respective count.

Also calculating vocabulary length by merging all mega docs and then counting total number of unique words in train set. This vocabulary length will be used to add Laplace correction in testing.

Step 2: Testing

Now for testing I am considering remaining 500 docs from each subfolder (docs which were not used for training). I am considering doc one by one and calculating likelihood of it belong to each class.

I am neglecting prior probability since it will be same for each class.

Likelihood of a document for particular class is calculated by below formula.

Here,

w => word in the doc

c => word count of word w in mega doc of class

W => total word count of mega doc of class

V => vocabulary length calculated earlier

c1 => word count of w in doc d i.e. current doc

For each word in the document I am finding it’s count in mega doc of class, adding 1 as Laplace correction. Dividing it by summing total number of words in the mega doc and vocabulary length as Laplace correction denominator. Multiplying log of this ratio with the word frequency in the current doc. Basically converted product of likelihood of each word to logarithmic form.

I am getting 20 values of likelihood by above formula one for each class. I am then considering max value to predict class. Class with max value of likelihood of doc will be predicted class.

Step 3: Calculate Accuracy

I am comparing actual class label of that doc to find whether it was correctly classified. Checking for all 10,000 docs that belongs to test data and finding total accuracy.

Accuracy =

Instructions to Run : Please just change path variable which represents location of data set folder and run.

Output:

correct\_class\_dict {'comp.os.ms-windows.misc': 500, 'rec.sport.baseball': 500, 'sci.med': 499, 'comp.graphics': 500, 'rec.autos': 500, 'comp.sys.ibm.pc.hardware': 500, 'sci.electronics': 500, 'rec.sport.hockey': 500, 'talk.politics.guns': 500, 'soc.religion.christian': 497, 'talk.politics.mideast': 500, 'talk.politics.misc': 499, 'alt.atheism': 500, 'sci.crypt': 500, 'misc.forsale': 500, 'comp.windows.x': 500, 'rec.motorcycles': 500, 'comp.sys.mac.hardware': 499, 'sci.space': 500, 'talk.religion.misc': 500}

csum 9994

total test docs 9997

acc 99.96999099729919

dttime elapsed 0:34:14.505783

Note : Accuracy is very high. Assuming this is due to quality of input data.