INFORMATION RETRIEVAL

ASSIGNMENT 4

Submitted by –

Vaibhav Varshney(MT17065)

Dataset

- The dataset consist of five folders namely, “comp.graphics”, “rec.sport.hockey”, “sci.med”, “sci.space”, “talk.politics.misc”.
- The index of the folder names is considered as their labelled class.
- Each folder consist of 1000 files each.

Assumption

- It is assumed that the distribution followed for the conditional probability is binomial distribution. Hence the decision function for classification followed will be: (**for Naïve Bayes**)

$$c_{\text{map}} = \underset{c \in \mathbb{C}}{\arg\max} \left[\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c) \right]$$

- It is assumed that all the classes are equally probable.
- The test data doesn’t contains numerical terms.
- For **distance** calculation between two vectors, **cosine distance** has been used.

Methodology

(1) Preprocessing

Preprocessing is done in the following sequence:

- Punctuation removal
- Tokenization
- Stop words removal
- Stemming(using Porter Stemmer Algorithm)

(2) Feature Selection

For optimization of the performance of the model, the following heuristic is implemented. I have calculated the tf-idf values of all the words of the complete corpus. Further, on the basis certain percentage, I have selected the words with high tf-idf value from each document. Taking the union of all the words, a feature vector has been selected.

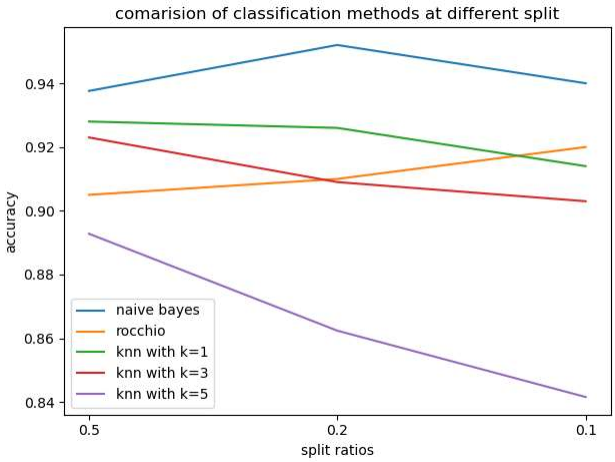
(3) Splitting of the Data

For the splitting of the dataset, “train_test_split” function has been used from the sklearn’s library. It provides the split into train size and test size as provided input by the user.

Analysis

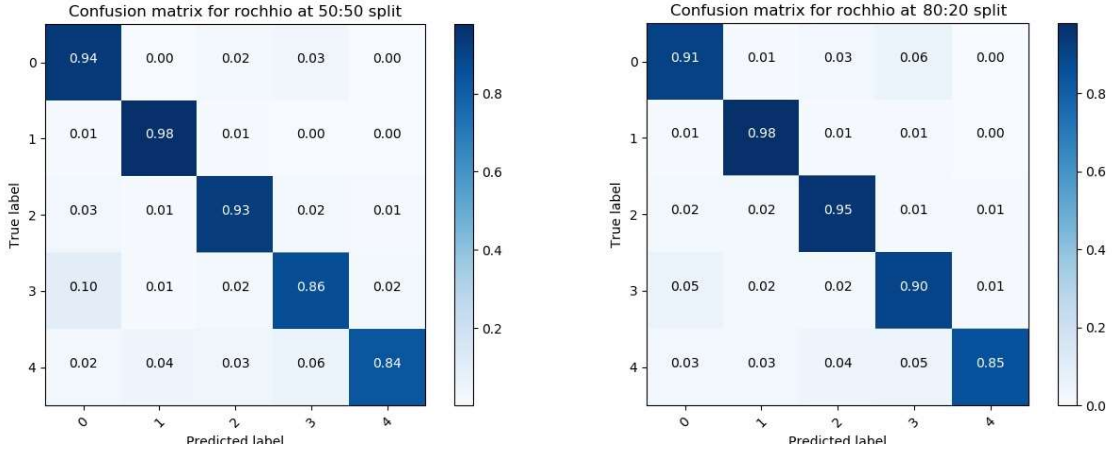
Split ratio	Method	Test Accuracy
50:50	Naïve Bayes	0.9376
	Rocchio	0.905
	KNN with k=1	0.928
	KNN with k=3	0.923
	KNN with k=5	0.8928
80:20	Naïve Bayes	0.952
	Rocchio	0.91
	KNN with k=1	0.926
	KNN with k=3	0.909
	KNN with k=5	0.8624
90:10	Naïve Bayes	0.94
	Rocchio	0.92
	KNN with k=1	0.914
	KNN with k=3	0.903
	KNN with k=5	0.8416

Plots

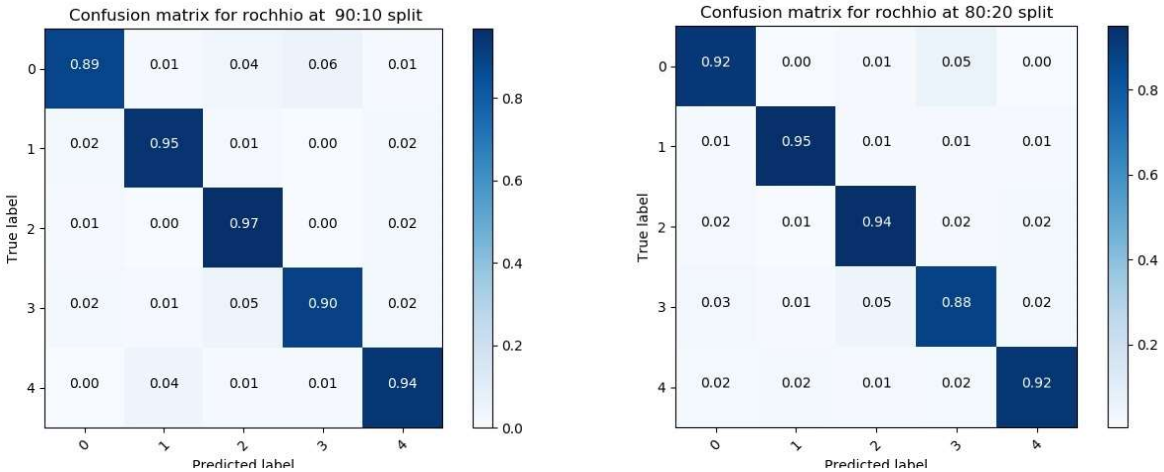


Confusion Matrices

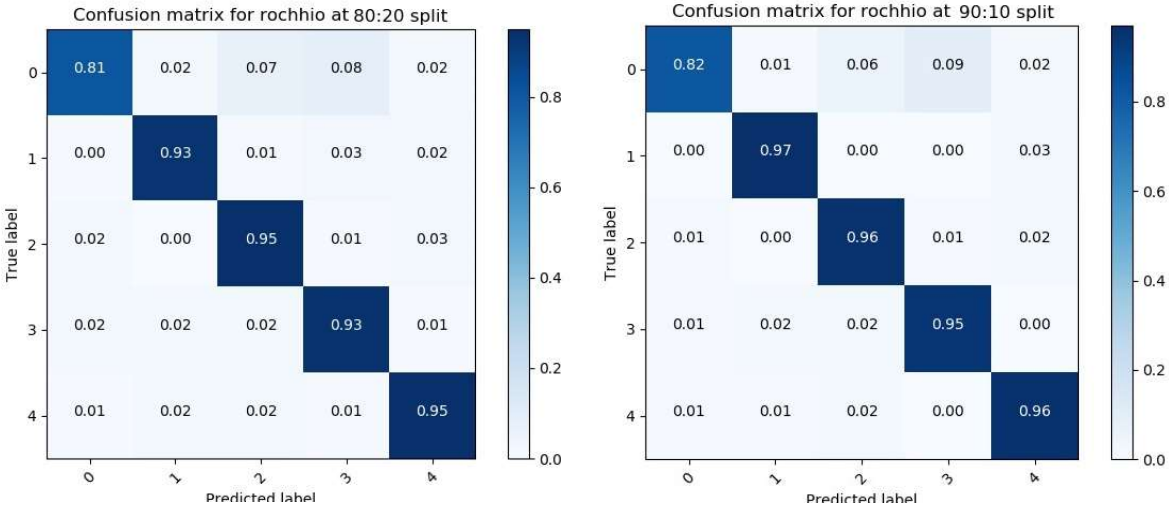
(1) Rocchio Algorithm

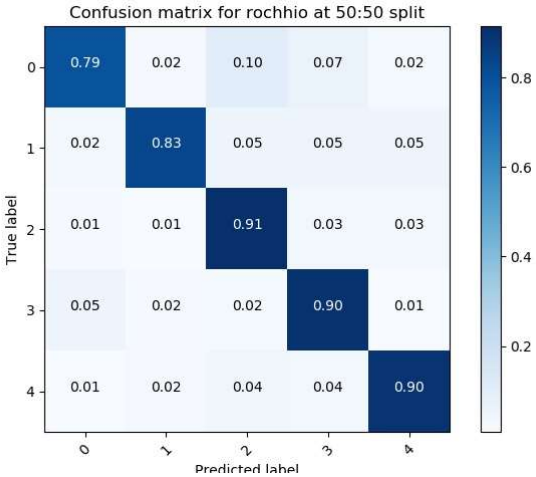


(2) KNN with k=1

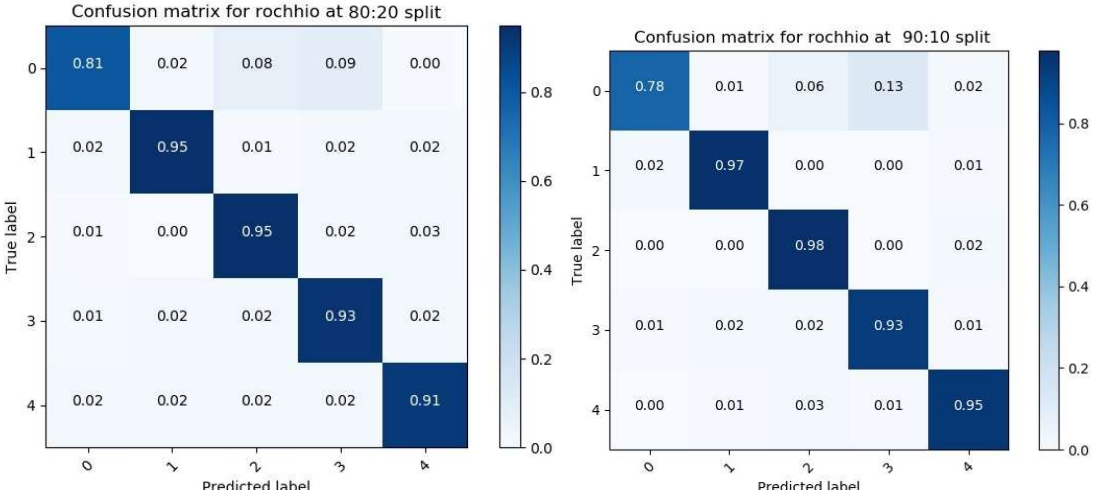


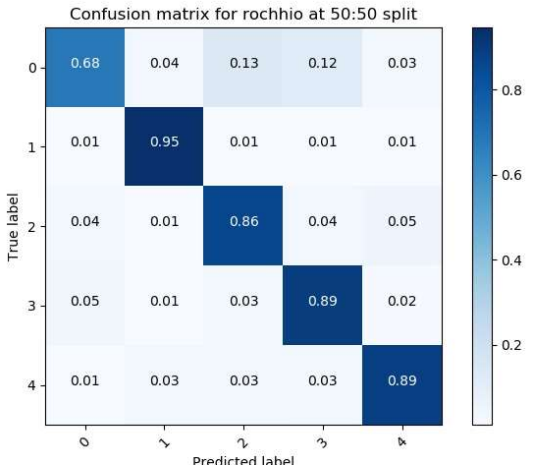
(3) KNN with k=3




Confusion matrix for rochhio at 50:50 split
0.79 0.02 0.10 0.07 0.02
0.02 0.83 0.05 0.05 0.05
0.01 0.01 0.91 0.03 0.03
0.05 0.02 0.02 0.90 0.01
0.01 0.02 0.04 0.04 0.90
True label
Predicted label
0 1 2 3 4
0.8 0.6 0.4 0.2

(4) KNN with k=5


Confusion matrix for rochhio at 80:20 split
0.81 0.02 0.08 0.09 0.00
0.02 0.95 0.01 0.02 0.02
0.01 0.00 0.95 0.02 0.03
0.01 0.02 0.02 0.93 0.02
0.02 0.02 0.02 0.02 0.91
True label
Predicted label
0 1 2 3 4
0.8 0.6 0.4 0.2 0.0
Confusion matrix for rochhio at 90:10 split
0.78 0.01 0.06 0.13 0.02
0.02 0.97 0.00 0.00 0.01
0.00 0.00 0.98 0.00 0.02
0.01 0.02 0.02 0.93 0.01
0.00 0.01 0.03 0.01 0.95
True label
Predicted label
0 1 2 3 4
0.8 0.6 0.4 0.2 0.0


Confusion matrix for rochhio at 50:50 split
0.68 0.04 0.13 0.12 0.03
0.01 0.95 0.01 0.01 0.01
0.04 0.01 0.86 0.04 0.05
0.05 0.01 0.03 0.89 0.02
0.01 0.03 0.03 0.03 0.89
True label
Predicted label
0 1 2 3 4
0.8 0.6 0.4 0.2

Observations

The following facts have been observed by seeing the results of the experiment:

- In different split ratios of train and test data, it has been observed that on increasing the training data size, an increase in the test data's accuracy is seen.
- Naïve Bayes outperformed all of the other approaches in each of the split of the dataset.
- In case of KNN, on increasing the value of k, a decrease in accuracy is observed.

Inference

- Naïve Bayes will work better as compared to Rocchio approach.
- With increase in test size, accuracy starts to decay, hence it can be inferred that the model tends to over-fit with increase in train size.
- It is also observed that with increase in the value of k, accuracy is decreasing at any of the given splits. It can be inferred that with increase in value of k, the variance of the model tends to move to zero. Hence the accuracy tends to decrease.
