EE 219 Project 1 Report

Cheng Ma 105033453
Jinxi Zou 605036454

1. Introduction

In this project, we are required to train the data set and use the result to identify the category of the test data. Classification is an essential element of data analysis, especially when dealing with large amounts of data. The categories we use in this project are listed in Table 1.

Table 1: the eight categories to be analysed

Computer technology	Recreational activity
comp.graphics	rec.autos
comp.os.ms-windows.misc	rec.motorcycles
comp.sys.ibm.pc.hardware	rec.sport.baseball
comp.sys.mac.hardware	rec.sport.hockey

2. Question a

When solving classification problem, we should make sure that the sizes of the data sets belonging to different classes should be equal. In question a, we plot the histogram of the number of the training documents per class and the resulted histogram is as follows:

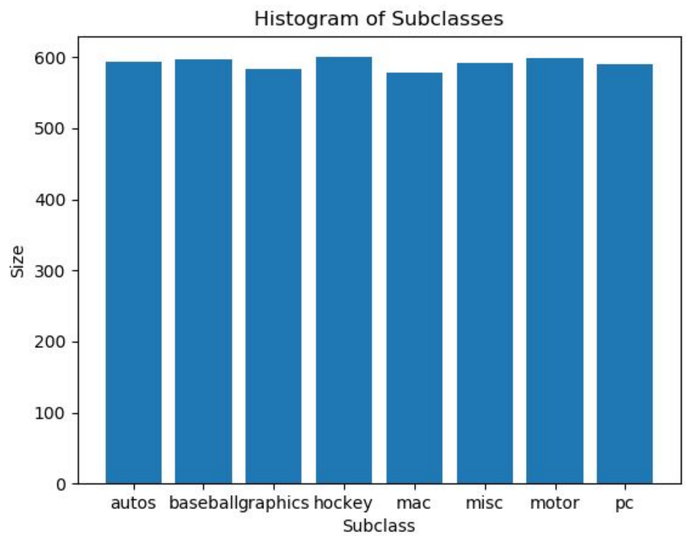

Fig 1: the histogram of the number of the train documents

And the frequency of these documents are 584, 591, 590, 578, 594, 598, 597, 600, accordingly, from which we could conclude that these documents are distributed equally, and thus no further balancing is required.

3. Question b

TFxIDF is the Term Frequency-Inverse Document Frequency, and it's used to capture the importance of a word to a specific document. In this task, we create a TFxIDF vector representations to the eight classes mentioned above.

Stop words are used to get rid of the nonsense words, such as 'that', 'is'. In the implementation, we used *text.ENGLISH_STOP_WORDS* defined in sklearn.feature_extraction. Stemming is used to convert same stemmed words into the original one, and we used *PorterStemmer* from nltk.stem to stem all the words in the documents.

The result we get are as follows, when the parameter min_df = 2 and min_df = 5:

Table 2: the number of terms when min_df = 2

categories	**number of terms**
comp.graphics	5151

comp.os.ms-windows.misc	9132
comp.sys.ibm.pc.hardware	4509
comp.sys.mac.hardware	4224
rec.autos	5327
rec.motorcycles	5419
rec.sport.baseball	5190
rec.sport.hockey	6312

Table 3: the number of terms when min_df = 5

categories	number of terms
comp.graphics	2005
comp.os.ms-windows.misc	3115
comp.sys.ibm.pc.hardware	1772
comp.sys.mac.hardware	1732
rec.autos	2377
rec.motorcycles	2496
rec.sport.baseball	2362
rec.sport.hockey	2639

From the results we could get that when the number of terms dropped dramatically when min_df changes from 2 to 5.

4. Question c

This task is similar to question b. The only difference is that in the previous one, we use terms in the documents of the same class, and in this task, we gather all the documents in each class and treat them as a single 'big' documents to the the TFxIDF. And the following is the 10 most significant terms in these four classes.

Table 4: the 10 most significant terms

ibm.pc.hardware	mac.hardware	misc.forsale	soc.religion.christian
floppy	lciii	hulk	clh
isa	c650	printer	christian
drives	powerbook	cd	faith
pc	duo	obo	christianity
disk	simms	hiram	athos
bus	nubus	condition	bible
bios	centris	forsale	christ
controller	scsi	wolverine	church
ide	quadra	dos	jesus
scsi	mac	shipping	christians

5. Question d

In this part, we reduce the dimension of the TFxIDF matrix from task b in order to get a better performance when using learning algorithms. We use two ways for dimension reduction, one is LSI (Latent Semantic indexing), and the other is NMF (Non-Negative matrix Factorization).

6. Question e

We use learning algorithms to get the desired classifier, which will then be used to classify the test data. And the classes we need to separate documents into are 'Computer Technology' and 'Recreation Activity', which is a binary class. To achieve this, first we need to put all the documents in the subclasses of these two classes into one class, and use this to do the training.

When using hard margin SVM classifier (SVC) by setting gamma to 1000, the ROC curve we get is

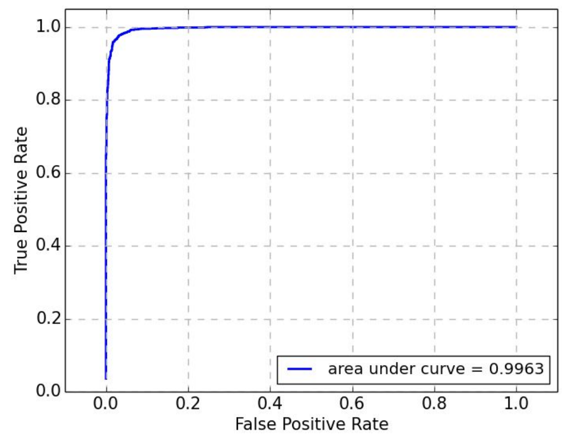

Fig 2: ROC curve when applying hard margin SVC

And the confusion matrix, accuracy, recall and precision are listed in the Table 6

Table 5: the classification report when gamma is 1000

Confusion matrix	accuracy	recall	precision
[1511 49] [37 1553]	0.973	0.97	0.97

When using soft margin SVC and gamma is 0.001, the results are as follows:

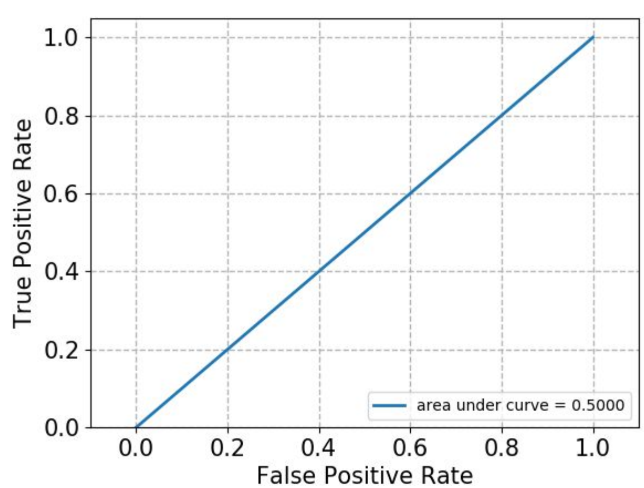

Fig 3: ROC curve when applying soft margin SVC

Table 6: the classification report when gamma is 0.01

Confusion matrix	accuracy	recall	precision
[0 1560] [0 1590]	0.505	0.50	0.25

7. Question f

After sweeping the value of k and comparing the mean of scores five fold cross validation, we found that when C is 100, the accuracy of prediction is highest at 0.971.

Table 7: the classification report when gamma is 100

Confusion matrix	accuracy	recall	precision
[[1504 56] [36 1554]]	0.971	0.97	0.97

8. Question g

In this task, we use naive Bayes algorithm to do the classification. since Multinomial naive Bayes requires non-negative features, thus we use NMF to reduce the dimension of the matrix.

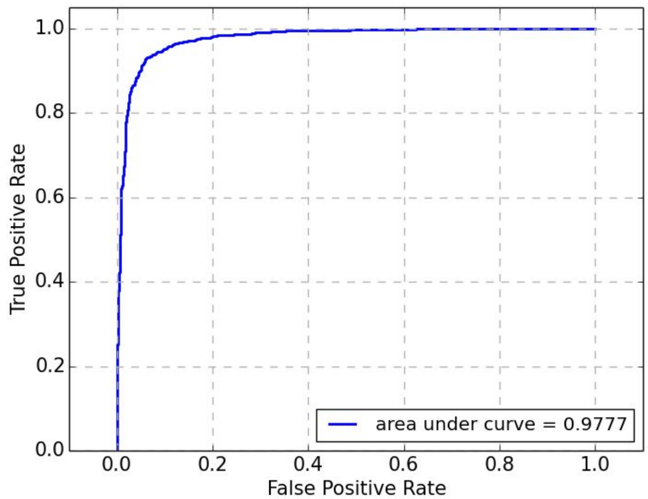

Fig 5: ROC curve when classifying by naive Bayes

Table 8: the classification report using naive Bayes

Confusion matrix	accuracy	recall	precision
[1369 191] [56 1534]	0.922	0.92	0.92

9. Question h

We use the logistic regression classifier in this part. The results are:

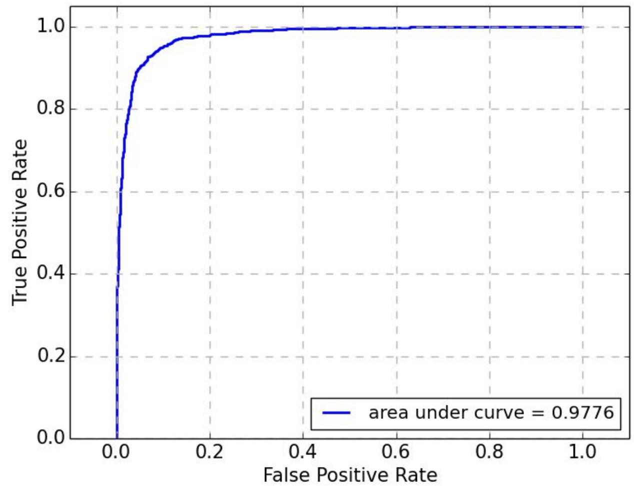

Fig 6: ROC curve when classifying by logistic regression

Table 9: the classification report using logistic regression

Confusion matrix	accuracy	recall	precision
[1424 136] [92 1498]	0.928	0.93	0.93

10. Question i_1

We repeat part h using both L1 and L2 regularization, and the penalty parameter C ranges from 0.01 to 1000. Followings are the classification report we get.

Table 10: the classification report when using L1 regularization and C = 0.01

Confusion matrix	accuracy	recall	precision
[1560 0] [1590 0]	0.495	0.25	0.50

Table 11: the classification report when using L1 regularization and C = 0.1

Confusion matrix	accuracy	recall	precision
[1246 314] [752 838]	0.662	0.66	0.68

Table 12: the classification report when using L1 regularization and C = 1

Confusion matrix	accuracy	recall	precision
[1452 108] [43 1547]	0.952	0.95	0.95

Table 13: the classification report when using L1 regularization and C = 10

Confusion matrix	accuracy	recall	precision
[1477 83] [47 1543]	0.959	0.96	0.96

Table 14: the classification report when using L1 regularization and C = 100

Confusion matrix	accuracy	recall	precision
[1478 82] [47 1543]	0.959	0.96	0.96

Table 15: the classification report when using L1 regularization and C = 1000

Confusion matrix	accuracy	recall	precision
[1479 81] [49 1541]	0.959	0.96	0.96

Table 16: the classification report when using L2 regularization and C = 0.01

Confusion matrix	accuracy	recall	precision
[41 1519] [0 1590]	0.518	0.75	0.52

Table 17: the classification report when using L2 regularization and C = 0.1

Confusion matrix	accuracy	recall	precision
[1269 291] [41 1549]	0.895	0.90	0.89

Table 18: the classification report when using L2 regularization and C = 1

Confusion matrix	accuracy	recall	precision
[1424 136] [92 1498]	0.928	0.93	0.93

Table 19: the classification report when using L2 regularization and C = 10

Confusion matrix	accuracy	recall	precision
[1445 115] [78 1512]	0.939	0.94	0.94

Table 20: the classification report when using L2 regularization and C = 100

Confusion matrix	accuracy	recall	precision
[1463 97] [56 1534]	0.951	0.95	0.95

Table 21: the classification report when using L2 regularization and C = 1000

Confusion matrix	accuracy	recall	precision
[1470 90] [50 1540]	0.956	0.96	0.96

From the tables above we can know that, when penalty parameter is small, the accuracy is pretty low in both L1 and L2 regularization. The accuracy is acceptable when using L1 regularization and C equals to 1, and the rate is 0.952; In L2 regularization, the accuracy is 0.895 when C is 0.1. That means L2 regularization performs better when C is low. And when C is large enough, typically when C is greater than 1, both of the two regularizations have a good performance.

11. Question i_2

We set the svc and bayes classifiers to one vs one classifier and one vs rest classifier and using two kinds of method to do dimension reduction and record the final result in following tables.

Table 22: the classification report when using ONE VS ONE SVC classifier and LSI

Confusion matrix	accuracy	recall	precision
[[319 45 27 1] [39 322 24 0] [23 15 350 2] [3 1 2 392]]	0.88	0.88	0.88

Table 23: the classification report when using ONE VS ONE Bayes classifier and LSI

Confusion matrix	accuracy	recall	precision
[[250 41 89 12] [89 161 124 11] [31 39 316 4] [1 1 7 389]]	0.71	0.71	0.72

Table 24: the classification report when using ONE VS REST SVC classifier and LSI

Confusion matrix	accuracy	recall	precision
[[312 54 26 0] [35 325 25 0] [20 13 354 3] [4 1 1 392]]	0.88	0.88	0.88

Table 25: the classification report when using ONE VS REST BAYES classifier and LSI

Confusion matrix	accuracy	recall	precision
[[242 41 101 8] [82 161 134 8] [28 33 324 5] [1 1 6 390]]	0.71	0.71	0.72

Table 26: the classification report when using ONE VS ONE SVC classifier and NMF

Confusion matrix	accuracy	recall	precision
[[337 32 22 1] [64 302 16 3] [44 14 331 1] [10 0 3 385]]	0.87	0.87	0.87

Table 27: the classification report when using ONE VS ONE Bayes classifier and NMF

Confusion matrix	accuracy	recall	precision
[[276 27 83 6] [50 251 81 3] [44 28 307 11] [5 1 7 385]]	0.78	0.78	0.79

Table 28: the classification report when using ONE VS REST SVC classifier and NMF

Confusion matrix	accuracy	recall	precision
[[321 41 28 2] [46 314 20 5] [27 17 341 5] [2 1 3 392]]	0.87	0.87	0.87

Table 29: the classification report when using ONE VS REST Bayes classifier and NMF

Confusion matrix	accuracy	recall	precision
[[288 31 68 5] [50 266 66 3] [41 18 322 9] [4 1 6 387]]]	0.81	0.81	0.81

From above tables, it is obvious that NMF has higher accuracy, precision and recall when the method of dimension reduction is the only variable. This was caused by of the non-negativity of NMF, since we used non-negative number to represent term frequency. One vs one classifier and one vs rest classifier had the simliar accuracy while SVC classifier always had a better performance than Bayes classifier.