Project 2 [*]

February 22, 2016

1 Dataset and Problem Statement

1.1 Part A

2 Modeling Text Data and Feature Extraction

2.1 Part A

2.2 Part C

10 Most significant features with TFICF scores:

```
   Class:comp.sys.ibm.pc.hardware
('scsi', 0.42809947405535098),
('drive', 0.31974715073739174),
('use', 0.22321146724973656),
('mb', 0.1968842917792481),
('ide', 0.18884045322355955),
('card', 0.15550544489414134),
('disk', 0.1477153861736914),
('control', 0.13410160556882411),
('dos', 0.12739571821010173),
('jumper', 0.11502197860943709)

Class:comp.sys.mac.hardware
('mac', 0.29656004455958551),
('use', 0.2251638723660454),
('scsi', 0.1890126678045923),
('appl', 0.18559985263916312),
('drive', 0.17315418917163491),
('mb', 0.17113312542860176),
('simm', 0.1637222404222877),
('problem', 0.14905214086203006),
('quadra', 0.13825360012263685),
('nubus', 0.12476544401311132)
```

[*] EE 239AS ; Winter 2016

```
Class:misc.forsale
('dos', 0.23204826904651923),
('new', 0.19885758974308329),
('sale', 0.18905428351306636),
('offer', 0.17954408282558959),
('use', 0.17596750747049816),
('includ', 0.17096030197337017),
('ship', 0.15768604272172954),
('price', 0.14163238406162046),
('wolverin', 0.1270563291946013),
('sell', 0.11874230178903535)

Class:soc.religion.christian
('god', 0.37385757805445796),
('christian', 0.27290603508238487),
('jesus', 0.23172842957647949),
('church', 0.20108044082283888),
('christ', 0.16786399940330063),
('peopl', 0.14539203709627244),
('say', 0.14475435272304318),
('bibl', 0.13177856565674298),
('believ', 0.12944992776554082),
('think', 0.12307308403324817)
```

3 Feature Selection

3.1 Part D

On applying LSI to the TFIDF matrix with k=50, each document was mapped to a 50 dimensional vector.

4 Learning Algorithms

4.1 Part E: Linear SVM

	Accuracy Predicted Computer Technology	Predicted Recreation Activity
Actual Computer Technology	1581	9
Actual Recreation Activity	236	1324

Table 1: Confusion Matrix: Linear SVM

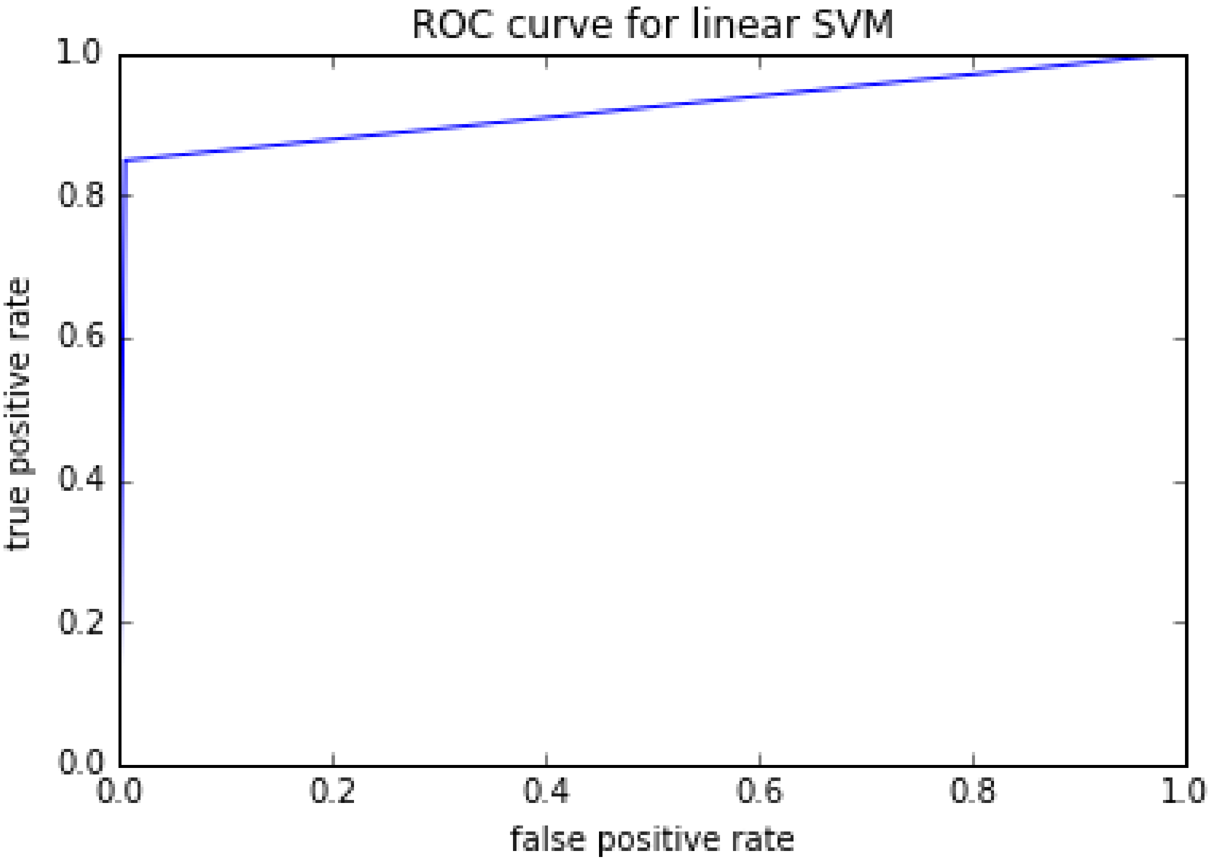


Figure 1: ROC curve for linear SVM

Learning Algorithm	Accuracy	Precision	Recall
Linear SVM	92.22	99.32	84.87

Table 2: Liner SVM

4.2 Part F: Soft Margin SVM

4.3 Part G Naive Bayes

4.4 Part G Naive Bayes

5 Multi-class Classification

5.1 Part I

The results for Multi-class classification are shown in the tables below. Table 9 contains the results for One vs Rest method and Table 10 contains the results for One vs One method.

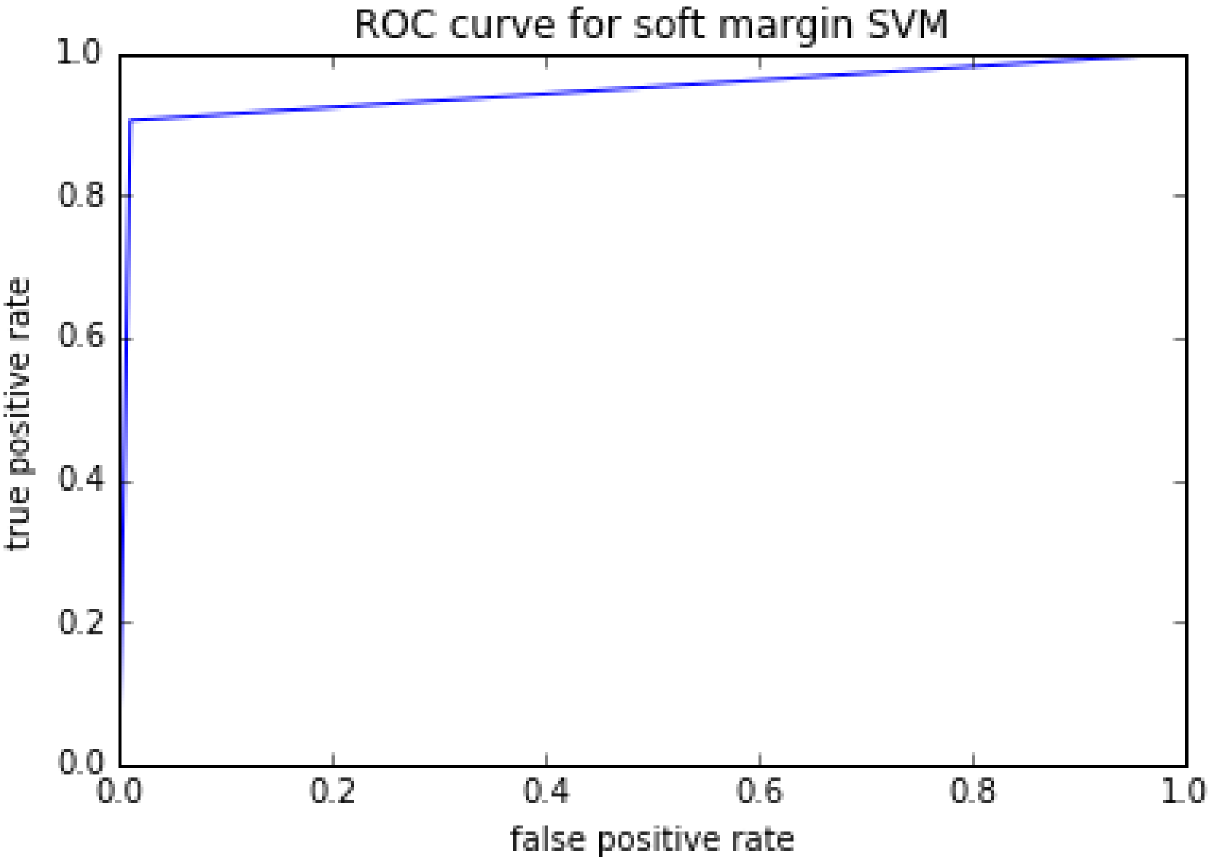


Figure 2: ROC curve for soft margin SVM

	Accuracy Predicted Computer Technology	Predicted Recreation Activity
Actual Computer Technology	1573	17
Actual Recreation Activity	148	1412

Table 3: Confusion Matrix: soft margin SVM

The confusion matrix for One vs One methods are shown below in figure 5 and Confusion matrix for One vs Rest methods are in figure 6

Learning Algorithm	Accuracy	Precision	Recall
Soft margin SVM	94.76	98.81	90.51

Table 4: soft margin SVM

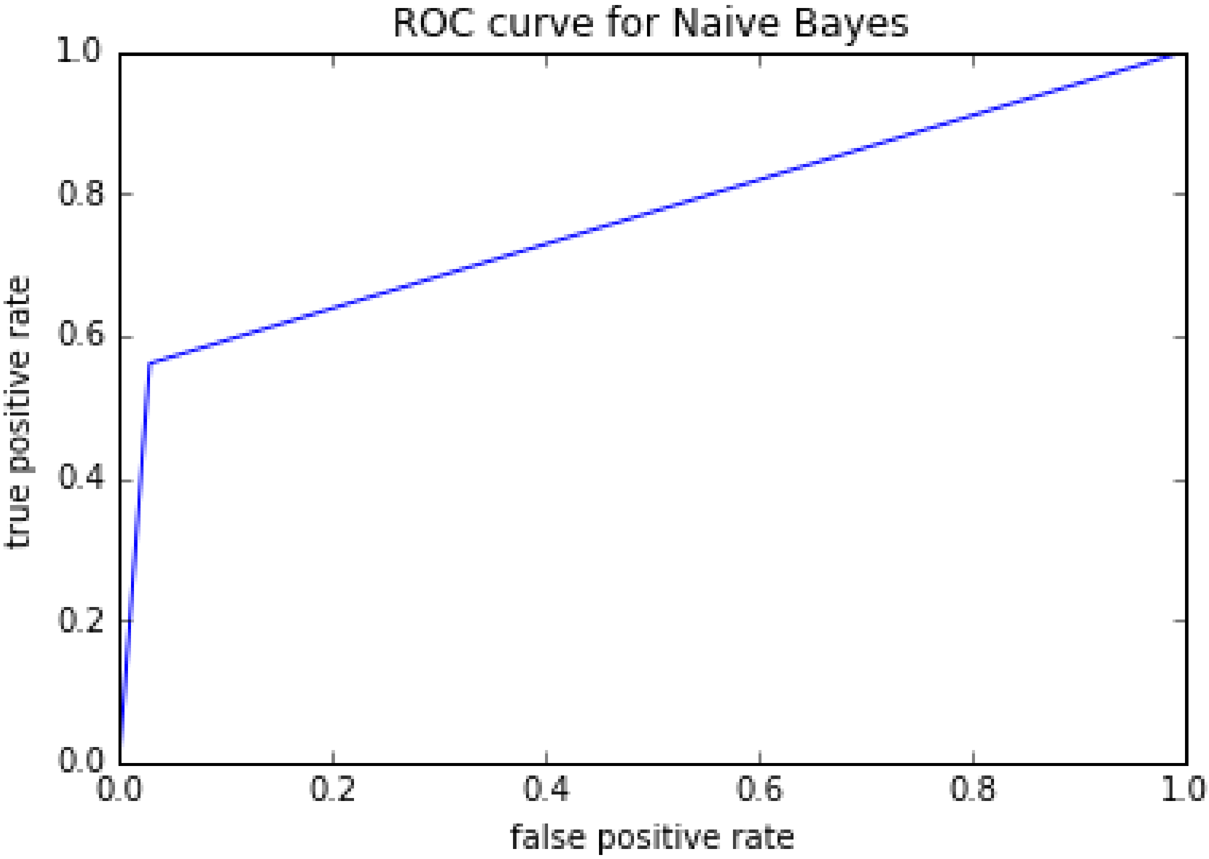


Figure 3: ROC curve for Naive Bayes

	Accuracy Predicted Computer Technology	Predicted Recreation Activity
Actual Computer Technology	1544	46
Actual Recreation Activity	685	875

Table 5: Confusion Matrix: Naive Bayes

Learning Algorithm	Accuracy	Precision	Recall
Gaussian Naive Bayes	76.79	95.00	56.08

Table 6: Naive Bayes

	Accuracy Predicted Computer Technology	Predicted Recreation Activity
Actual Computer Technology	1519	71
Actual Recreation Activity	23	1537

Table 7: Confusion Matrix: Logistic Regression

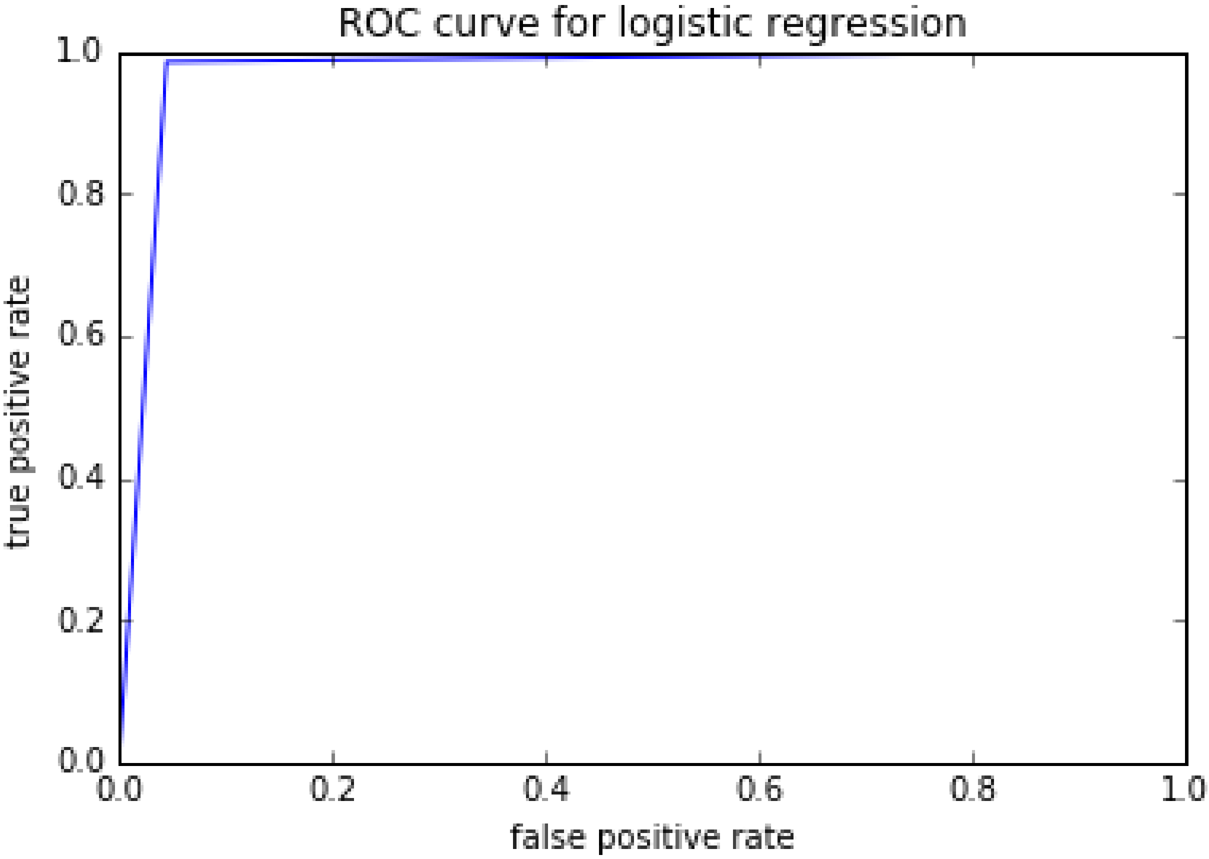


Figure 4: ROC curve for Logistic Regression

Learning Algorithm	Accuracy	Precision	Recall
Logistic Regression	97.01	95.58	98.52

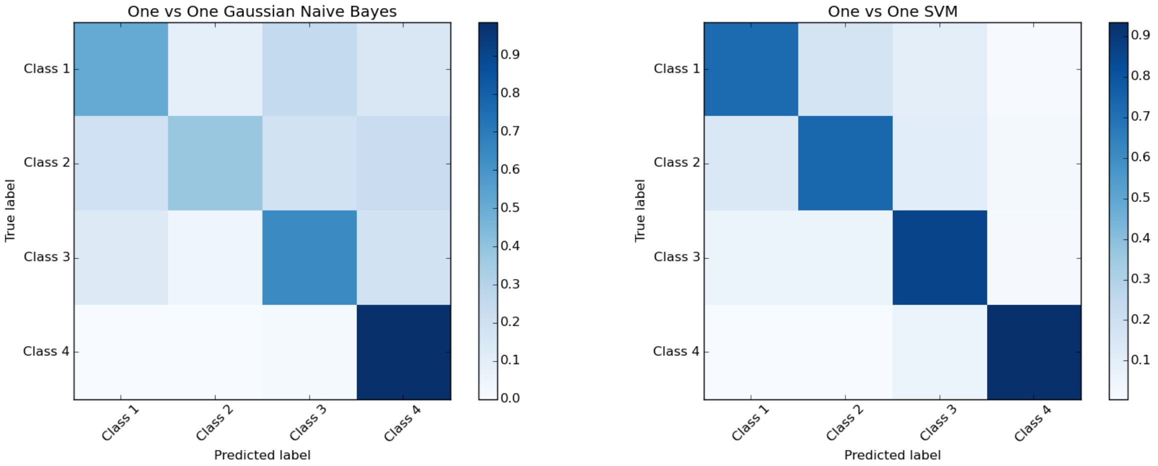
Table 8: Logistic Regression

Learning Algorithm	Accuracy	Precision	Recall
Gaussian Naive Bayes	63.32	64.50	63.32
Linear SVM	81.40	81.50	81.40

Table 9: One vs Rest

Learning Algorithm	Accuracy	Precision	Recall
Gaussian Naive Bayes	64.53	65.47	64.53
Linear SVM	80.89	81.28	80.89

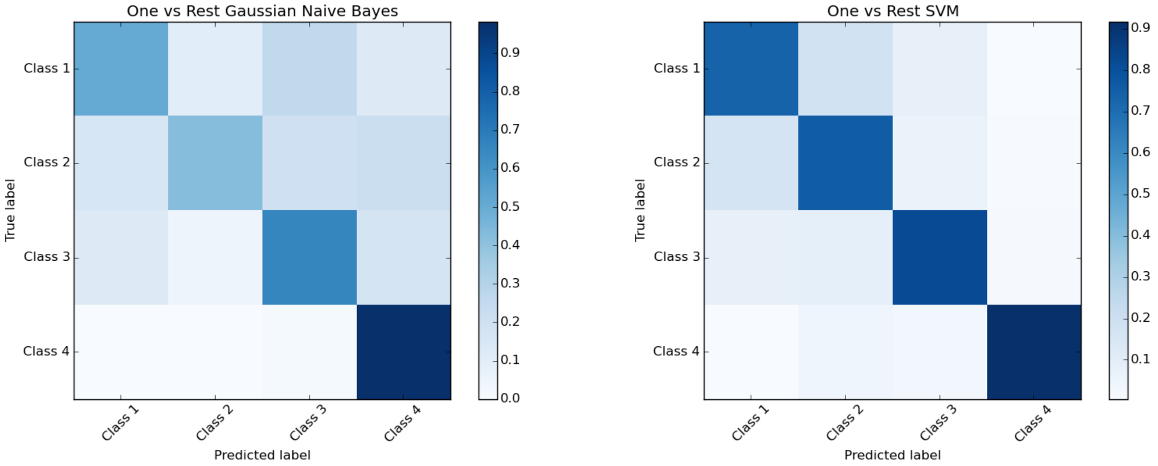
Table 10: One vs One



(a) Gaussian Naive Bayes

(b) Linear SVM

Figure 5: Confusion Matrix for One vs One Method



(a) Gaussian Naive Bayes

(b) Linear SVM

Figure 6: Confusion Matrix for One vs Rest Method