

Project 2 Report *

February 22, 2016

Tushar Sudhakar Jee (004589213)
Shubham Agarwal (204587029)
Pulkit Aggarwal(804587205)
Ishan Upadhyaya(204617837)

1 Dataset and Problem Statement

1.1 Part A

Number of documents in Recreational Activity are 2389
Number of documents in Computer Technology are 2343

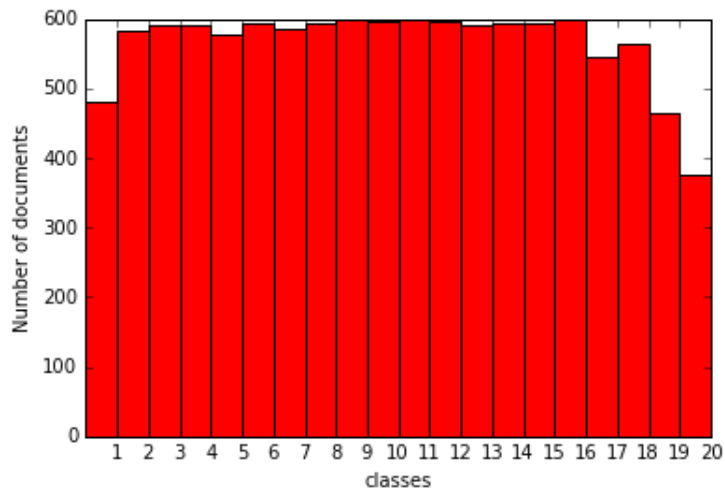


Figure 1: Histogram for the number of documents

2 Modeling Text Data and Feature Extraction

2.1 Part B

Final number of terms extracted are 34792.

*EE 239AS ; Winter 2016

2.2 Part C

Ten Most significant features with TFICF scores:

	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	misc.forsale	soc.religion.christian
1	scsi	mac	dos	god
2	drive	use	new	christian
3	use	scsi	sale	jesus
4	mb	appl	offer	church
5	ide	drive	use	christ
6	card	mb	includ	peopl
7	disk	simm	ship	say
8	control	problem	price	bibl
9	dos	quadra	wolverin	believ
10	jumper	nubus	sell	think

Figure 2: 10 most significant words for the given classes.

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

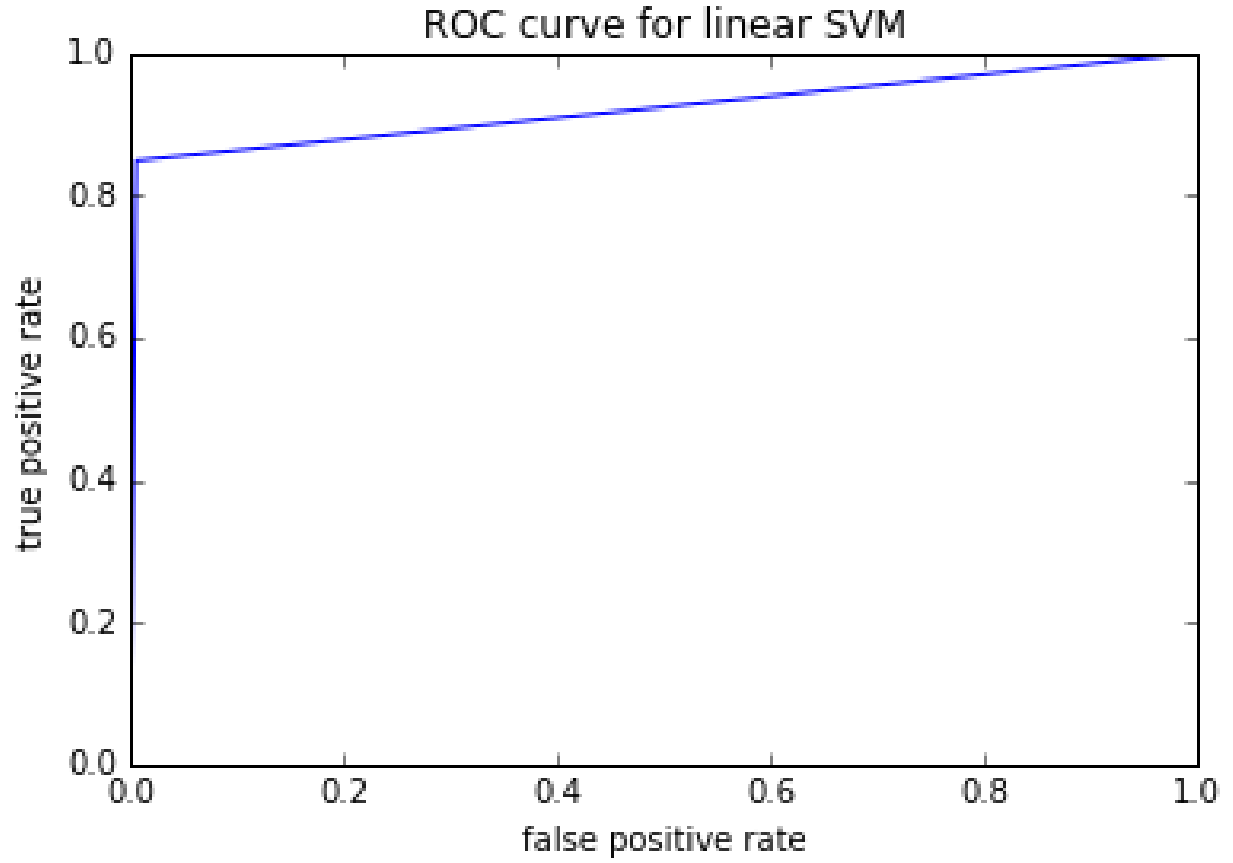


Figure 3: ROC curve for linear SVM

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

Table 1: Confusion Matrix: Linear SVM

Learning Algorithm	Accuracy	Precision	Recall
Linear SVM	92.22	99.32	84.87

Table 2: Liner SVM

4.2 Part F: Soft Margin SVM

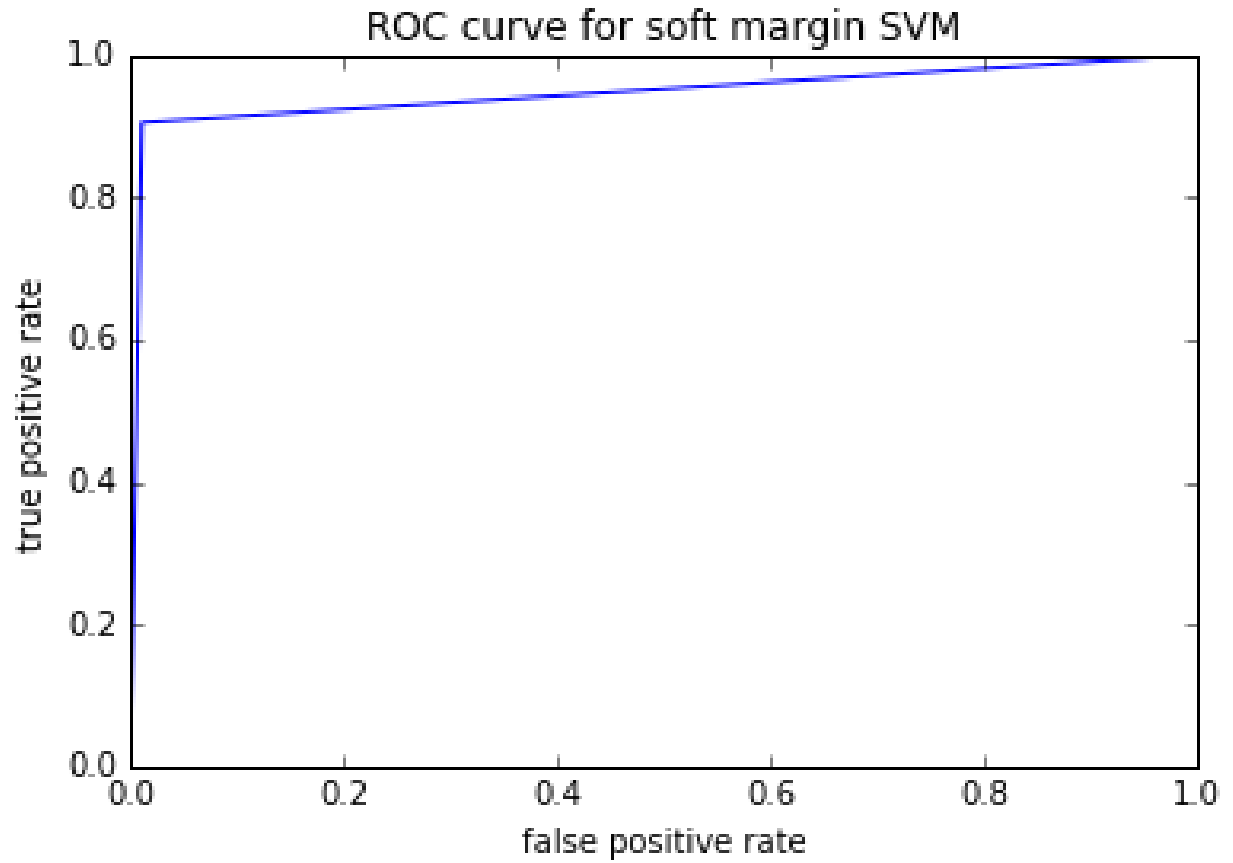


Figure 4: ROC curve for soft margin SVM

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

Table 3: Confusion Matrix: soft margin SVM

Learning Algorithm	Accuracy	Precision	Recall
Soft margin SVM	94.76	98.81	90.51

Table 4: soft margin SVM

4.3 Part G Naive Bayes

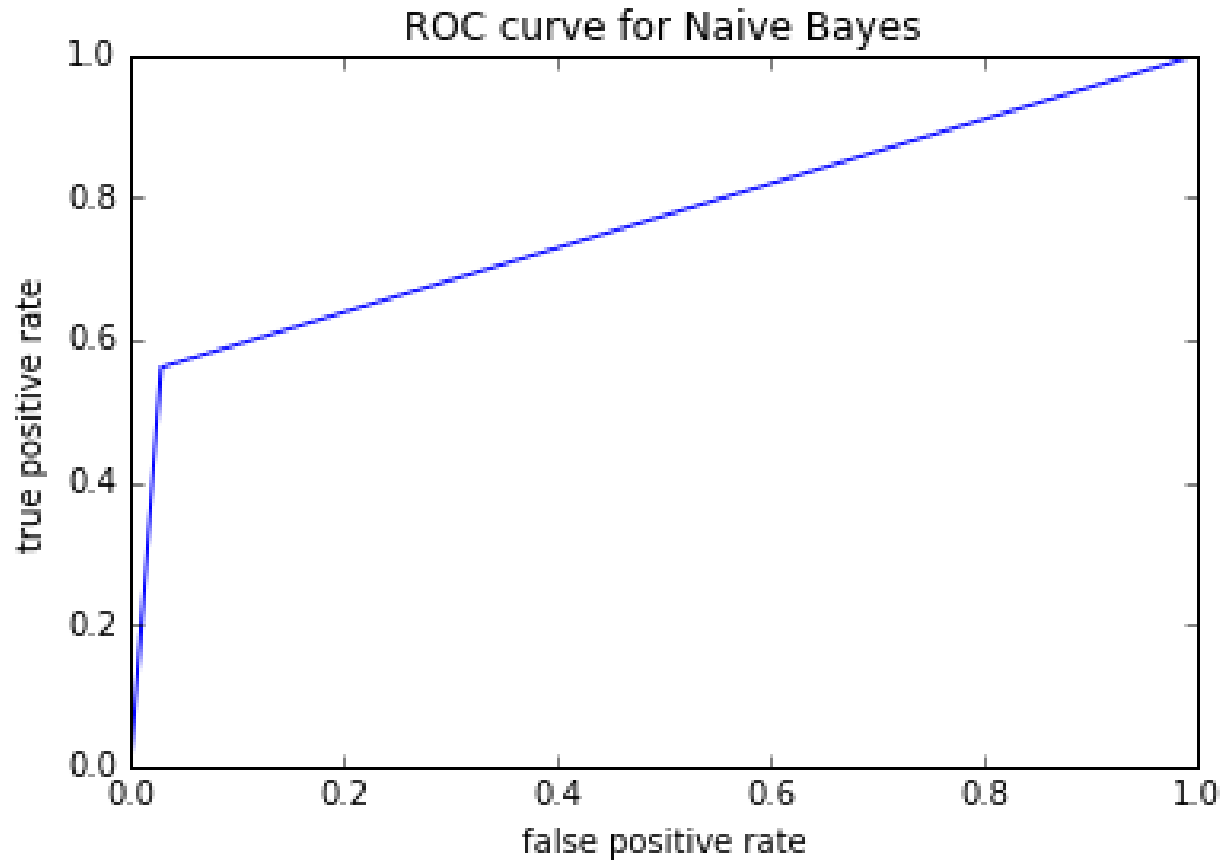


Figure 5: ROC curve for Naive Bayes

	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

4.4 Part H Logistic Regression

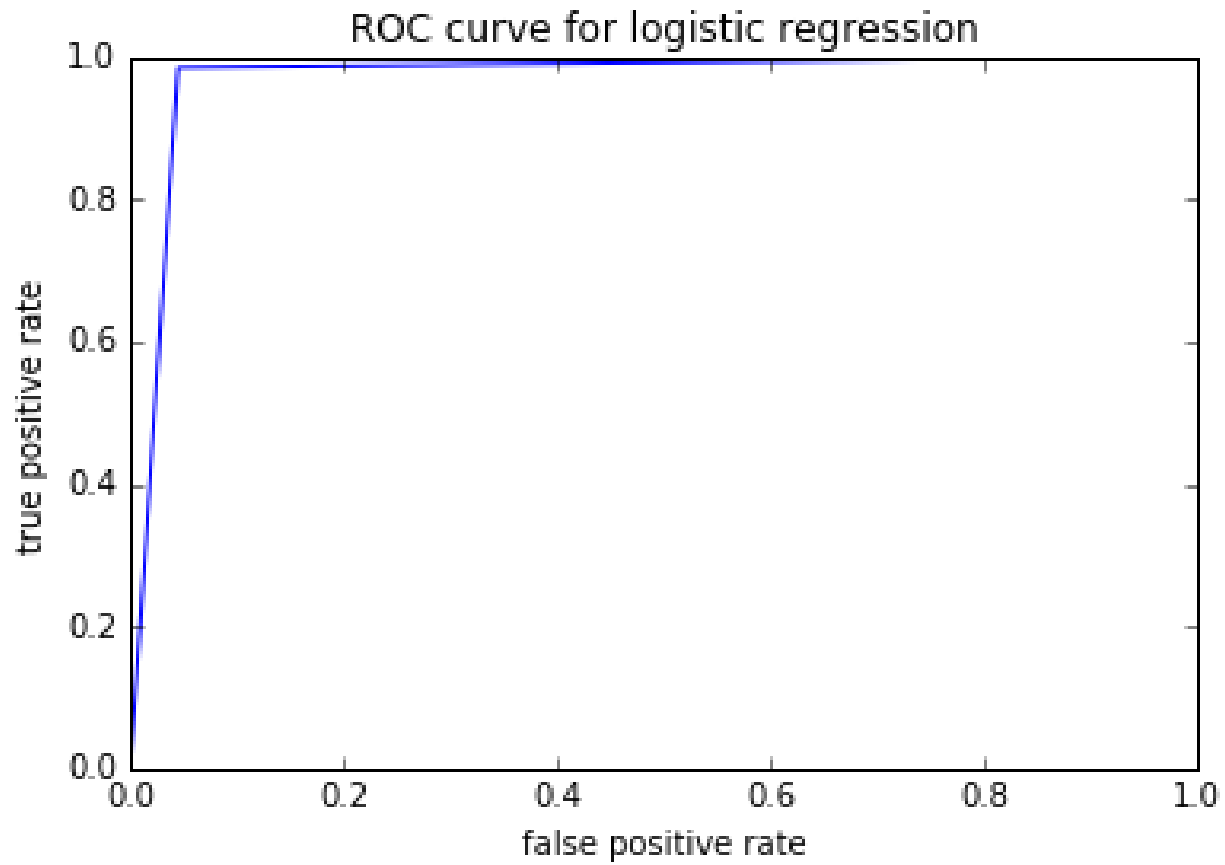


Figure 6: ROC curve for Logistic Regression

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

Table 7: Confusion Matrix: Logistic Regression

Learning Algorithm	Accuracy	Precision	Recall
Logistic Regression	97.01	95.58	98.52

Table 8: Logistic Regression

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.

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

The confusion matrix for One vs One methods are shown below in figure 7 and Confusion matrix for One vs Rest methods are in figure 8

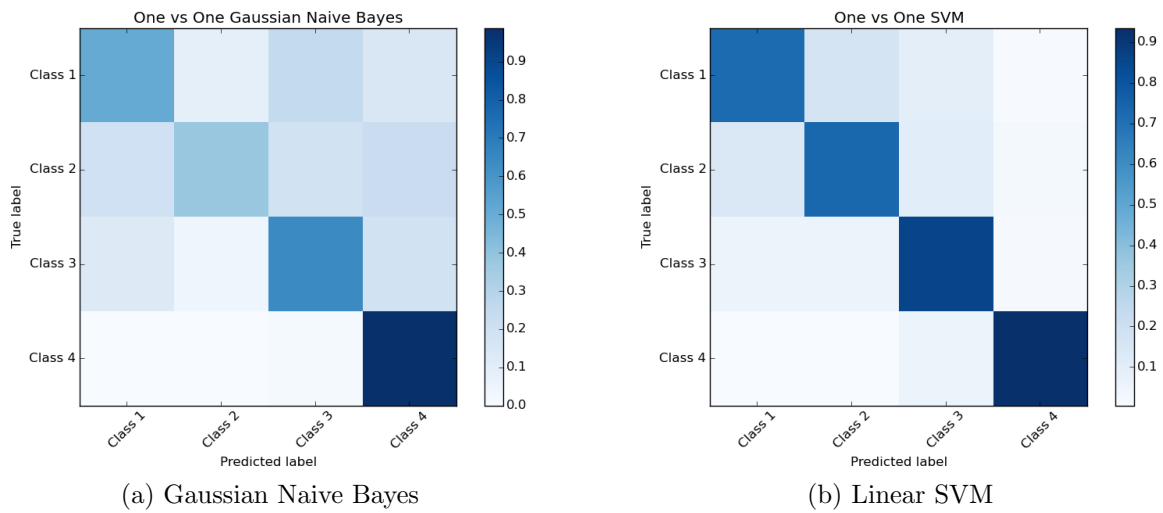


Figure 7: Confusion Matrix for One vs One Method

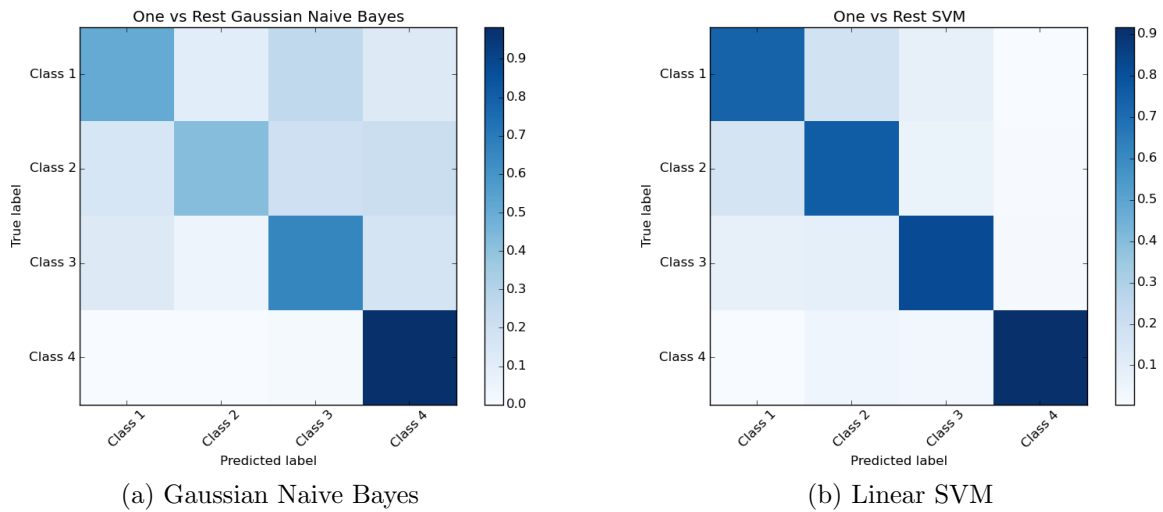


Figure 8: Confusion Matrix for One vs Rest Method