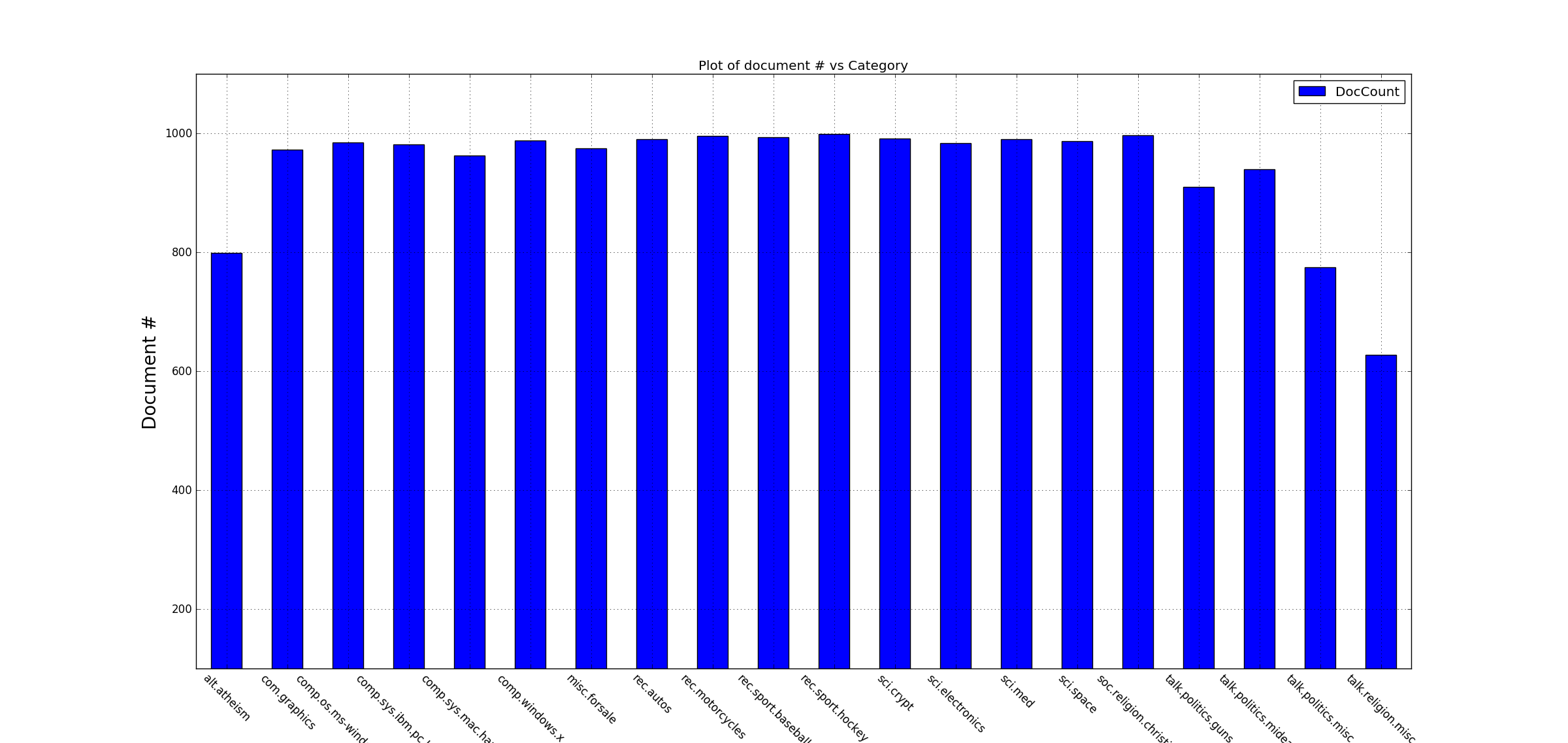
**EE219 Project 2 – Classification Analysis**– *Shubham Mittal* (104774903), *Swati Arora* (404758379), *Anshita Mehrotra* (904743371)

**Dataset and Problem Statement:**

1. The number of documents in ***Recreational Activity*** are ***3979***.

The number of documents in ***Computer Technology*** are ***3903***.  
Histogram of the number of documents per topic is shown below. We see that the distribution of the training samples is almost the same, except for tapering down on the sides for a few categories (which are not included in the later questions).

**Modeling Text Data and Feature Extraction:**

1. The final number of terms extracted equals **70465**
2. The most significant terms for the different classes are as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **comp.sys.ibm.pc.hardware** | | **comp.sys.mac.hardware** | | **misc.forsale** | | **soc.religion.christian** | |
| **Words** | **Counts** | **Words** | **Counts** | **Words** | **Counts** | **Words** | **Counts** |
| *edu* | 1423 | *edu* | 1899 | *edu* | 1751 | *god* | 2577 |
| *Drive* | 1403 | *line* | 1073 | *00* | 1215 | *christian* | 1760 |
| *line* | 1101 | *mac* | 1020 | *line* | 1044 | *edu* | 1638 |
| *com* | 1080 | *subject* | 997 | *subject* | 1008 | *church* | 937 |
| *subject* | 1024 | *organ* | 934 | *Sale* | 955 | *subject* | 1176 |
| *use* | 1010 | *use* | 803 | *Organ* | 981 | *jesus* | 904 |
| *scsi* | 1000 | *quadra* | 270 | *univers* | 564 | *Homosexu* | 653 |
| *organ* | 972 | *appl* | 664 | *com* | 548 | *peopl* | 1073 |
| *card* | 769 | *Problem* | 611 | *new* | 542 | *Sin* | 795 |
| *ide* | 573 | *centri* | 223 | *10* | 509 | *Line* | 1052 |

**Learning Algorithms**

We first load the training and test dataset for the categories which need to be classified. Classifiers need to classify documents into two classes: Computer technology and Recreational activity.

**Computer technology (indicated as class 0)** include subcategories *comp.graphics , comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware*.

**Recreational activity (indicated as class 1)** include subcategories *rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey*

**Ans e) Linear Support Vector Machine**

In this problem, Linear Support Vector Machine is trained to fit the test dataset. We used linear kernel to train the classifier. Statistics obtained are as follows:

**Confusion matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Computer technology (Class 0)** | **Predicted :** Recreational activity (Class 1) |
| **Actual : Computer technology (Class 0)** | 1501 | 59 |
| **Actual** : Recreational activity(Class 1) | 38 | 1552 |

**Recall and Precision score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 | 0.98 | 0.96 | 0.97 | 1560 |
| 1 | 0.96 | 0.98 | 0.97 | 1590 |

**avg / total** 0.97 0.97 0.97 3150

**Accuracy: 0.969206349206 (96.92%)**

**ROC Curve**



**Ans f) Soft Margin Support Vector Machine with 5-fold cross validation**

In this problem, Soft margin Support Vector Machine with different values of gamma is trained to fit the test dataset. It is done in order to minimize training error and avoid overfitting of data. To obtain best results, 5-fold cross validation is performed.

The best results were obtained at gamma = 0.1

**Confusion matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Computer technology (Class 0)** | **Predicted :** Recreational activity (Class 1) |
| **Actual : Computer technology (Class 0)** | 1501 | 59 |
| **Actual** : Recreational activity(Class 1) | 38 | 1552 |

**Recall and Precision score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 | 0.98 | 0.96 | 0.97 | 1560 |
| 1 | 0.96 | 0.98 | 0.97 | 1590 |

**avg / total** 0.97 0.97 0.97 3150

**Accuracy: 0.969206349206 (96.92%)**

**ROC Curve**



**Ans g) Naïve Bayes Algorithm**

**Confusion matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Computer technology (Class 0)** | **Predicted :** Recreational activity (Class 1) |
| **Actual : Computer technology (Class 0)** | 1293 | 267 |
| **Actual** : Recreational activity(Class 1) | 56 | 1534 |

**Recall and Precision score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 | 0.96 | 0.83 | 0.89 | 1560 |
| 1 | 0.85 | 0.96 | 0.90 | 1590 |

**avg / total** 0.90 0.90 0.90 3150

**Accuracy : 0.89746031746 (89.74 %)**

**ROC Curve**

****

**Ans h) Logistic Regression Classifier**

In this problem, Logistic Regression Classifier with different penalty function is trained to fit the test dataset.

Statistics for Logistic Regression Classifier with **‘l2’ penalty function** are as follows:

**Confusion matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Computer technology (Class 0)** | **Predicted :** Recreational activity (Class 1) |
| **Actual : Computer technology (Class 0)** | 1486 | 74 |
| **Actual** : Recreational activity(Class 1) | 35 | 1555 |

**Recall and Precision score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 | 0.98 | 0.95 | 0.96 | 1560 |
| 1 | 0.95 | 0.98 | 0.97 | 1590 |

**avg / total** 0.97 0.97 0.97 3150

**Accuracy: 0.965396825397 (96.53 %)**

**ROC Curve**

****

Statistics for Logistic Regression Classifier with **‘l1’ penalty function** are as follows:

**Confusion matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Computer technology (Class 0)** | **Predicted :** Recreational activity (Class 1) |
| **Actual : Computer technology (Class 0)** | 1491 | 69 |
| **Actual** : Recreational activity(Class 1) | 35 | 1555 |

**Recall and Precision score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 | 0.98 | 0.96 | 0.97 | 1560 |
| 1 | 0.96 | 0.98 | 0.97 | 1590 |

**avg / total** 0.97 0.97 0.97 3150

**Accuracy: 0.966984126984 (96.69 %)**

**ROC Curve**



Using ‘L1’ penalty function provides slight improvement in accuracy only when C=1, but reduces when C=0.01 as discussed below.

1. **Discussion:**

We enumerated through different values of regularization parameters for both ‘L1’ and ‘L2’ regularization. The accuracy results are summarized in the table below: (C is the inverse of the regularization strength)

|  |  |  |  |
| --- | --- | --- | --- |
| Penalty type | C=0.01 | C=1 | C=100 |
| L1 | 90.317 | 96.698 | 97.11 |
| L2 | 94.793 | 96.539 | 97.11 |
| L1 Sparsity | 96% | 22% | 0 |

Large values of C give more freedom to the model, while smaller values of C constrain the model. This leads to sparser solution in the L1 penalty case. As the value of C decreases, the coefficients of the fitted hyperplane become more and more sparse (for L1) affecting the performance of the model. L2 model, by its inherent design, remains almost unaffected by changing values of C as sparseness for all cases is 0.

As discussed above, using the L1 regularization improved the accuracy a bit in our case (for C=1). Both L1 and L2 regularizations aim to prevent overfitting in the data and make more generalized model which can perform better given a test data point. L2 regularization penalizes large values more than small values, thereby spreading the error across the training vector X. L1 regularization on the other hand, tries to have a sparse training vector, where some of the values are exactly zeros, and others may be large enough. L1 penalties are great at recovering truly sparse signals. In most cases where prediction is the ultimate goal, L2 regularization is preferred over L1 regularization; since if say two predictors are correlated, L1 simply picks any one, but L2 takes both of them and jointly shrinks the corresponding coefficients. That being said, L1 regularization is mostly used for feature selection in sparse feature spaces.

**Comparison of performance between different Regressors**



As seen in the figure, SVM and Logistic Regression classify documents with almost similar accuracy while Naïve Bayes is less accurate for this type of classification as compared to other two algorithms.

**MultiClass Classification**

**Ans(i) – Naïve Bayes and SVM**

The classifiers were trained for the following classes:

1. comp.sys.ibm.pc.hardware
2. comp.sys.mac.hardware
3. misc.forsale
4. soc.religion.christian.

OneVsOne and OneVsRest classification techniques used to train our classifiers.

**Results for OneVsOne Classification:**

1. ***Naïve Bayes Classifier***

|  |  |
| --- | --- |
|  | **Result\*100** |
| Recall | 73.5025918685 |
| Accuracy | 73.6741214058 |
| Precision | 77.0457299961 |

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| 278 | 18 | 94 | 2 |
| 70 | 186 | 125 | 4 |
| 47 | 19 | 323 | 1 |
| 0 | 0 | 32 | 366 |

1. ***SVM***

|  |  |
| --- | --- |
|  | **Result\*100** |
| Recall | 88.2601047851 |
| Accuracy | 88.3067092652 |
| Precision | 88.4576980617 |

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| 333 | 44 | 15 | 0 |
| 40 | 323 | 22 | 0 |
| 29 | 14 | 346 | 1 |
| 9 | 4 | 5 | 380 |

**Results for OneVsRestClassifier**

1. ***Naïve Bayes Classifier***

|  |  |
| --- | --- |
|  | **Result\*100** |
| Recall | 72.4044929407 |
| Accuracy | 72.5878594249 |
| Precision | 76.7645816972 |

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| 257 | 16 | 118 | 1 |
| 63 | 177 | 140 | 5 |
| 40 | 17 | 330 | 3 |
| 0 | 0 | 26 | 372 |

1. ***SVM***

|  |  |
| --- | --- |
|  | **Result\*100** |
| Recall | 88.8913870886 |
| Accuracy | 88.945686901 |
| Precision | 88.8656279393 |

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| 322 | 47 | 20 | 3 |
| 32 | 324 | 27 | 2 |
| 20 | 13 | 355 | 2 |
| 3 | 1 | 3 | 391 |