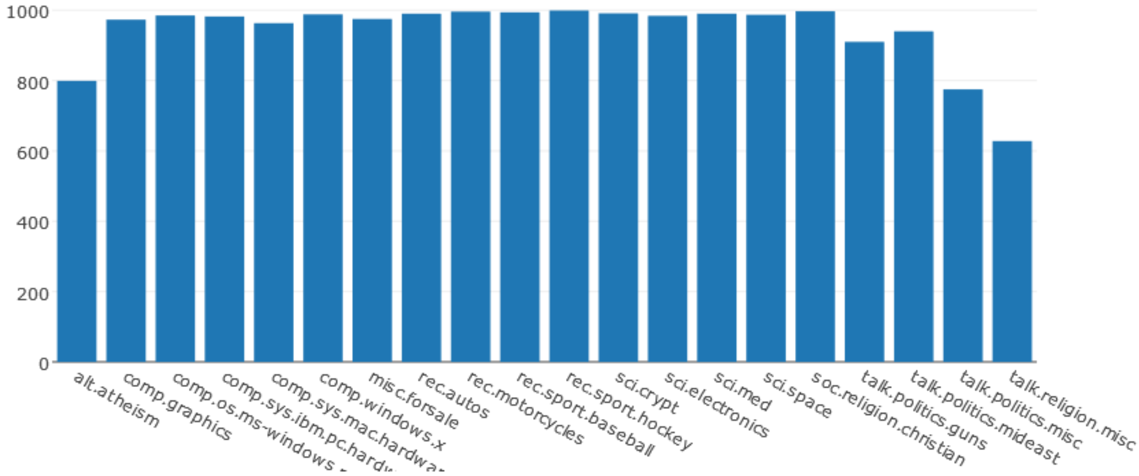
EE239AS - Project 2 Report

Group Members:

1) Rebecca Correia - 204587944
2) Salil Kanetkar - 704557096
3) Vedant Patil - 104590942

Problem a:

The figure below shows the number of documents in each of the 20 categories of the 20 News Group Dataset. The documents are roughly equally distributed in each category except the corner ones.



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

The number of documents in ***Recreational Activity*** are ***3979***.
The number of documents in ***Computer Technology*** are ***3903***.

Problem b:

The final number of terms are **79498.**

Problem c:

The first value is the TF-IDF score and the next is the frequent word.

The 10 most significant terms in **comp.sys.ibm.pc.hardware** are:

```
(0.12239993666838027, u'bio'),
(0.12701047726703196, u'system'),
(0.13301310101100342, u'bu'),
(0.15104401409532223, u'ide'),
(0.15189738362249117, u'use'),
(0.15256906198380346, u'control'),
(0.17057907198319602, u'disk'),
(0.18668316091455309, u'card'),
(0.31867893134542236, u'scsi'),
(0.36188525466168475, u'drive')
```

The 10 most significant terms in **comp.sys.mac.hardware** are:

```
(0.10655257867645329, u'bit'),
(0.12015027504407341, u'monitor'),
(0.12802572124297931, u'card'),
(0.13006999435940197, u'simm'),
(0.13013809858461353, u'problem'),
(0.1343206593494865, u'scsi'),
(0.14608894928492772, u'use'),
(0.18660408494578767, u'drive'),
(0.24046284303117602, u'appl'),
(0.3210999918577277, u'mac')
```

The 10 most significant terms in **misc.forsale** are:

```
(0.11896024837666791, u'game'),
(0.12877103639521825, u'condit'),
(0.13282075591329823, u'sell'),
(0.15544816758441793, u'price'),
(0.15684936166486935, u'includ'),
(0.15944674935819239, u'new'),
(0.17249014251892658, u'ship'),
(0.17444251552487361, u'do'),
(0.19752952756555531, u'offer'),
(0.20711836544519346, u'sale')
```

The 10 most significant terms in **soc.religion.christian** are:

```
(0.11525641639375862, u'say'),
(0.12021125025716464, u'would'),
(0.12312665118314295, u'believ'),
(0.12335335254530352, u'faith'),
(0.12368183841998988, u'one'),
(0.12456498898139731, u'christ'),
(0.16663523063137495, u'church'),
(0.1922538775880456, u'jesu'),
(0.22784459365436943, u'christian'),
(0.38063257451072369, u'god')
```

Problem d:

The dimensions after performing LSI are **50**

Problem e:

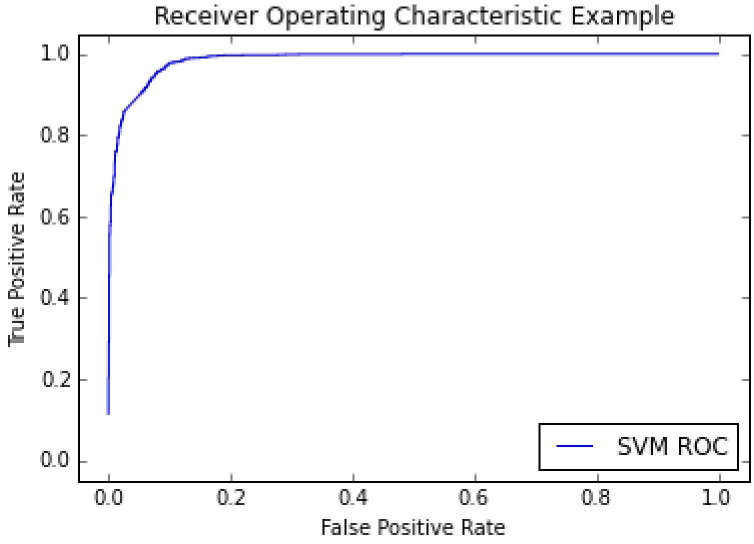
The accuracy for the model is **0.936508**
'0' is Computer Technology and '1' is Recreational Activity
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.93      1560
          1       0.92      0.96      0.94      1590
avg / total       0.94      0.94      0.94      3150
```

The confusion matrix is as shown below:

```
[[1428  132]
 [  68 1522]]
```

The ROC Curve is as shown below:



Problem f:

****

The value of Gamma is 0.001000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

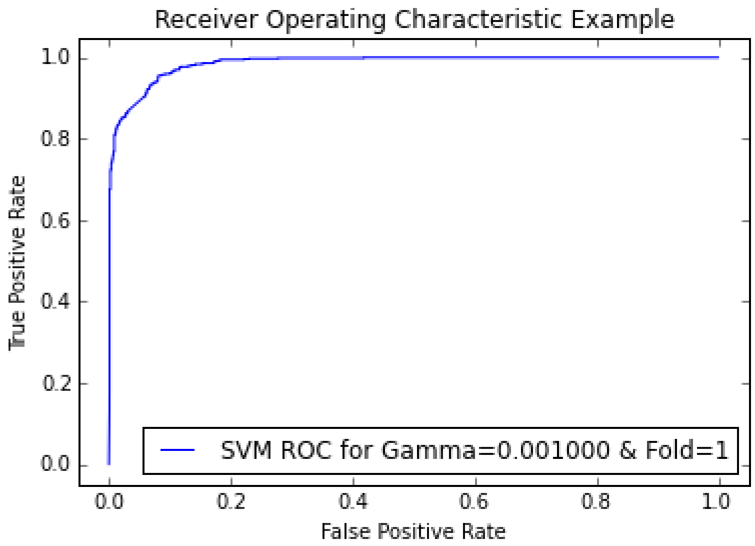
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796

avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

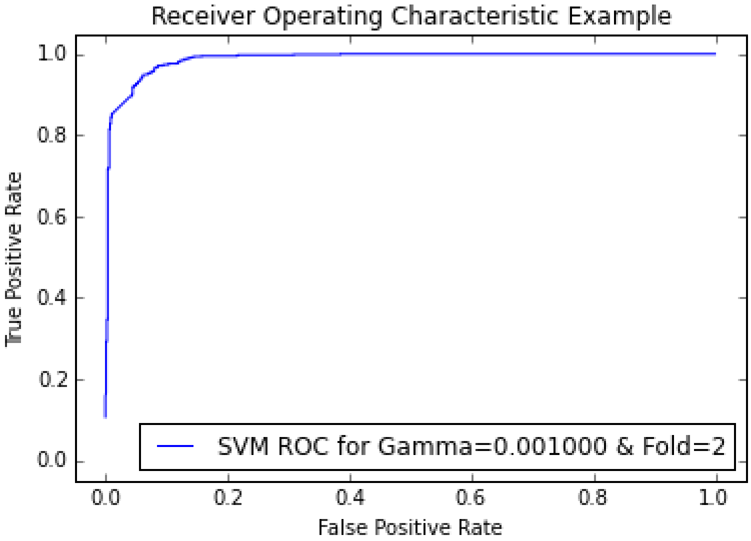
The precision and recall values are:

```
             precision    recall  f1-score   support
          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796

avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

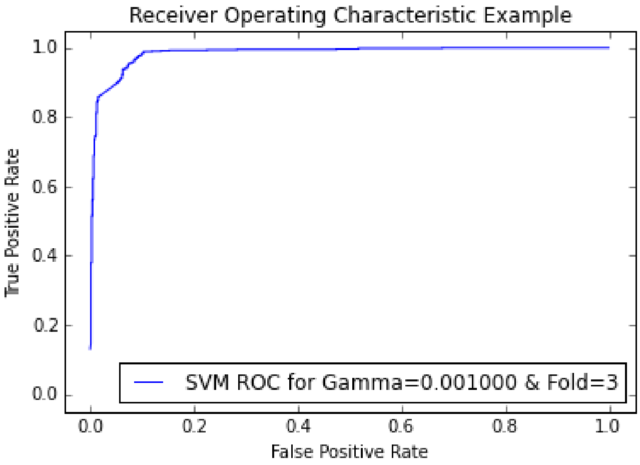
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.92      0.96      0.94       796

avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

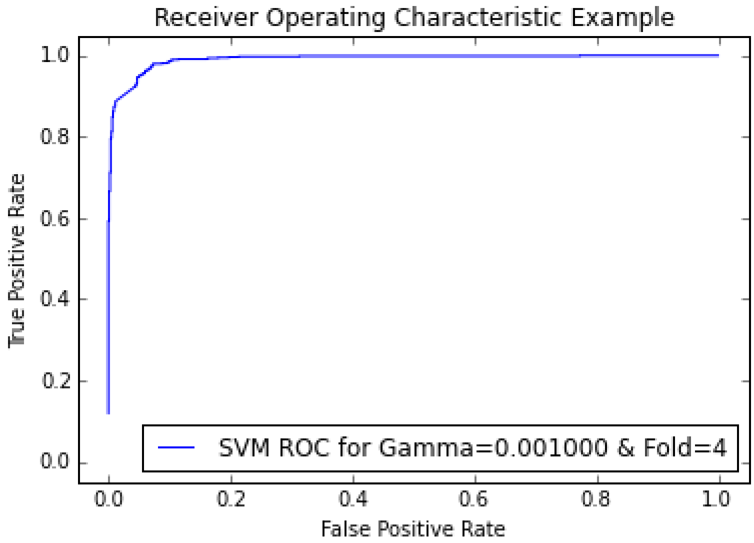
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.94      0.95       780
          1       0.94      0.96      0.95       796

avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

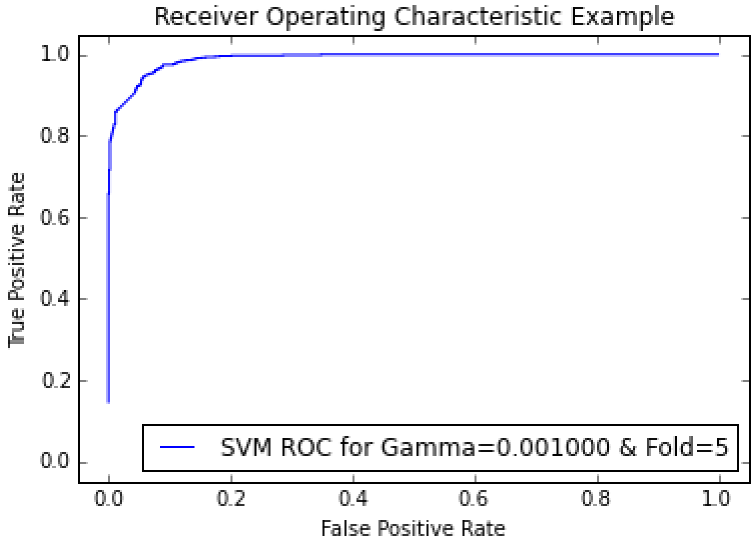
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.93      0.94       780
          1       0.93      0.96      0.94       795

avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 0.010000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

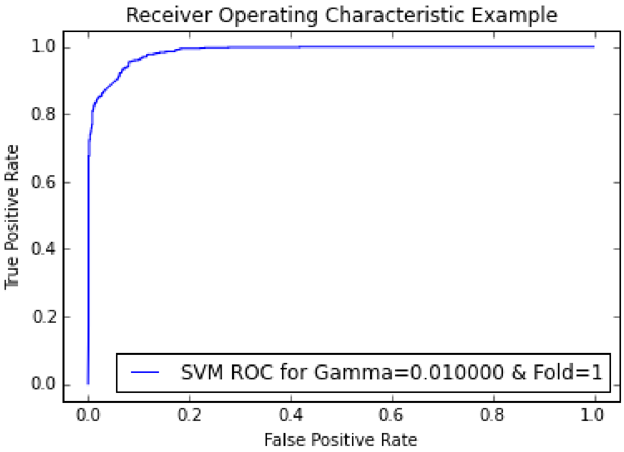
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796
avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

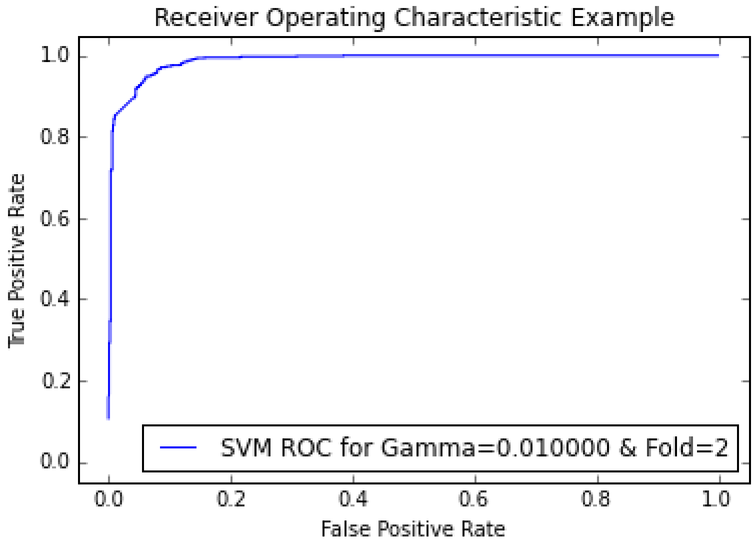
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

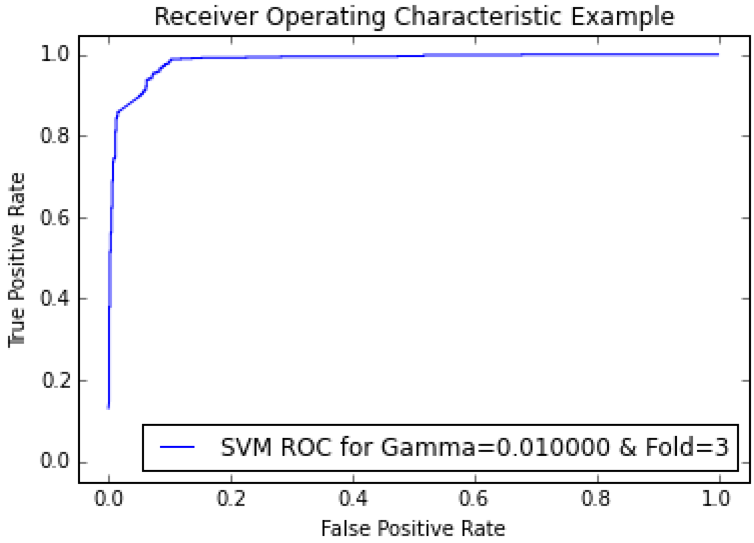
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.92      0.96      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

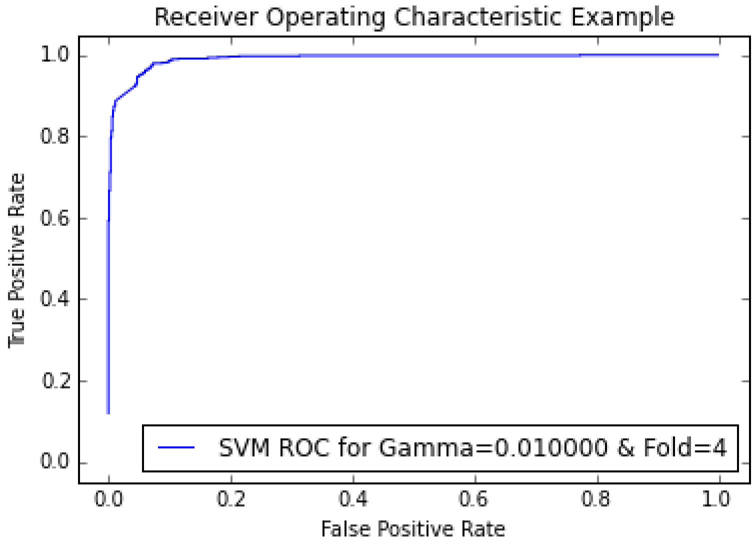
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.94      0.95       780
          1       0.94      0.96      0.95       796
avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

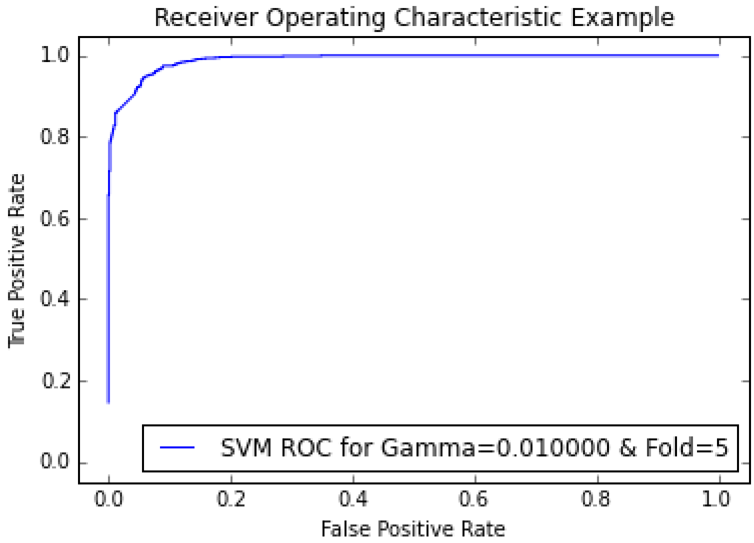
The precision and recall values are:

```
              precision    recall  f1-score   support

          0       0.96      0.93      0.94       780
          1       0.93      0.96      0.94       795
avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 0.100000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

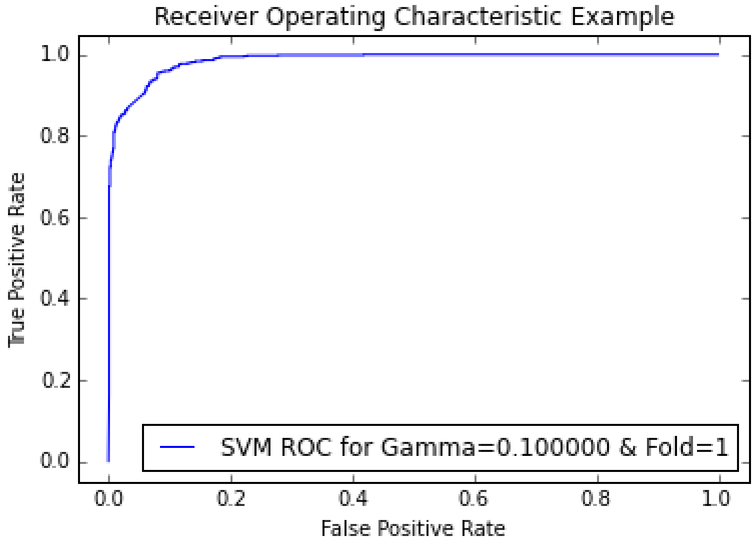
The precision and recall values are:

```
              precision    recall  f1-score   support
          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796
avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



<u>Fold Number: 2</u>

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

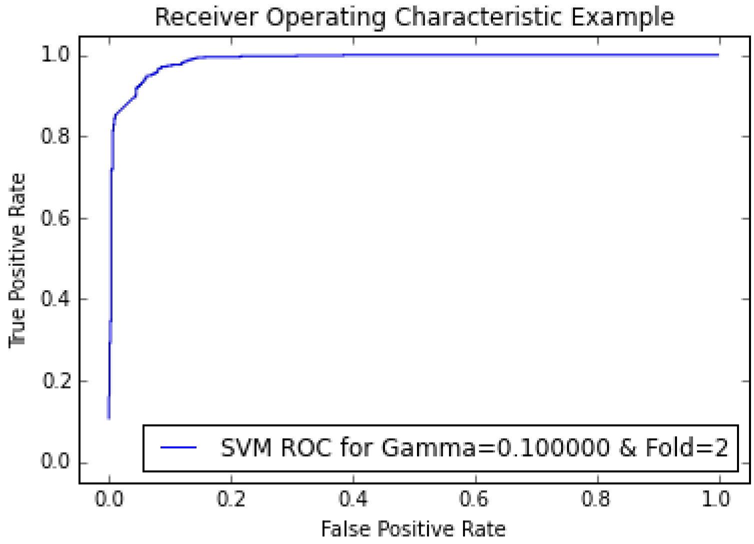
The precision and recall values are:

```
                     precision    recall  f1-score   support
                          0       0.95      0.92      0.94       781
                          1       0.93      0.95      0.94       796
           avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

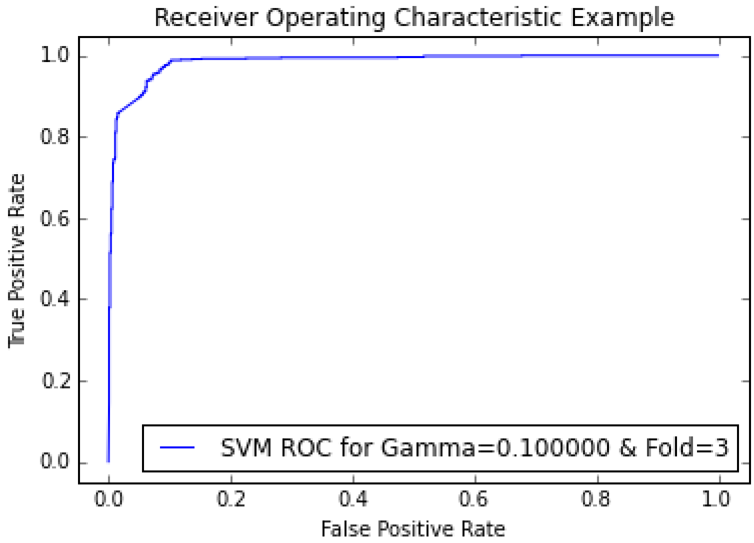
The precision and recall values are:

```
                   precision    recall  f1-score   support

                0       0.95      0.92      0.94       781
                1       0.92      0.96      0.94       796
      avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

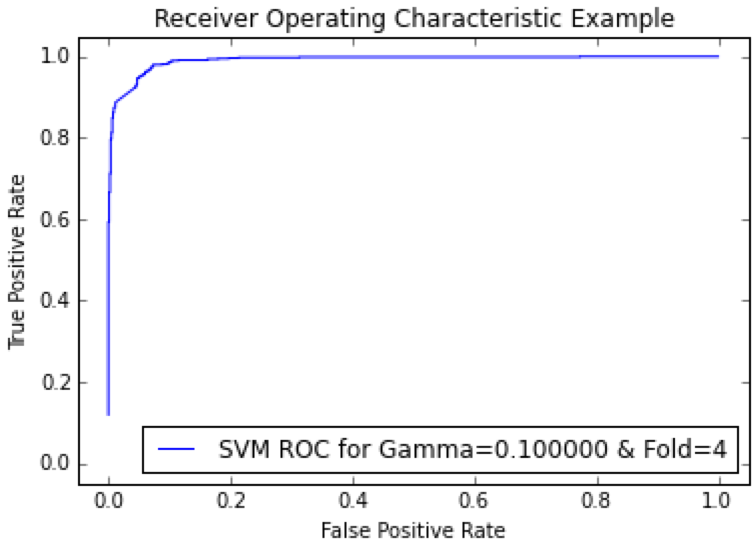
The precision and recall values are:

```
                   precision    recall  f1-score   support

                0       0.96      0.94      0.95       780
                1       0.94      0.96      0.95       796

      avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

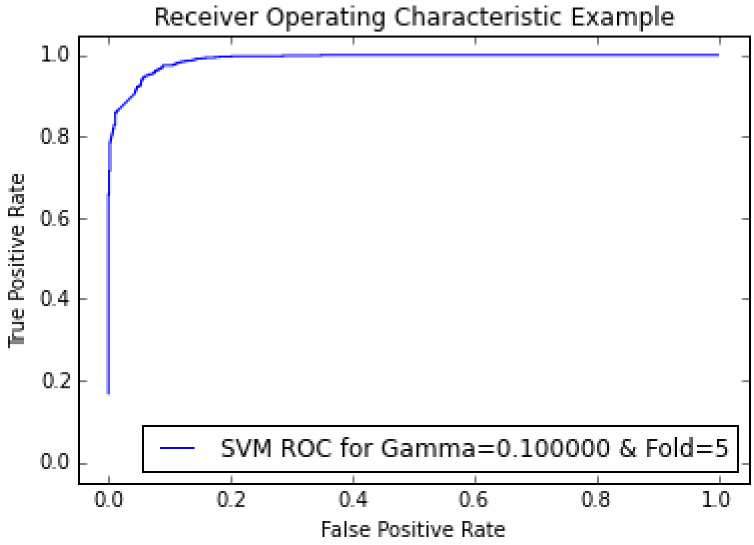
The precision and recall values are:

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.94	795
avg / total	0.94	0.94	0.94	1575

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 1.000000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

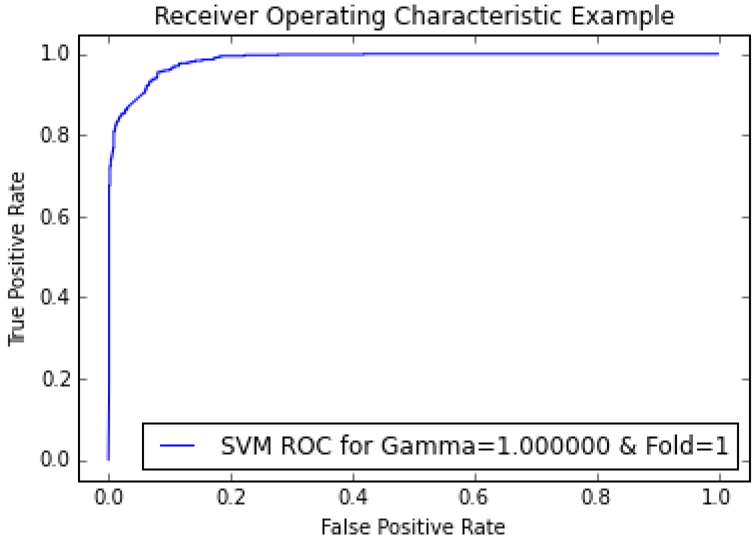
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796
avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

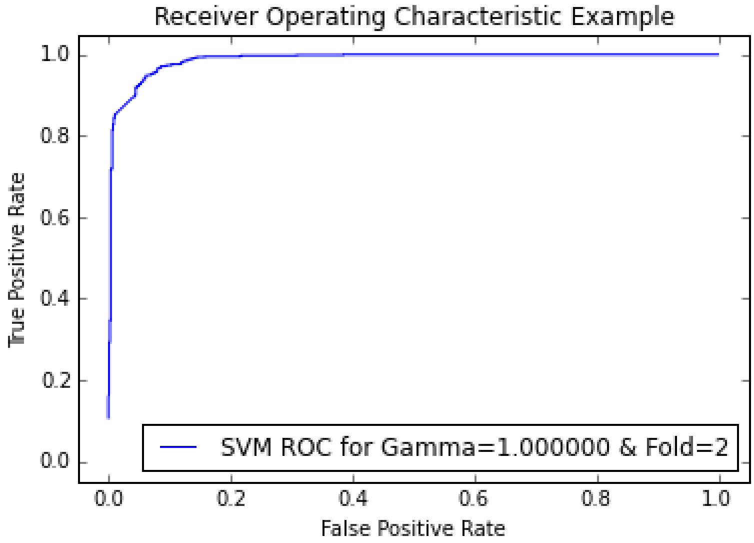
The precision and recall values are:

```
             precision    recall  f1-score   support
          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

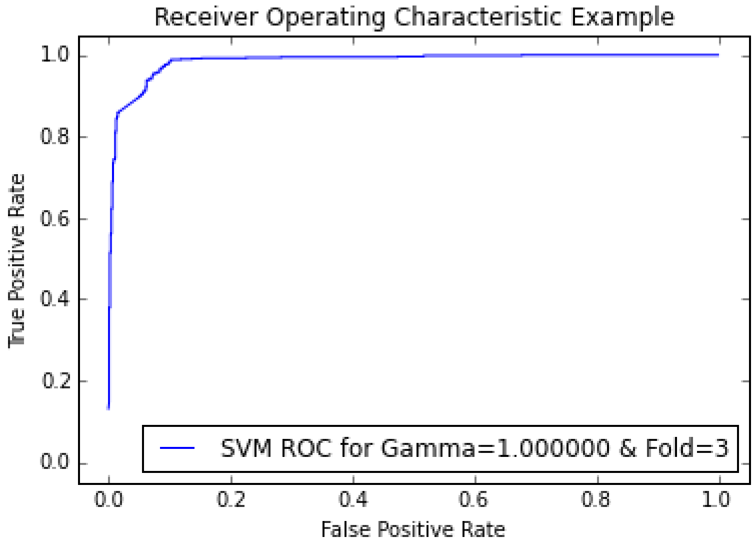
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.92      0.96      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

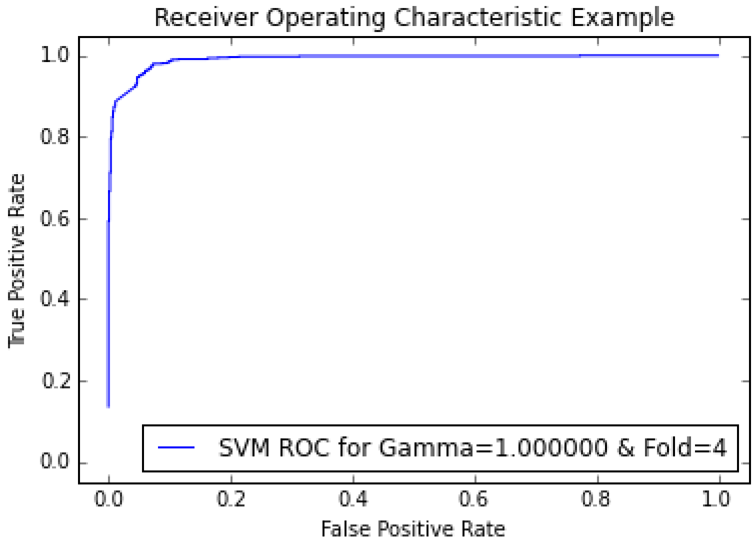
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.94      0.95       780
          1       0.94      0.96      0.95       796
avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

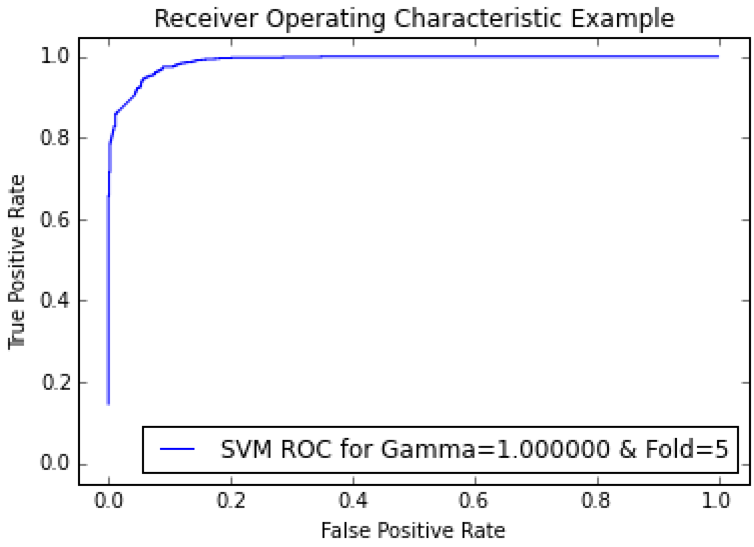
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.93      0.94       780
          1       0.93      0.96      0.94       795
avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 10.000000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

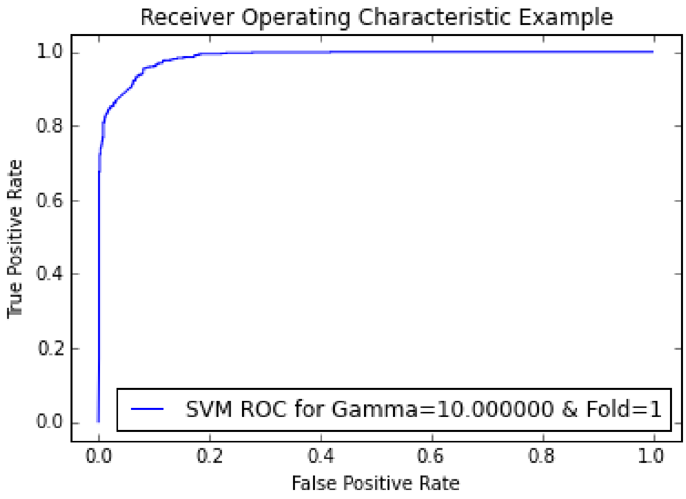
The precision and recall values are:

```
                  precision    recall  f1-score   support
                  0       0.96      0.90      0.93       781
                  1       0.91      0.96      0.93       796
        avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

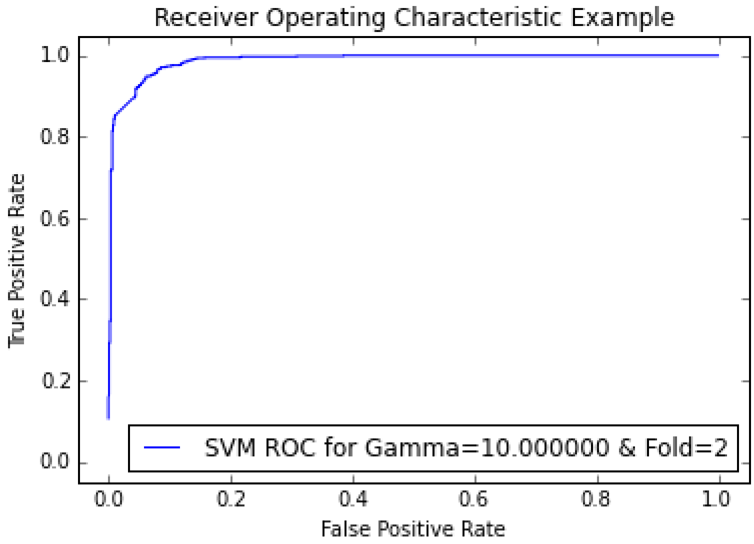
The precision and recall values are:

```
              precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

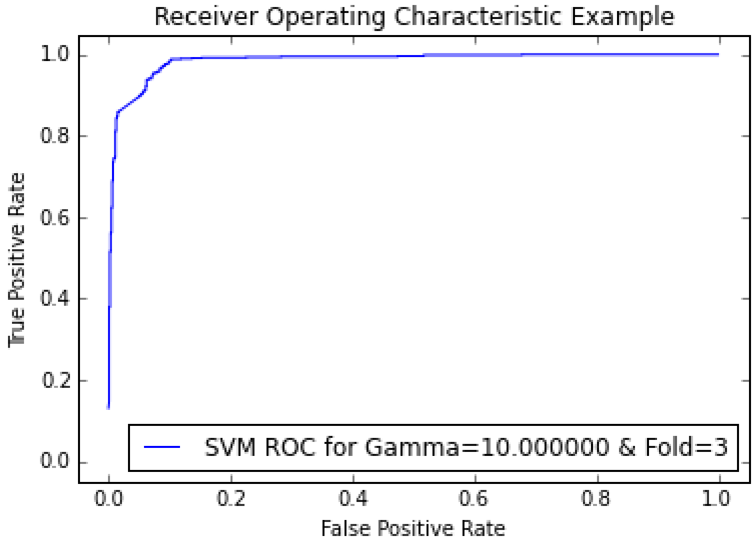
The precision and recall values are:

```
              precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.92      0.96      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

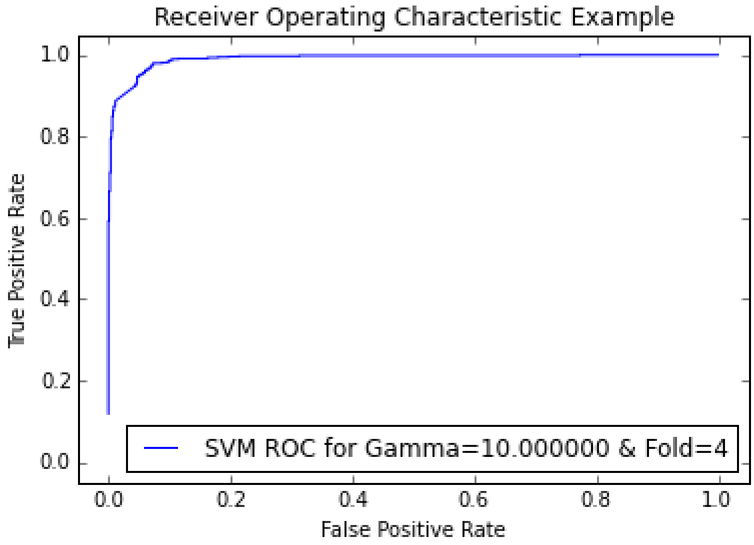
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.94      0.95       780
          1       0.94      0.96      0.95       796

avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

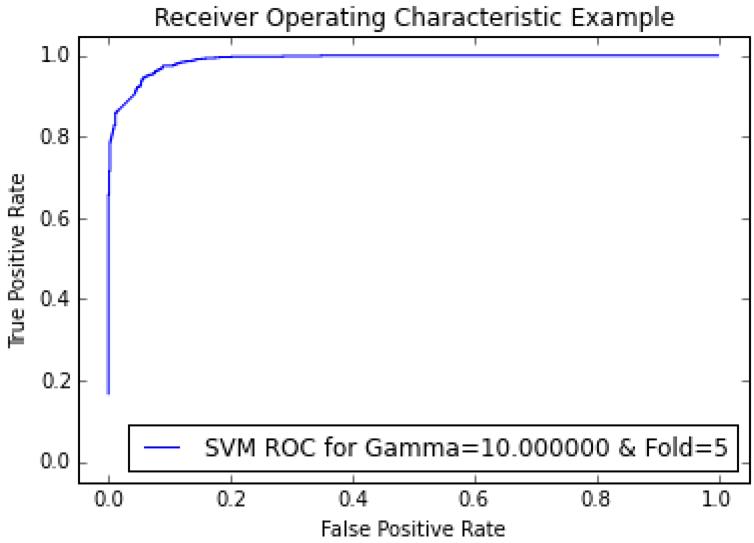
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.93      0.94       780
          1       0.93      0.96      0.94       795
avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 100.000000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

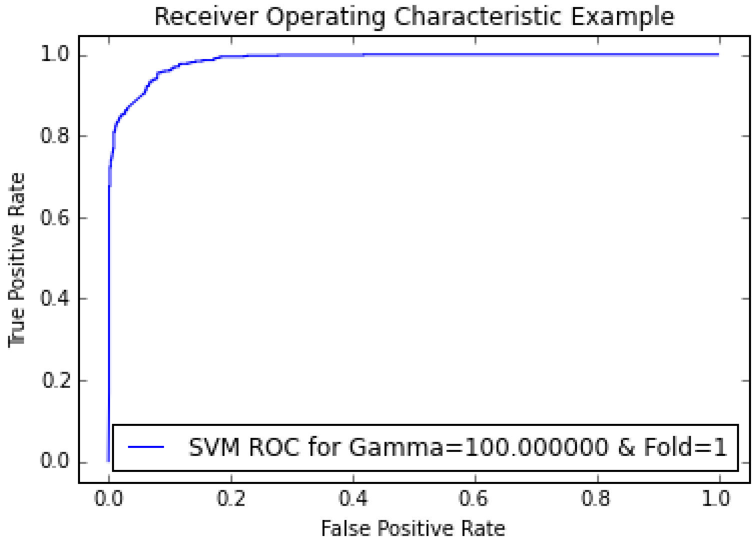
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796
avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

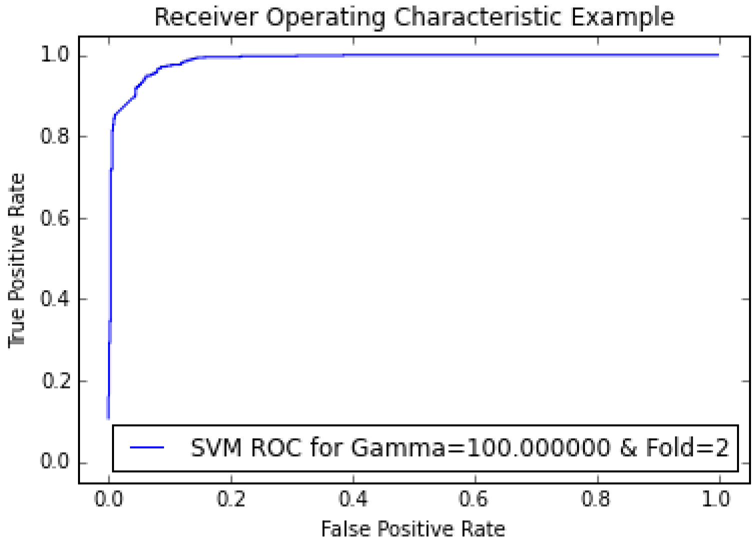
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796

avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

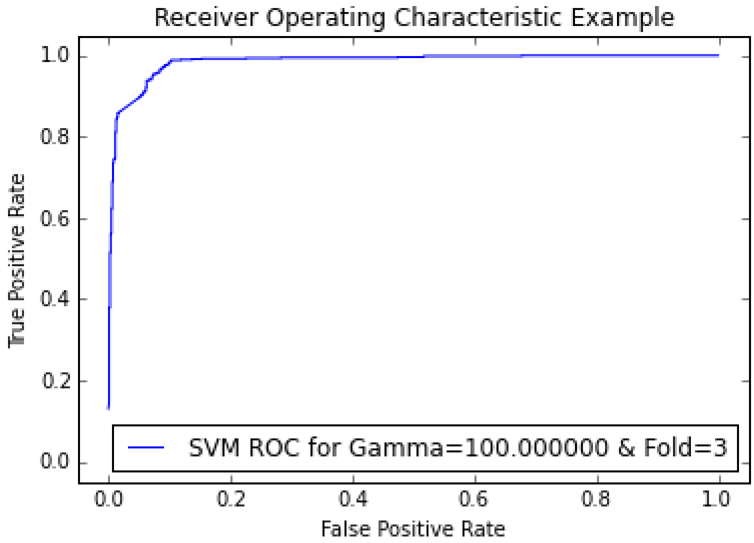
The precision and recall values are:

```
                precision    recall  f1-score   support

            0       0.95      0.92      0.94       781
            1       0.92      0.96      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

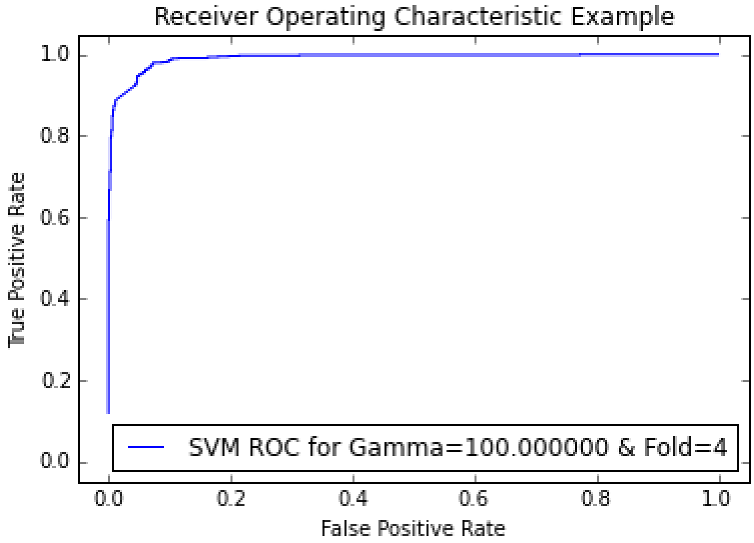
The precision and recall values are:

```
                precision    recall  f1-score   support

            0       0.96      0.94      0.95       780
            1       0.94      0.96      0.95       796

avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

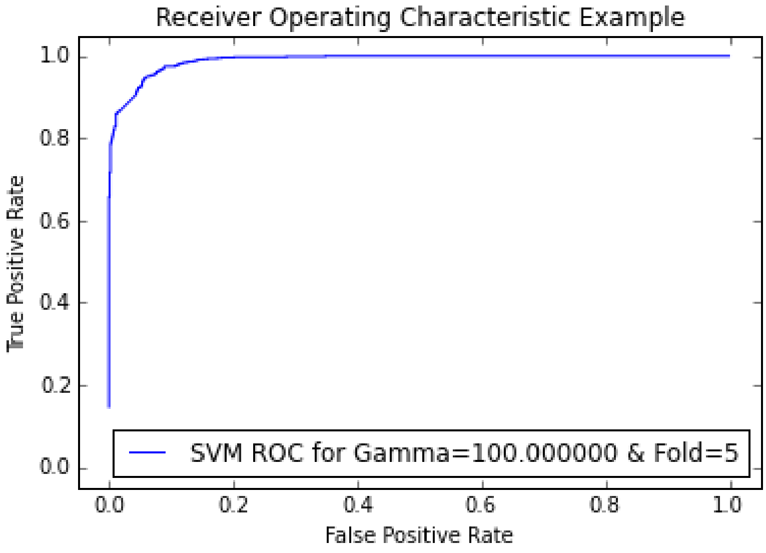
The precision and recall values are:

```
                  precision    recall  f1-score   support

            0       0.96      0.93      0.94       780
            1       0.93      0.96      0.94       795
  avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



****

The value of Gamma is 1000.000000

Fold Number: 1

The accuracy is **0.930881**

'0' is Computer Technology and '1' is Recreational Activity

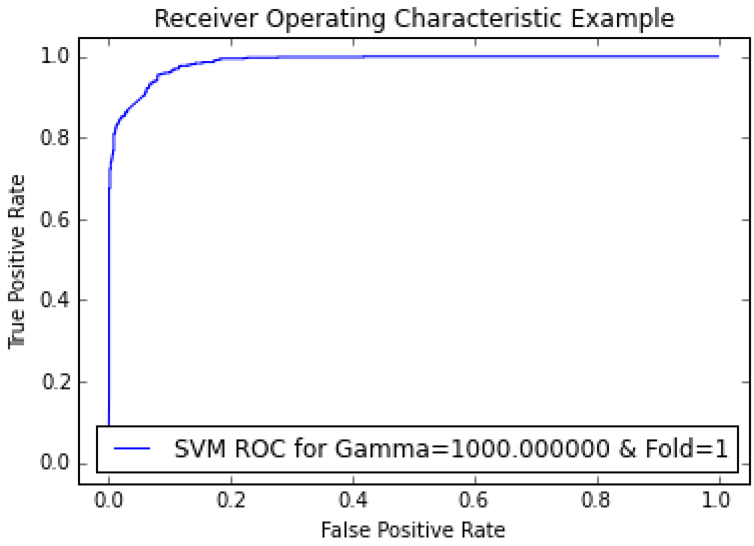
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.96      0.90      0.93       781
          1       0.91      0.96      0.93       796
avg / total       0.93      0.93      0.93      1577
```

The confusion matrix is as shown below:

```
[[702  79]
 [ 30 766]]
```

The ROC Curve is as shown below:



Fold Number: 2

The accuracy is **0.939759**

'0' is Computer Technology and '1' is Recreational Activity

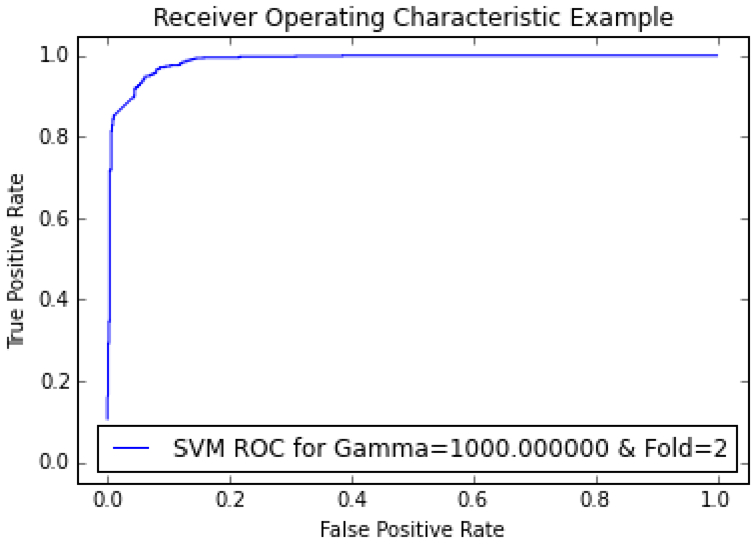
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.93      0.95      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[722  59]
 [ 36 760]]
```

The ROC Curve is as shown below:



Fold Number: 3

The accuracy is **0.937857**

'0' is Computer Technology and '1' is Recreational Activity

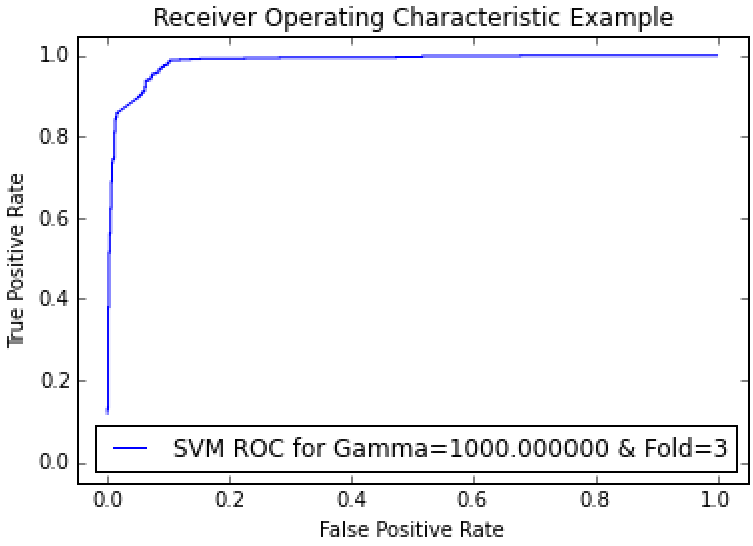
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.94       781
          1       0.92      0.96      0.94       796
avg / total       0.94      0.94      0.94      1577
```

The confusion matrix is as shown below:

```
[[718  63]
 [ 35 761]]
```

The ROC Curve is as shown below:



Fold Number: 4

The accuracy is **0.950508**

'0' is Computer Technology and '1' is Recreational Activity

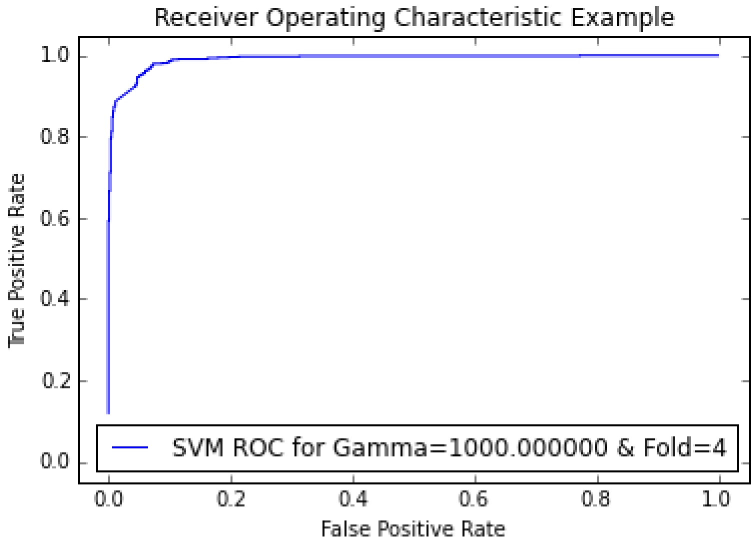
The precision and recall values are:

```
                    precision    recall  f1-score   support

                0       0.96      0.94      0.95       780
                1       0.94      0.96      0.95       796
      avg / total       0.95      0.95      0.95      1576
```

The confusion matrix is as shown below:

```
[[731  49]
 [ 29 767]]
```

The ROC Curve is as shown below:



Fold Number: 5

The accuracy is **0.941587**

'0' is Computer Technology and '1' is Recreational Activity

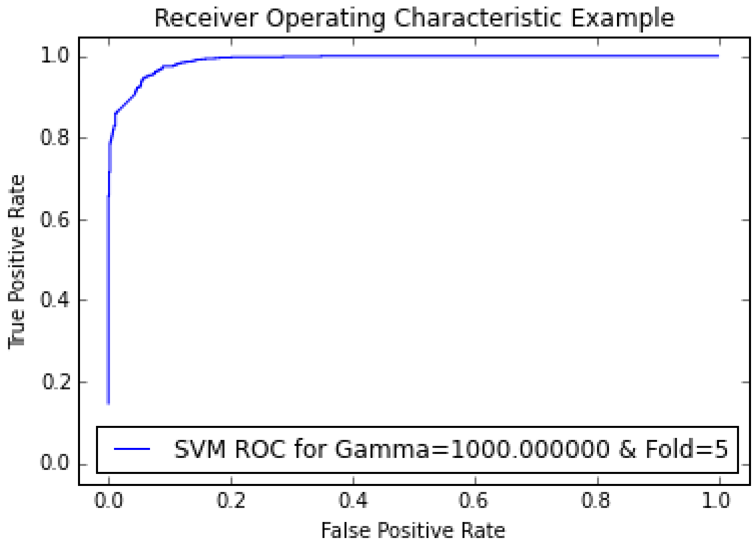
The precision and recall values are:

```
                    precision    recall  f1-score   support

                0       0.96      0.93      0.94       780
                1       0.93      0.96      0.94       795
      avg / total       0.94      0.94      0.94      1575
```

The confusion matrix is as shown below:

```
[[722  58]
 [ 34 761]]
```

The ROC Curve is as shown below:



Best Value of Gamma:
We did not observe much difference by varying the values of Gamma. The accuracies were almost in the same range. Gamma = 10 could be used as the best value.

Problem g:

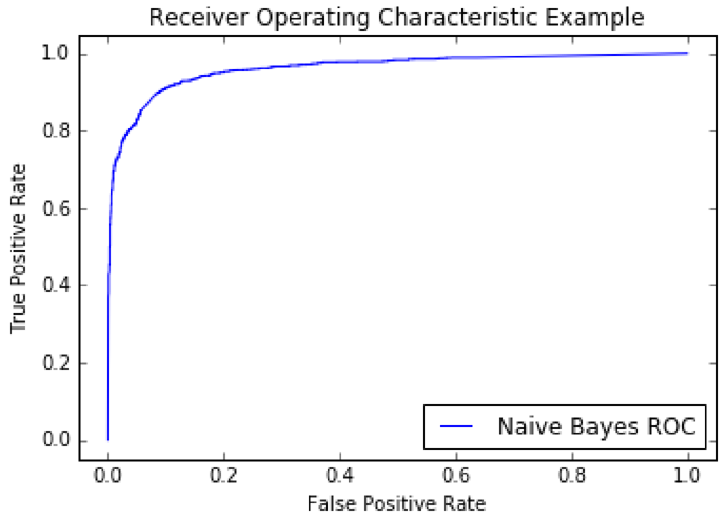
The accuracy for the **Naive Bayes** model is **0.898730**
'0' is Computer Technology and '1' is Recreational Activity
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.92      0.87      0.89      1560
          1       0.88      0.93      0.90      1590
avg / total       0.90      0.90      0.90      3150
```

The confusion matrix is as shown below:

```
[[1354  206]
 [ 113 1477]]
```

The ROC Curve is as shown below:



Problem h:

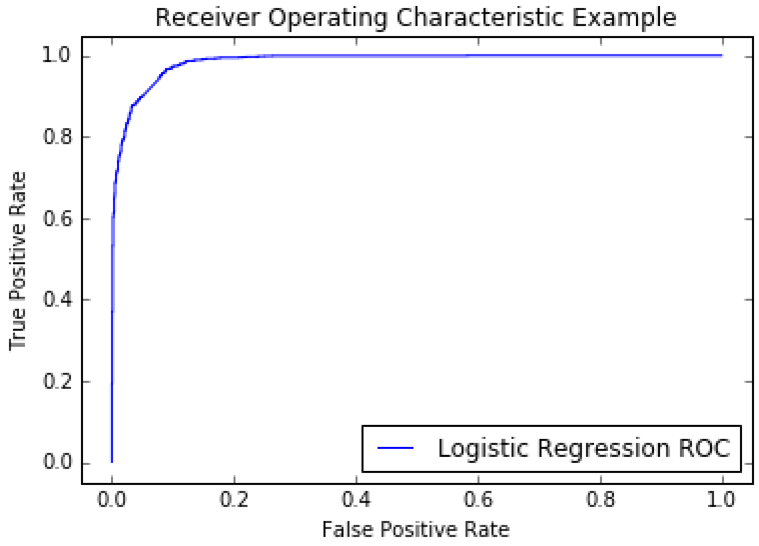
The accuracy for the **Logistic Regression** model is **0.935873**
'0' is Computer Technology and '1' is Recreational Activity
The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.95      0.92      0.93      1560
          1       0.92      0.95      0.94      1590

avg / total       0.94      0.94      0.94      3150
```

The confusion matrix is as shown below:

```
[[1430  130]
 [  72 1518]]
```

The ROC Curve is as shown below:



Problem i:

Multiclass SVM with One vs One Method:
The accuracy for the model is **0.819169**
The precision and recall values are:

```
                precision    recall  f1-score   support
                0       0.76      0.76      0.76       392
                1       0.72      0.77      0.74       385
                2       0.85      0.82      0.83       390
                3       0.95      0.93      0.94       398
      avg / total       0.82      0.82      0.82      1565
```

The confusion matrix is as shown below:

```
[[297  67  24   4]
 [ 58 295  26   6]
 [ 33  30 318   9]
 [  2  17   7 372]]
```

Multiclass SVM with One vs Rest Method:

The accuracy for the model is **0.821725**
The precision and recall values are:

```
                precision    recall  f1-score   support

                0       0.75      0.77      0.76       392
                1       0.72      0.76      0.74       385
                2       0.86      0.82      0.84       390
                3       0.97      0.93      0.95       398
      avg / total       0.83      0.82      0.82      1565
```

The confusion matrix is as shown below:

```
[[302  65  22   3]
 [ 63 294  24   4]
 [ 34  31 319   6]
 [  3  17   7 371]]
```

Multiclass Gaussian Naive Bayes with One vs Rest Method:
The accuracy for the model is **0.742492**
The precision and recall values are:

```
                precision    recall  f1-score   support
                0       0.61      0.74      0.67       392
                1       0.70      0.59      0.64       385
                2       0.79      0.71      0.75       390
                3       0.89      0.92      0.91       398
      avg / total       0.75      0.74      0.74      1565
```

The confusion matrix is as shown below:

```
[[290  65  27  10]
 [105 227  33  20]
 [ 68  30 277  15]
 [ 11   4  15 368]]
```

Multiclass Gaussian Naive Bayes with One vs One Method:

The accuracy for the model is **0.742492**

The precision and recall values are:

```
             precision    recall  f1-score   support

          0       0.61      0.73      0.66       392
          1       0.68      0.62      0.65       385
          2       0.81      0.70      0.75       390
          3       0.89      0.91      0.90       398
avg / total       0.75      0.74      0.74      1565
```

The confusion matrix is as shown below:

```
[[285  74  23  10]
 [100 239  27  19]
 [ 72  30 274  14]
 [ 12   7  15 364]]
```