EE219 – Project 2

Team Members :-

Omkar Patil : - 904760474
Shikhar Malhotra :- 504741656
Twinkle Gupta :- 804740325

Solution a :-

- Here a histogram of the number of documents per topic was to be plotted
- The number of documents in the groups Computer Technology and Recreational Activity is to be reported

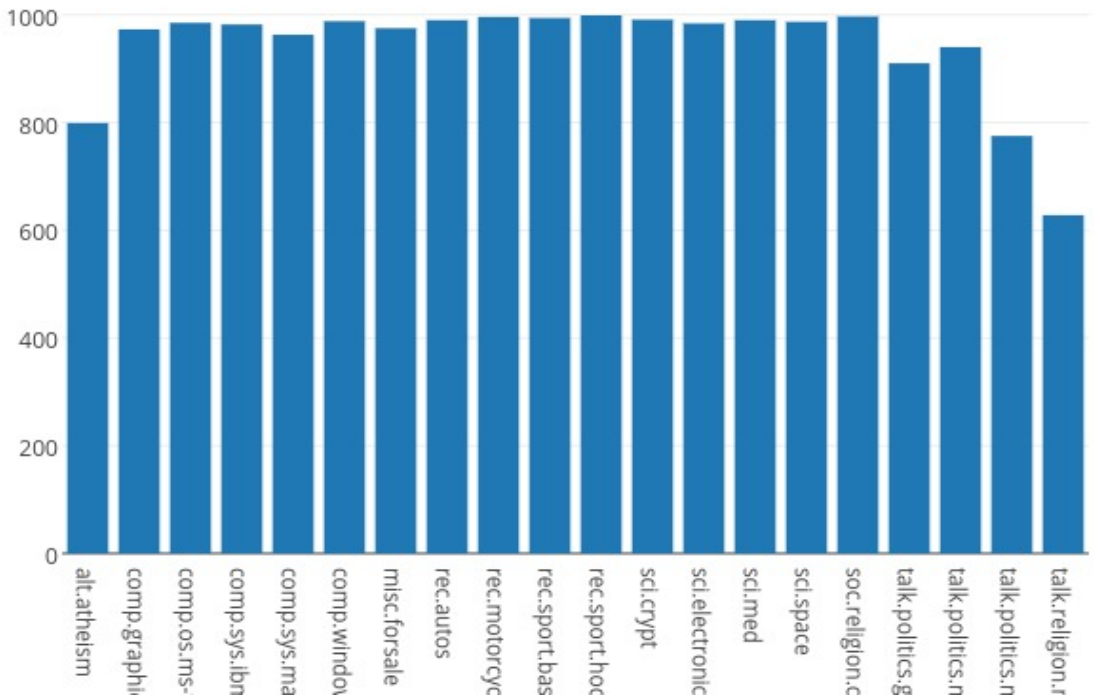


Fig 1 : -Histogram of Document Topics vs counts

Count of Computer Technology documents :- 4891
Count of Recreational Activity documents :- 3979

Solution b :-

- Here the documents in the data set are turned into numerical feature vectors.
- The words here are tokenized and unwanted tokens like punctuation marks, stop words, stemming words etc are removed for preprocessing the data .
- After this a TFxIDF vector representation is created and its terms are reported as

  Total number of terms are 63579

- In order to quantify significance of a word to a class/document a TFxICF measure is also created.
- The functionalities like cleaning, TFxIDF, TFxICF will be further used for other problems too

Solution c :-

Here by using the TFxICF measure created earlier 10 most significant terms from the mentioned classes are reported

Ten most significant terms for comp.sys.ibm.pc.hardware are :-
[(0.12149574483856736, 'bio'),
(0.13712650270852514, 'bus'),
(0.14769563274272957, 'control'),
(0.15166786099351987, 'ide'),
(0.15239173925032526, 'use'),
(0.1718255878443864, 'disk'),
(0.17865325942396079, 'mb'),
(0.18681748990415775, 'card'),
(0.32129500726798521, 'scsi'),
(0.36325877054986017, 'drive')]

Ten most significant terms for comp.sys.mac.hardware are :-
[(0.12213673102076499, 'monitor'),
(0.1268370372355741, 'simm'),
(0.13127662463471818, 'card'),
(0.1324889539758646, 'problem'),
(0.13808162967734547, 'scsi'),
(0.14979846692397503, 'use'),
(0.15115400811511046, 'mb'),
(0.19033930796619764, 'drive'),
(0.24479863157079601, 'appl'),
(0.3289097327283152, 'mac')]

Ten most significant terms for misc.forsale are :-
[(0.12156642668624081, 'game'),
(0.12793763445113682, 'condit'),
(0.1353640849873293, 'sell'),
(0.15816458061820357, 'price'),
(0.16161696538350262, 'includ'),
(0.16384702456785963, 'new'),
(0.17592755237036209, 'ship'),
(0.18519413558156761, 'dos'),
(0.20221883110912328, 'offer'),
(0.21380968731764752, 'sale')]

Ten most significant terms for soc.religion.christian are :-
[(0.1143105581204833, 'sin'),
(0.11829119184434948, 'peopl'),
(0.1221851759132737, 'say'),
(0.12767669389430575, 'believ'),
(0.12928491301481848, 'faith'),
(0.13194369309367127, 'christ'),
(0.17529890255649985, 'church'),
(0.20364218572149215, 'jesus'),
(0.2356363057102632, 'christian'),
(0.40424538051168046, 'god')]

Solution d :-

- After above solutions we get the TFxIDF vectors.
- The dimensionality of the TFxIDF vectors is huge.
- High dimensionality reduces the performance of many algorithms; hence the dimension of the data should be reduced.
- This is done via selecting a subset of the features, which are relevant to a given measure and which can be transformed further to a lower dimension space.
- Here Latent Semantic Indexing LSI, a dimension reducing transform is used.
- Here we represent the data in the term-document matrix, whose columns corresponds to TFxIDF representation of the documents.
- LSI is very much like Principal Component Analysis.
- Let D denote the $t \times d$ term-document matrix with rank r where each of the d columns represents a document vector of dimension t.

This functionality is further used for following solutions

```
After applying LSI normalized txidf matrix size :- 50
```

Solution e :-

- Linear Support Vector Machines are efficient when we want to deal with sparse high dimensional datasets, thus textual data.
- They have good accuracy and fairly good complexity too.
- Here basically they learn the vector of feature weights, $\hat{w}$, and the intercept, b, given the training data set.
- Once the weights are learned, data points can be classified with the equation $WT.\vec{x}+b$, with $\vec{x}$, being the vector representation of the new document to classify.
- By thresholding with zero and with the help of sign of the the equation $WT.\vec{x}+b$, one can classify documents fairly accurately
- Here we use SVM method to separate the documents into Computer Technology vs Recreational Activity groups.
- Here we also report the ROC curve(to understand tradeoff between false positive and true positive rate at various thresholds), confusion matrix, accuracy and precision.

****
Current model Accuracy :- 0.93873015873
'0' is Computer Technology and '1' is Recreational Activity
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.95	0.92	0.94	1560
1	0.93	0.95	0.94	1590
avg / total	0.94	0.94	0.94	3150

Current confusion matrix :-
[[1439 121]
[72 1518]]

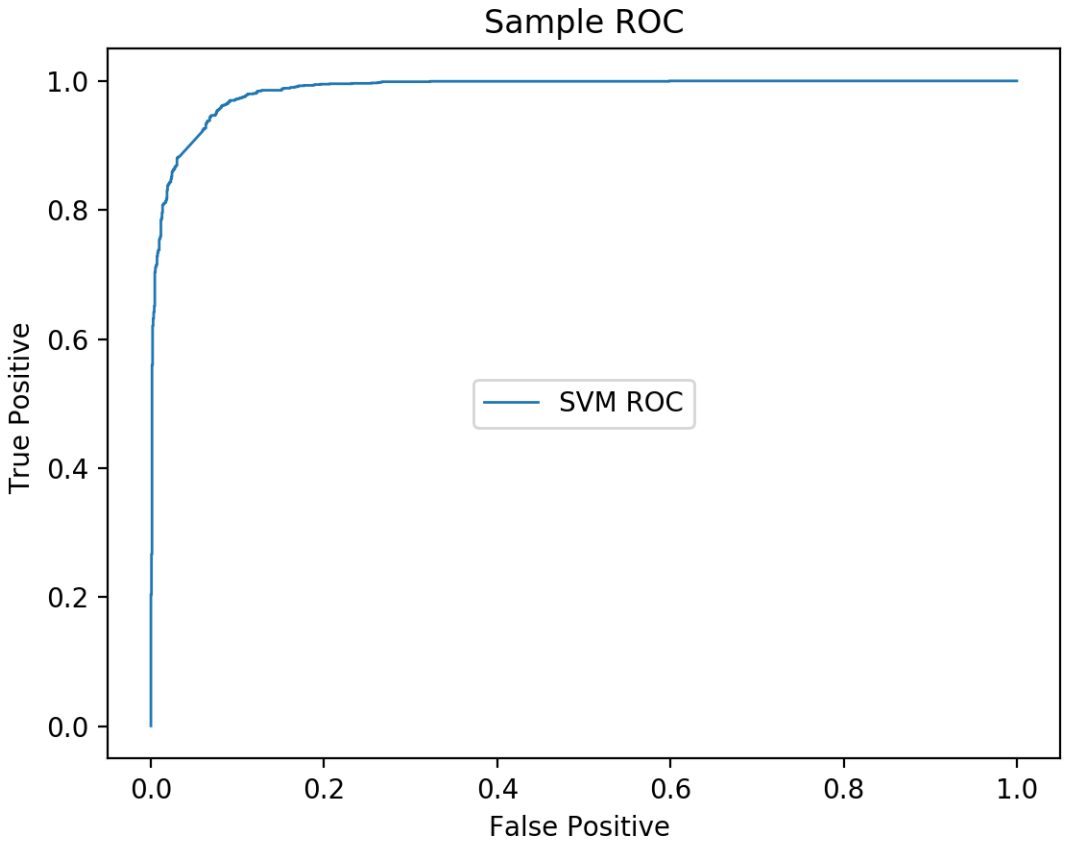


Fig 2 : -True positive vs False Positive

Solution f:-

- Here a Soft margin SVM is used for classification.
- Here we understand tradeoff between minimizing training data error function and maximizing margin between classified classes.
- The tradeoff parameter γ helps us understand importance of each parameter.
- We vary it between 0.001 and 1000 increasing in power of 10 to decide the best γ.

****
The value of Gamma is 0.001
****

Fold Number: 1

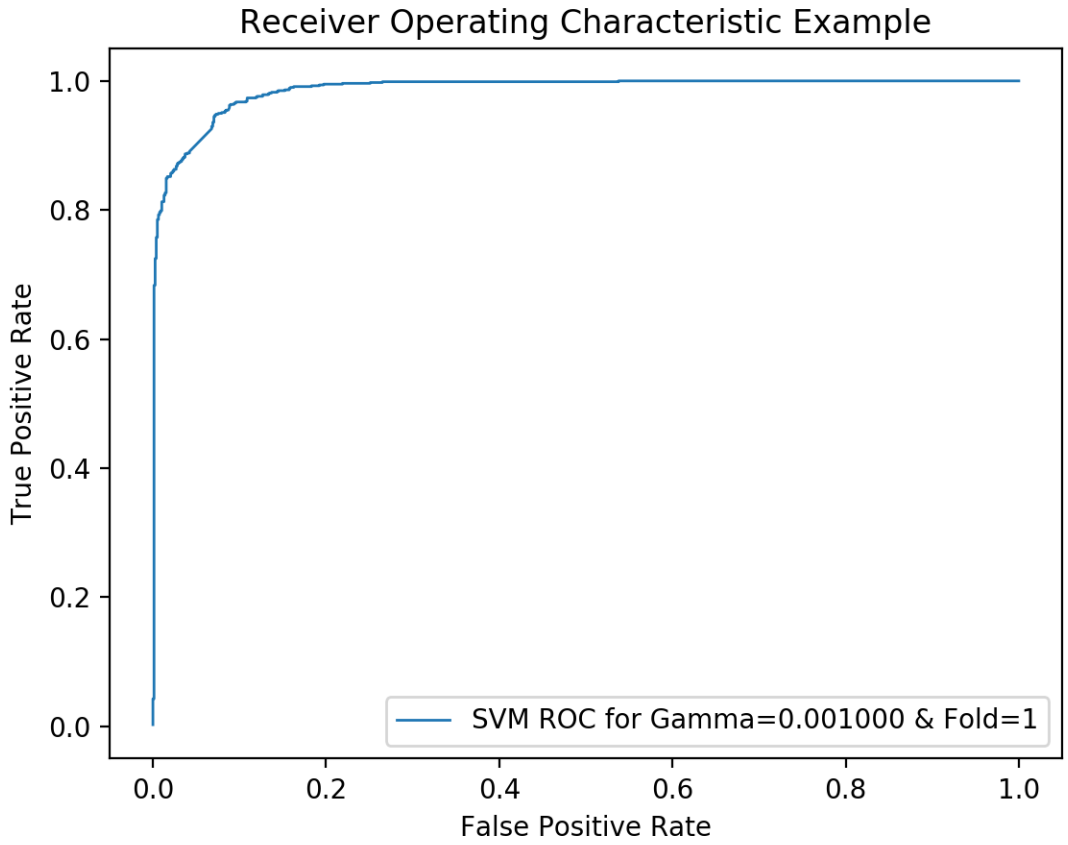
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
[29 767]]
Current ROC curve



##
##################

Fold Number: 2

Current model Accuracy 0.944831959417

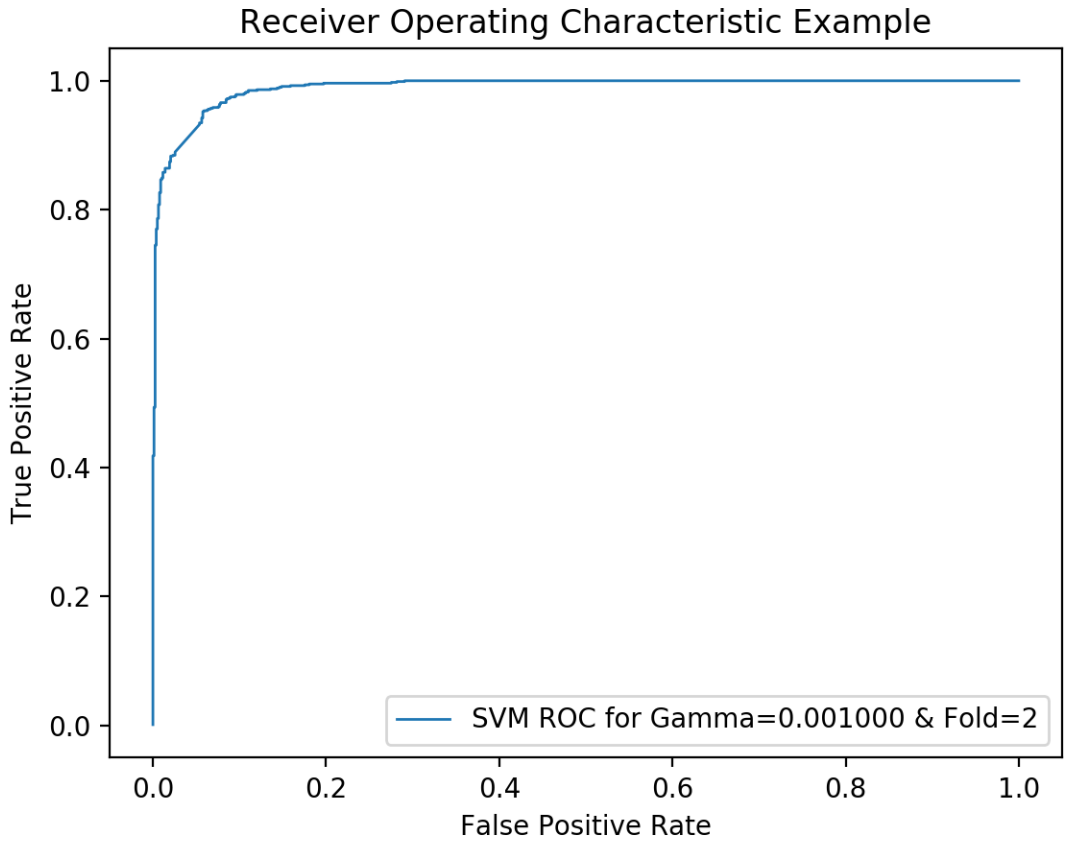
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-

[[727 54]
 [33 763]]
Current ROC curve



##
##################

Fold Number: 3

Current model Accuracy 0.940393151554

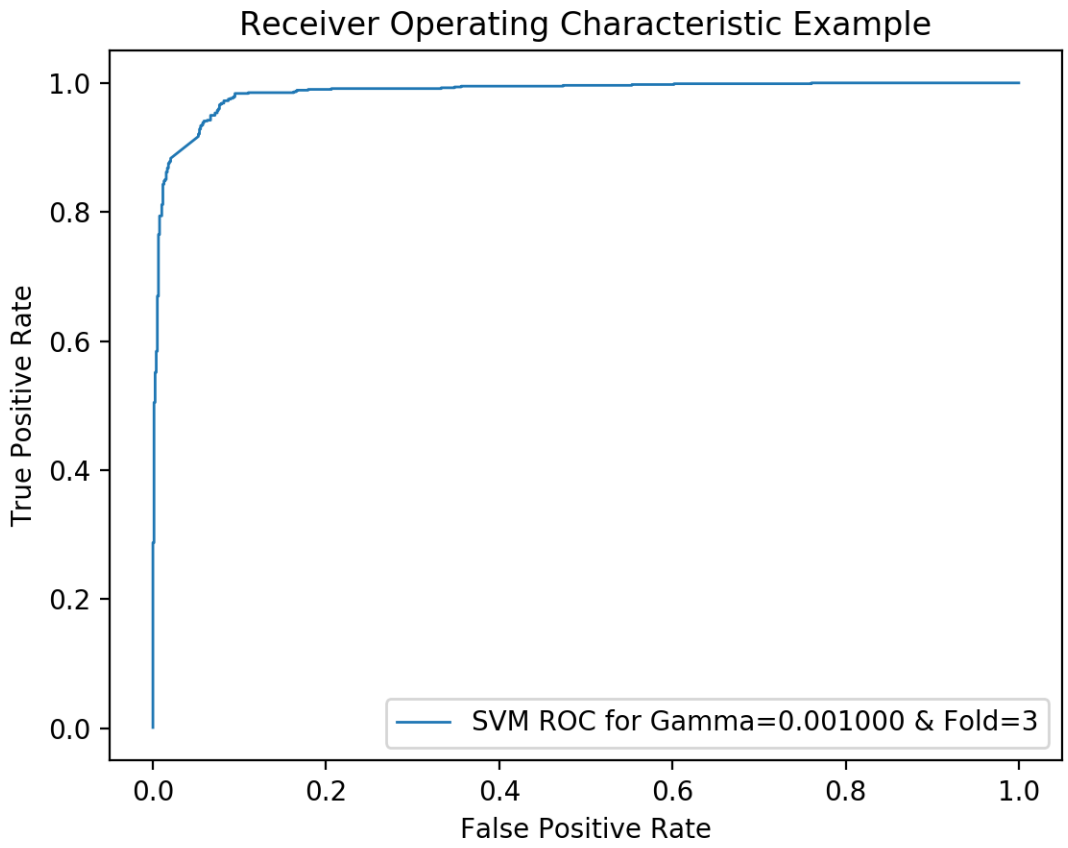
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support are :-

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

avg / total 0.94 0.94 0.94 1577

Current confusion matrix :-
[[723 58]
[36 760]]
Current ROC curve



##
##################

Fold Number: 4

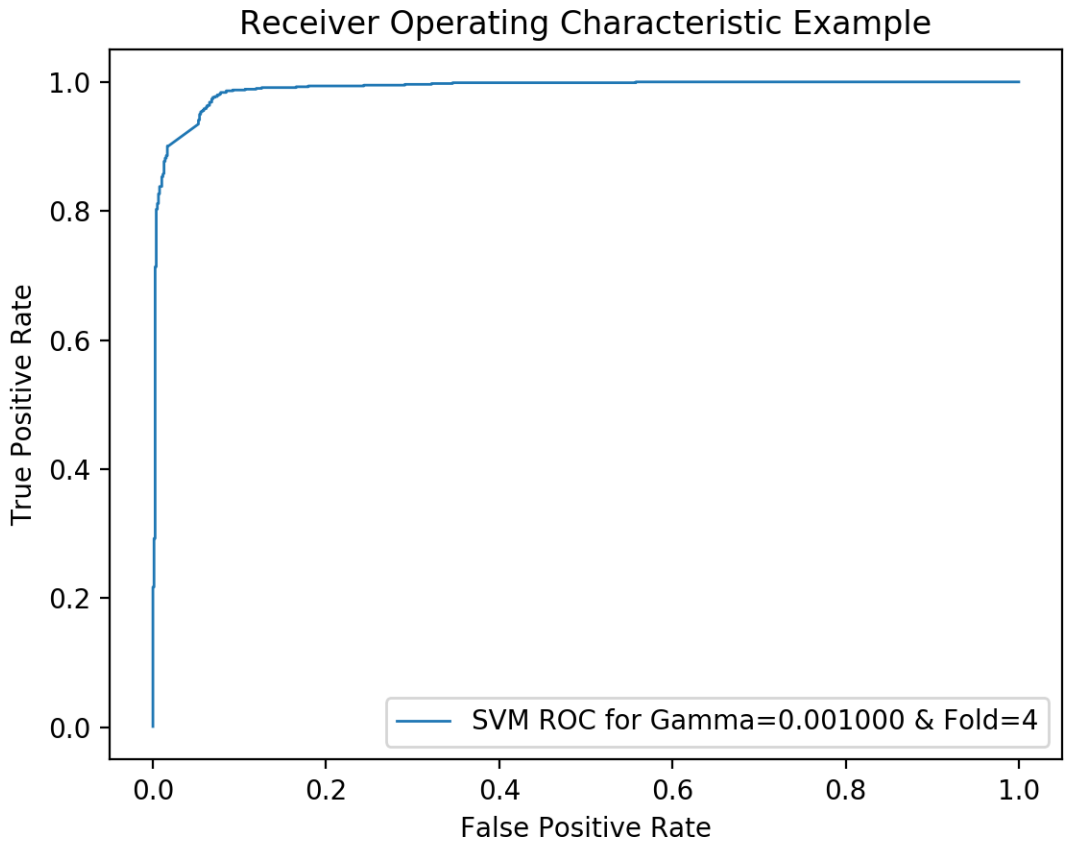
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
 [36 760]]
Current ROC curve



##
##################

Fold Number: 5

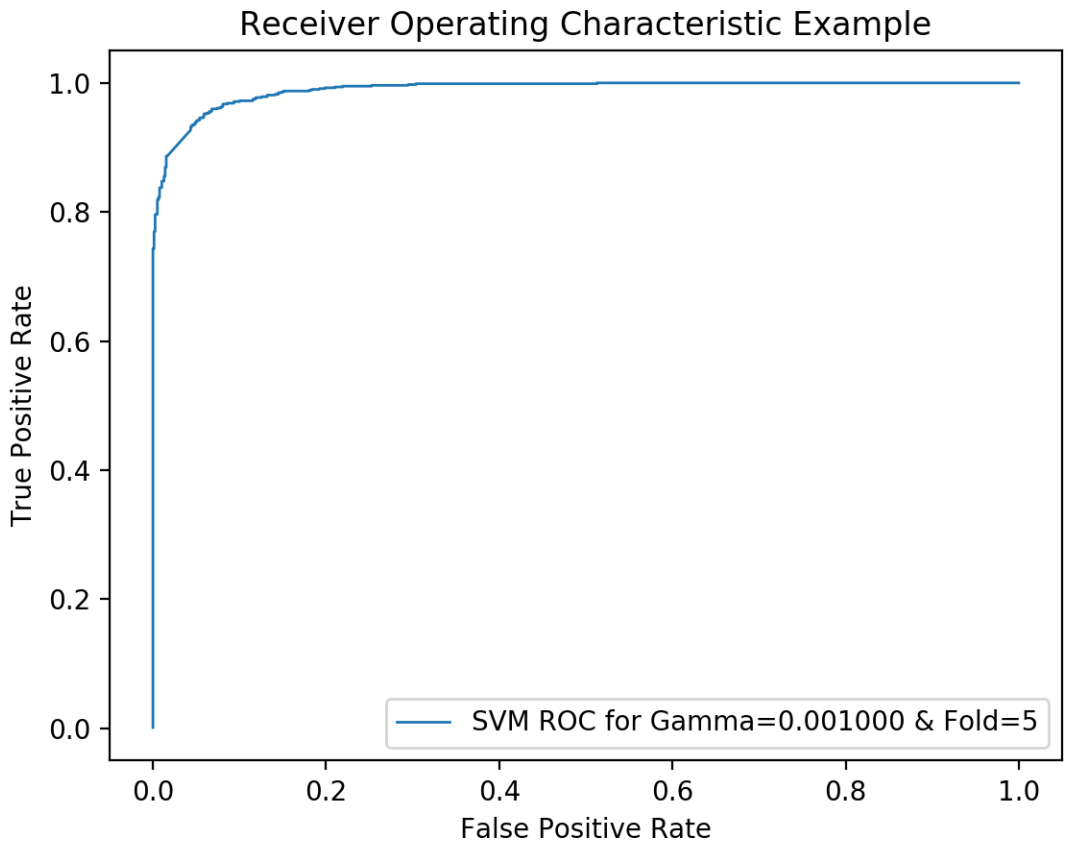
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
 [33 762]]
Current ROC curve



****
The value of Gamma is 0.01
****

Fold Number: 1

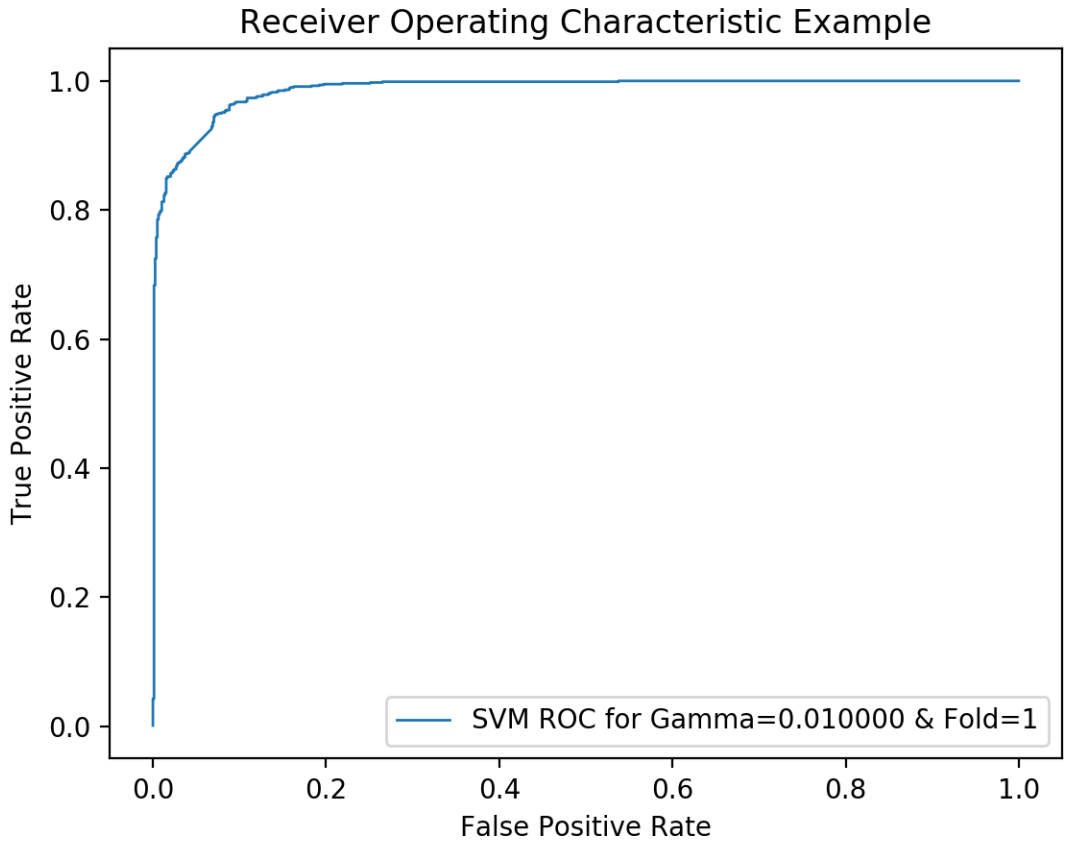
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
 [[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

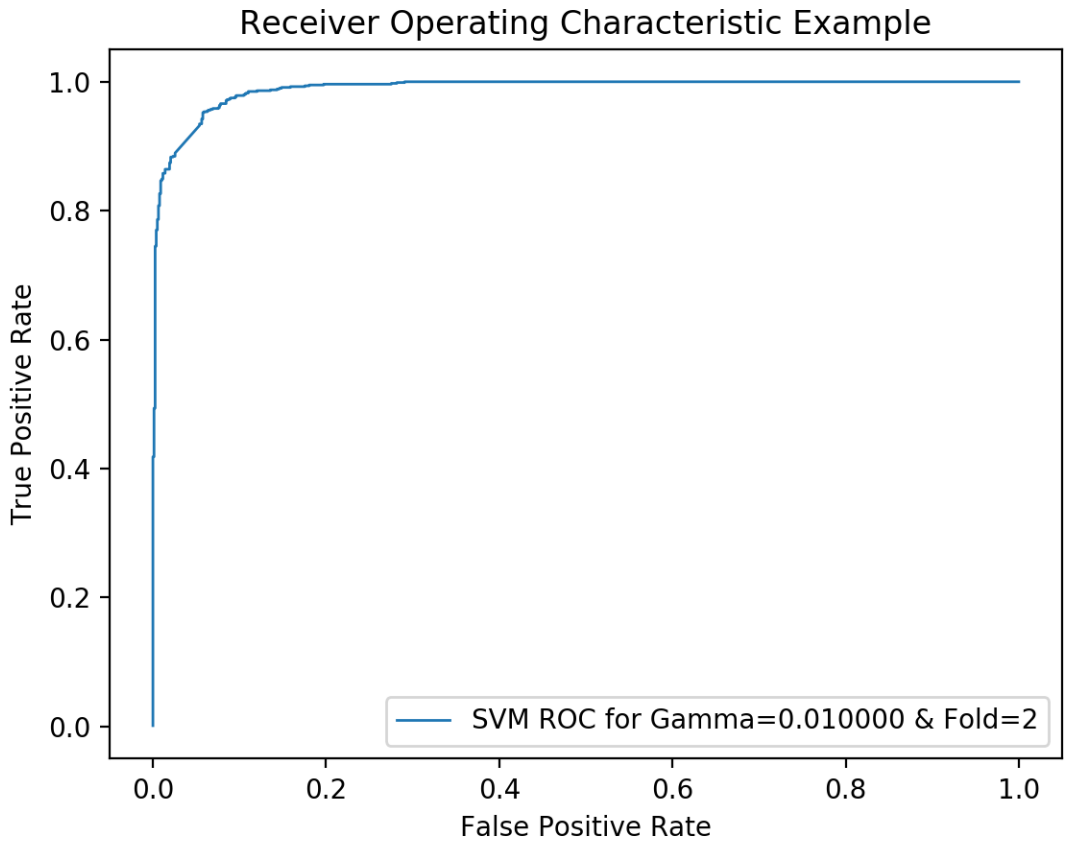
Current model Accuracy 0.944831959417

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-
[[727 54]
 [33 763]]
Current ROC curve



##
##################

Fold Number: 3

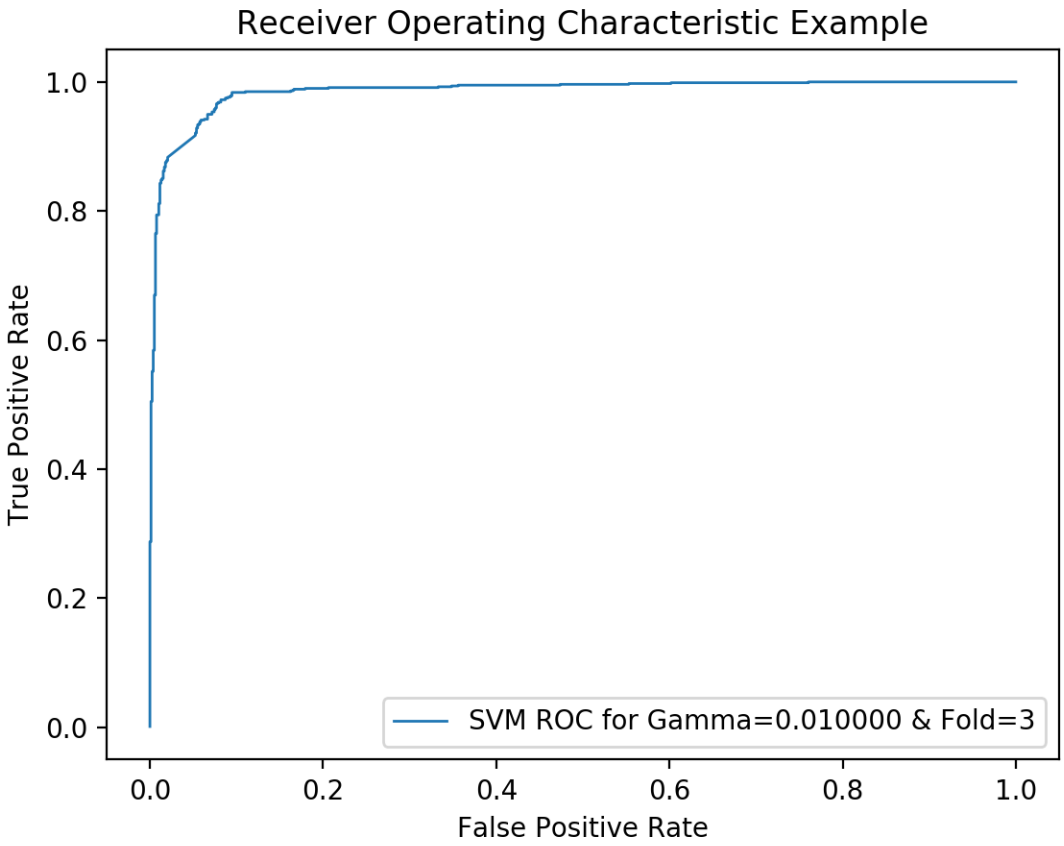
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
 [36 760]]
Current ROC curve



##
##################

Fold Number: 4

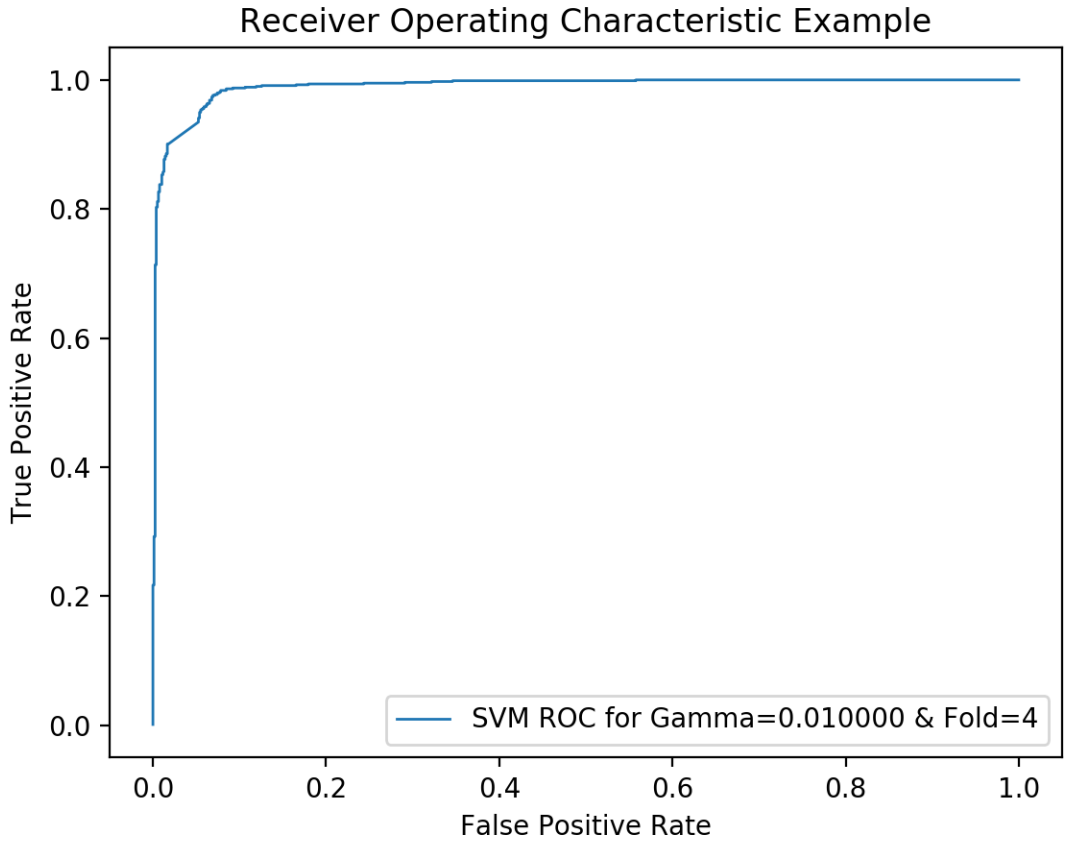
Current model Accuracy 0.94923857868

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

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[731 49]
 [31 765]]
Current ROC curve



##
##################

Fold Number: 5

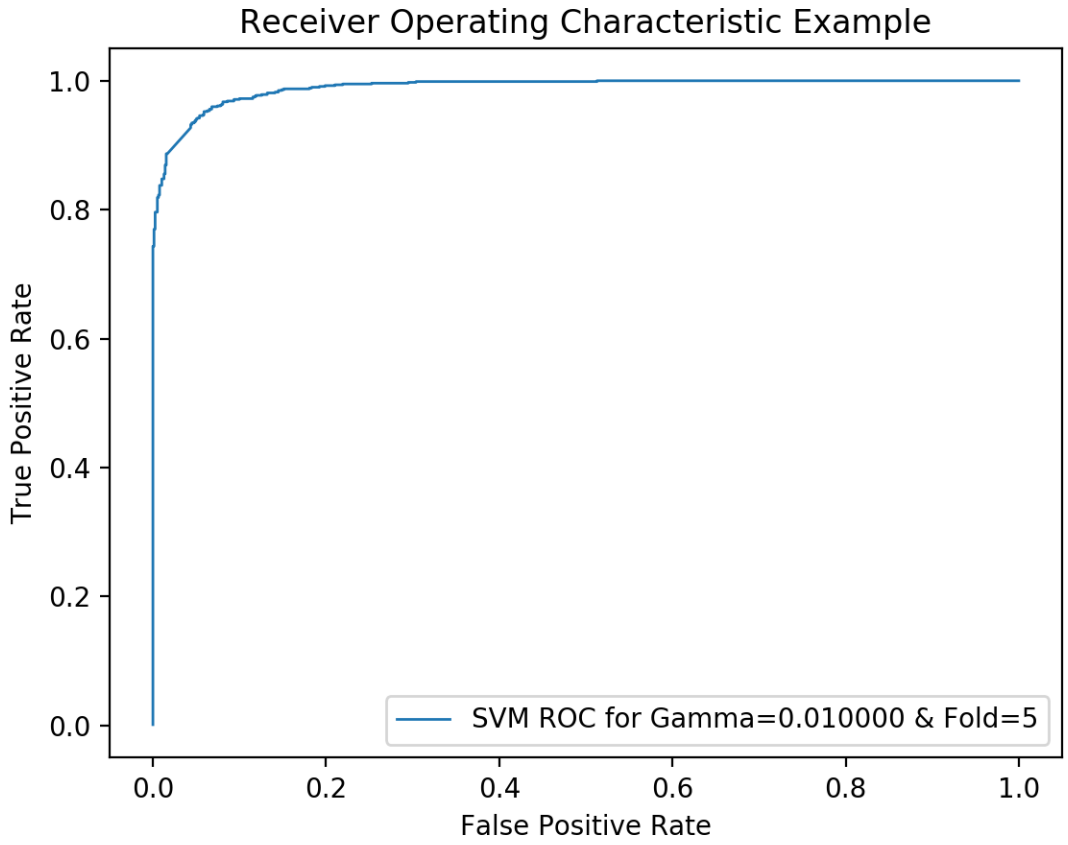
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
 [33 762]]
Current ROC curve



##
##################

****
The value of Gamma is 0.1
****

Fold Number: 1

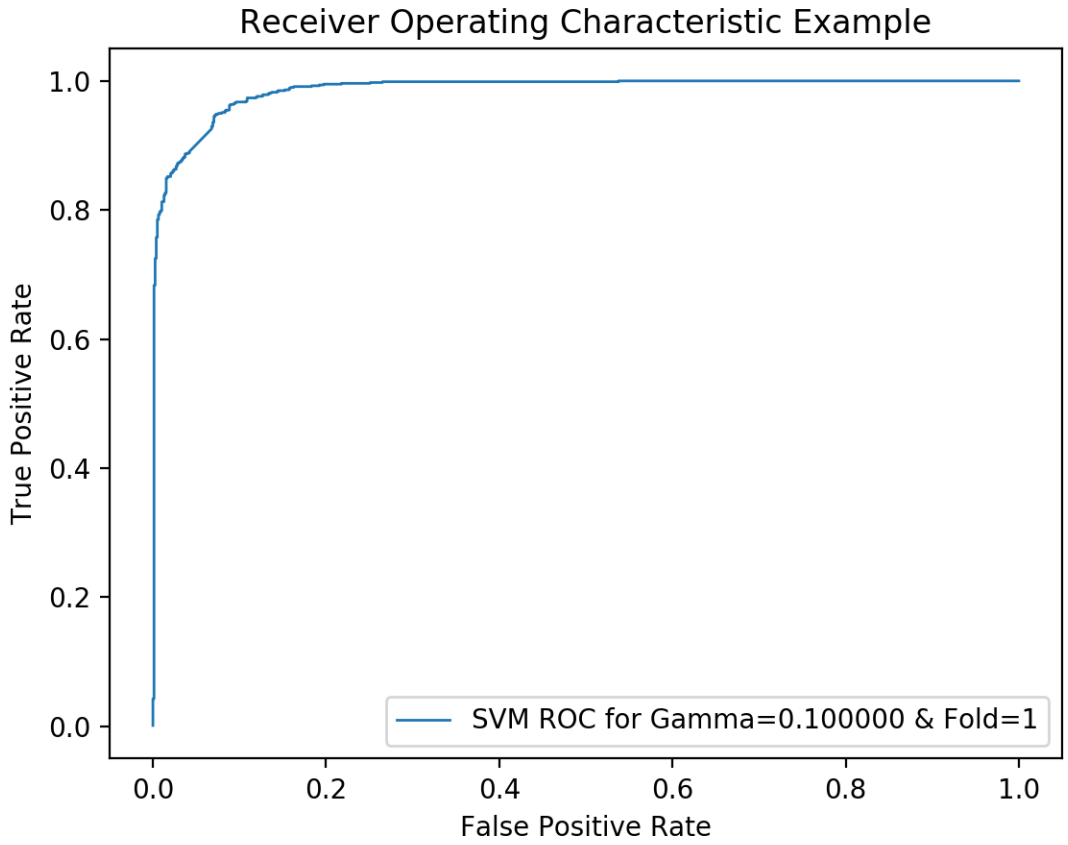
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

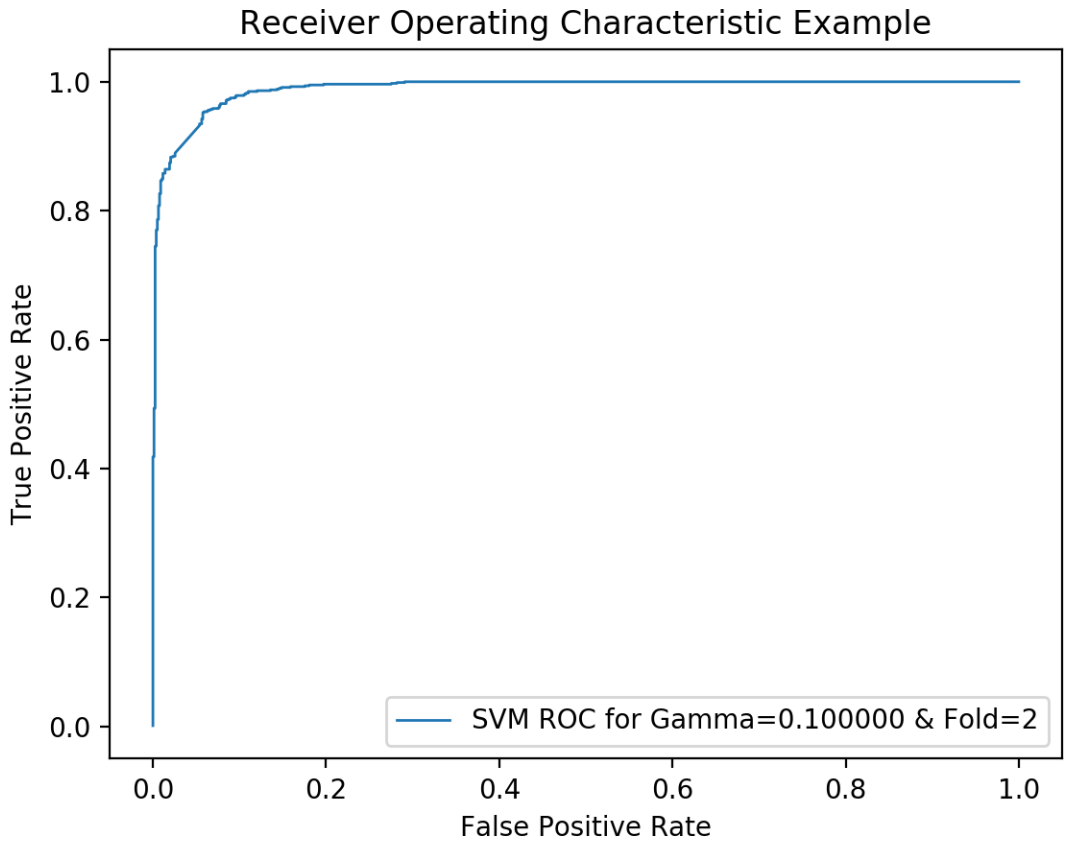
Current model Accuracy 0.944831959417

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-
[[727 54]
 [33 763]]
Current ROC curve



##
##################

Fold Number: 3

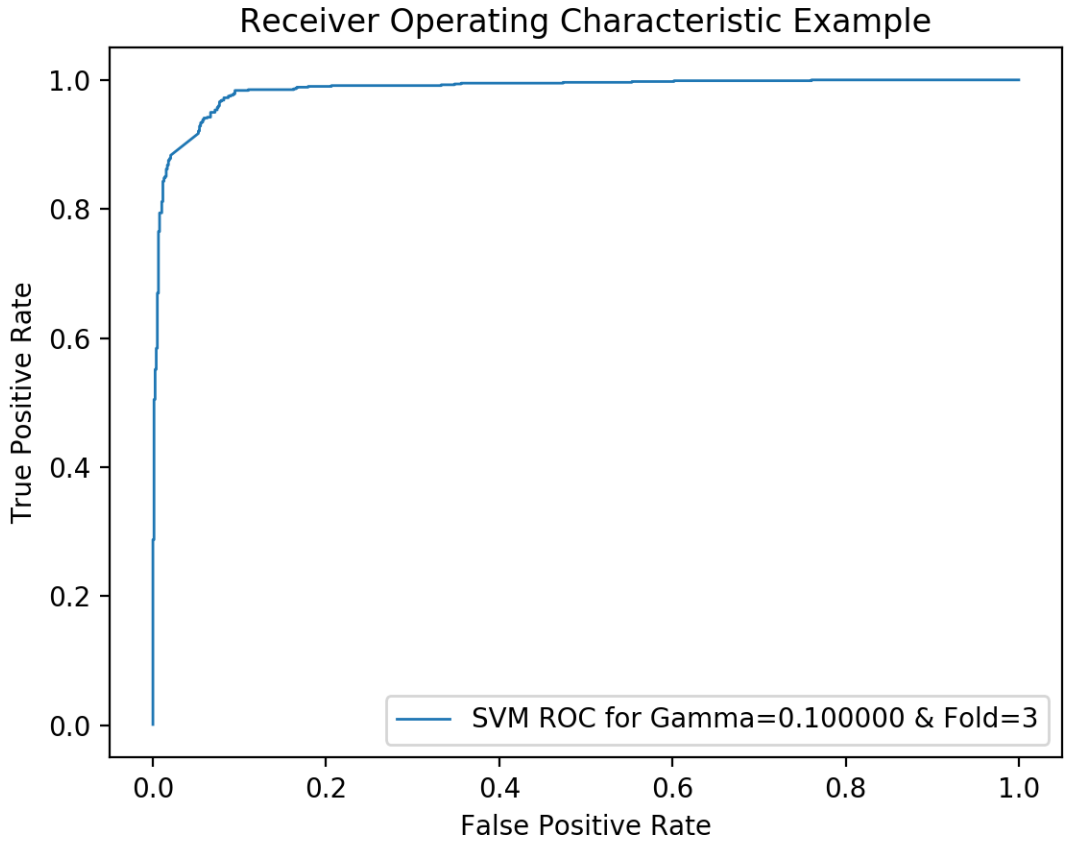
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
 [36 760]]
Current ROC curve



##
##################

Fold Number: 4

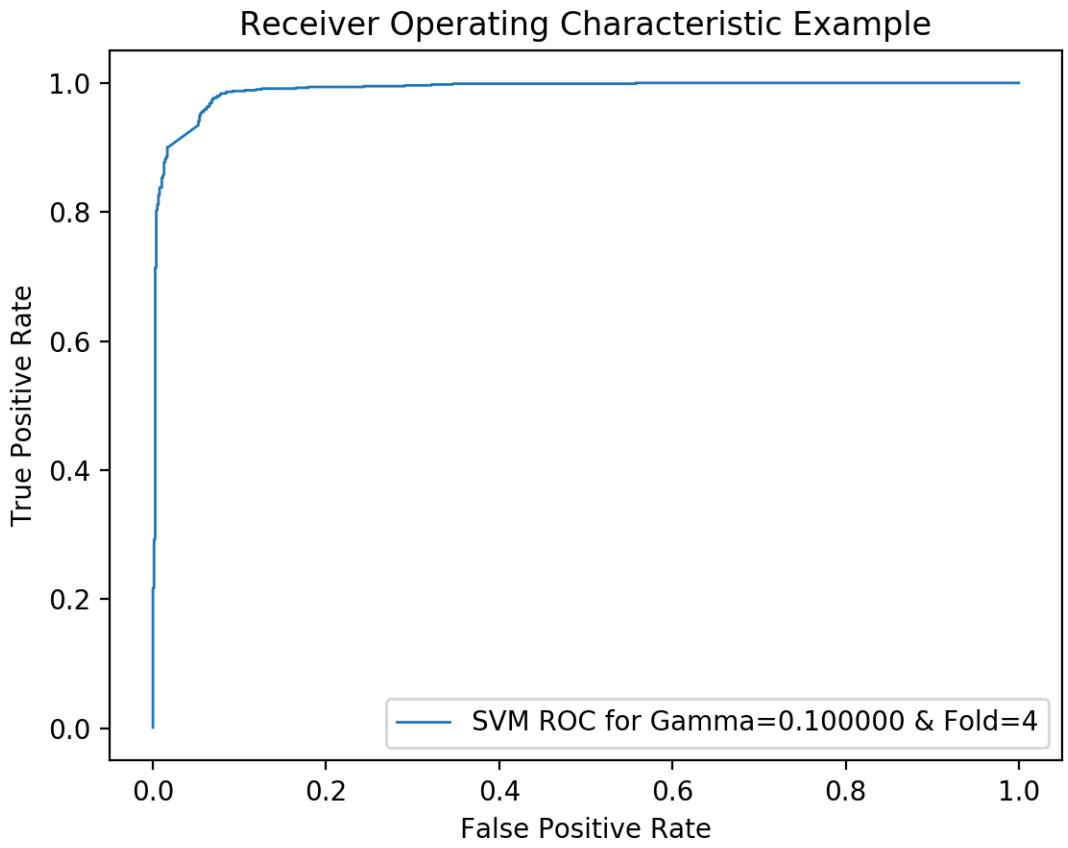
Current model Accuracy 0.94923857868

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

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[731 49]
 [31 765]]
Current ROC curve



##
##################

Fold Number: 5

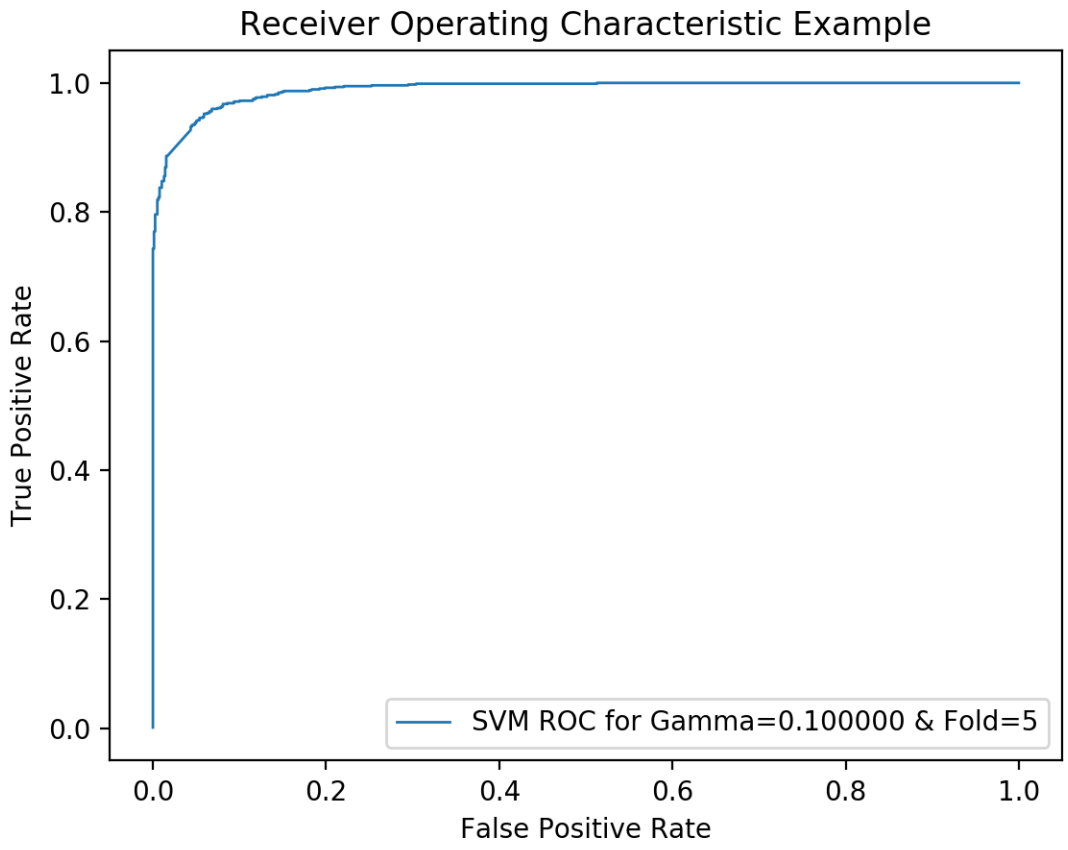
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
[33 762]]
Current ROC curve



****
The value of Gamma is 1
****

Fold Number: 1

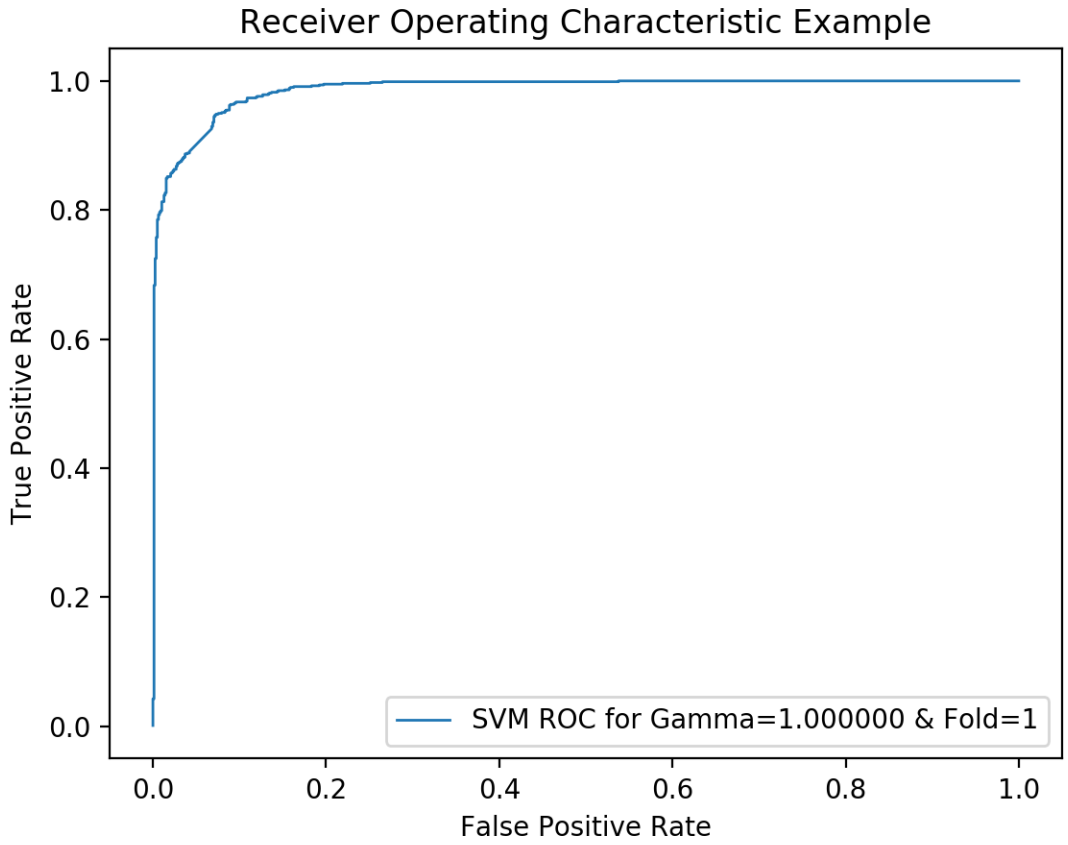
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

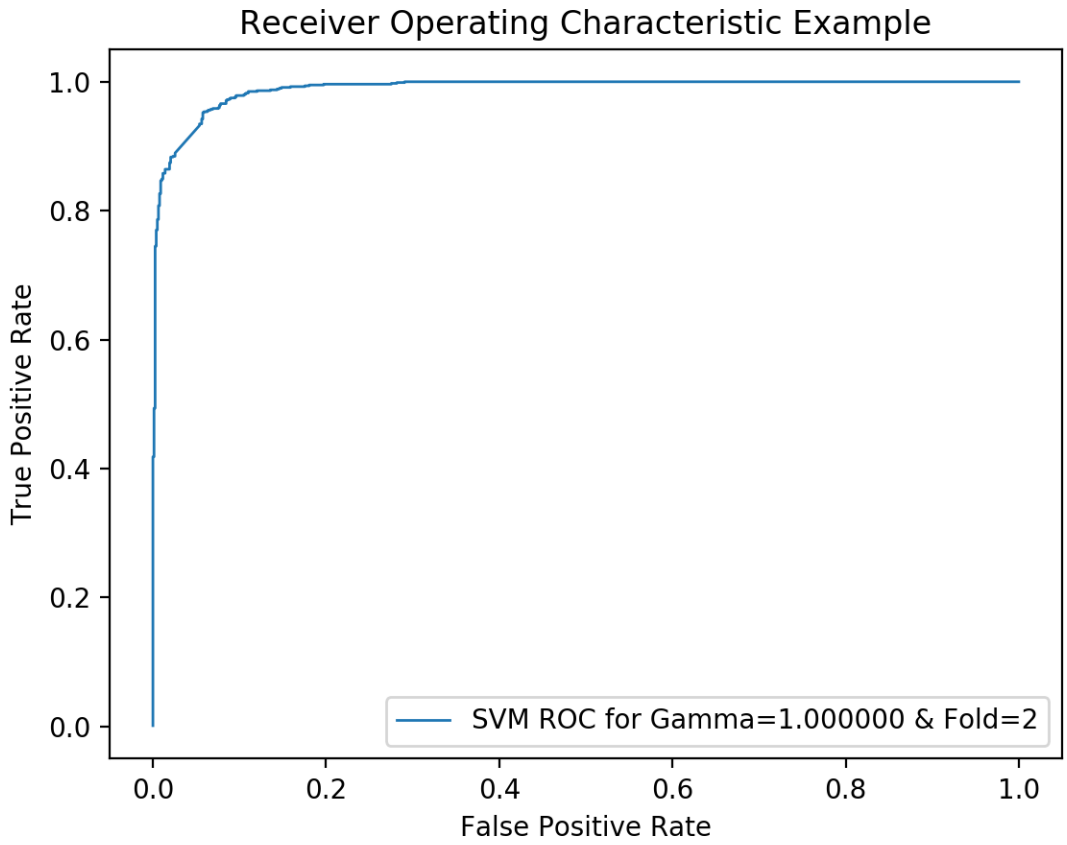
Current model Accuracy 0.944831959417

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-
[[727 54]
 [33 763]]
Current ROC curve



##
##################

Fold Number: 3

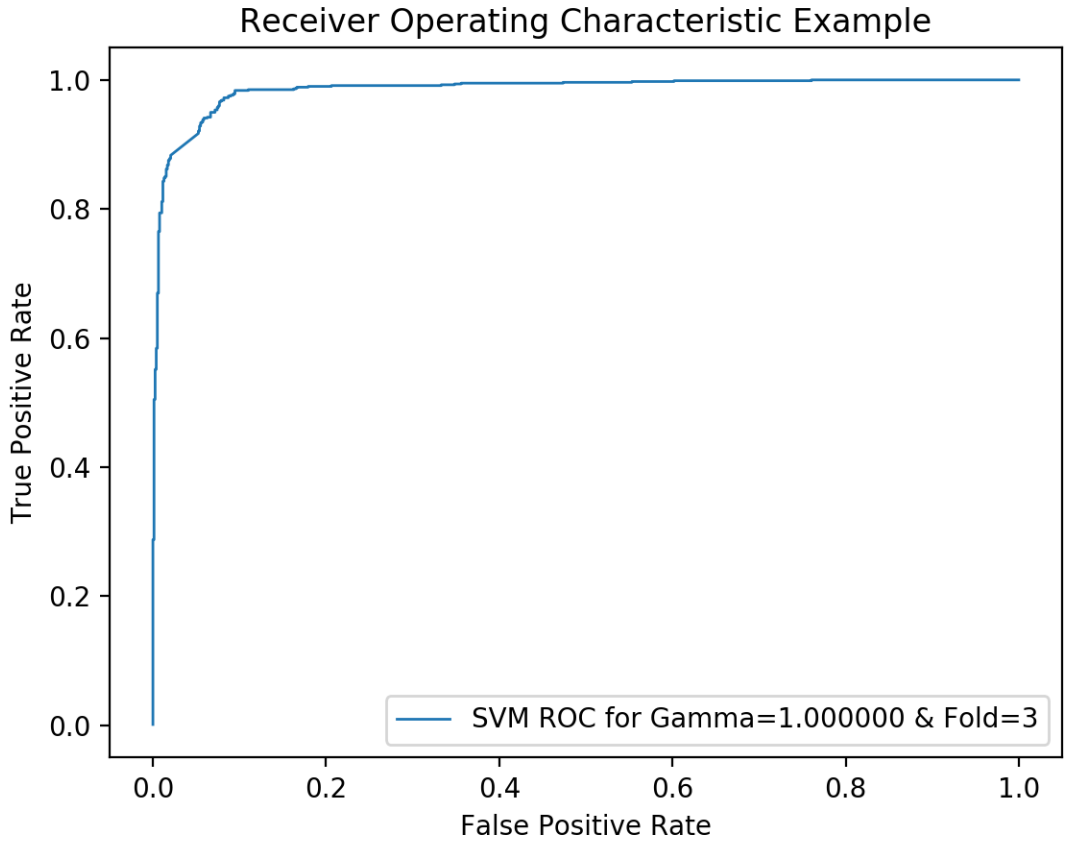
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
 [36 760]]
Current ROC curve



##
##################

Fold Number: 4

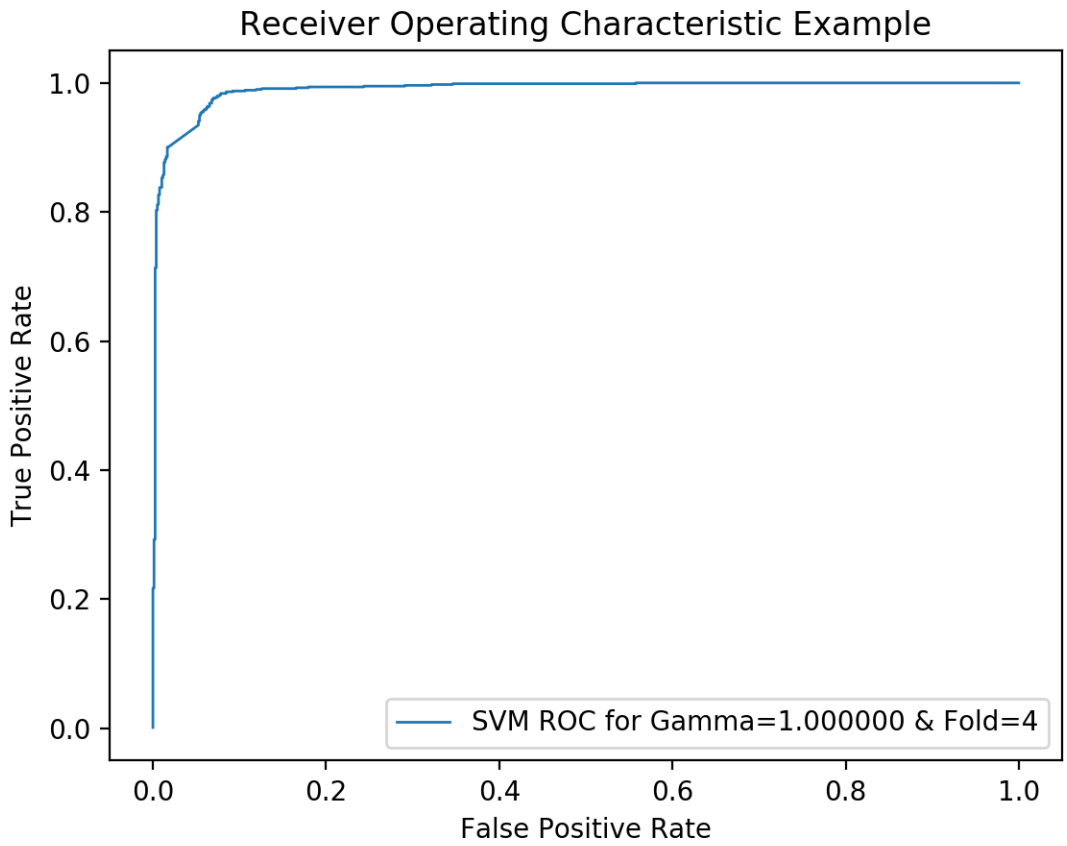
Current model Accuracy 0.94923857868

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

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[731 49]
 [31 765]]
Current ROC curve



##
##################

Fold Number: 5

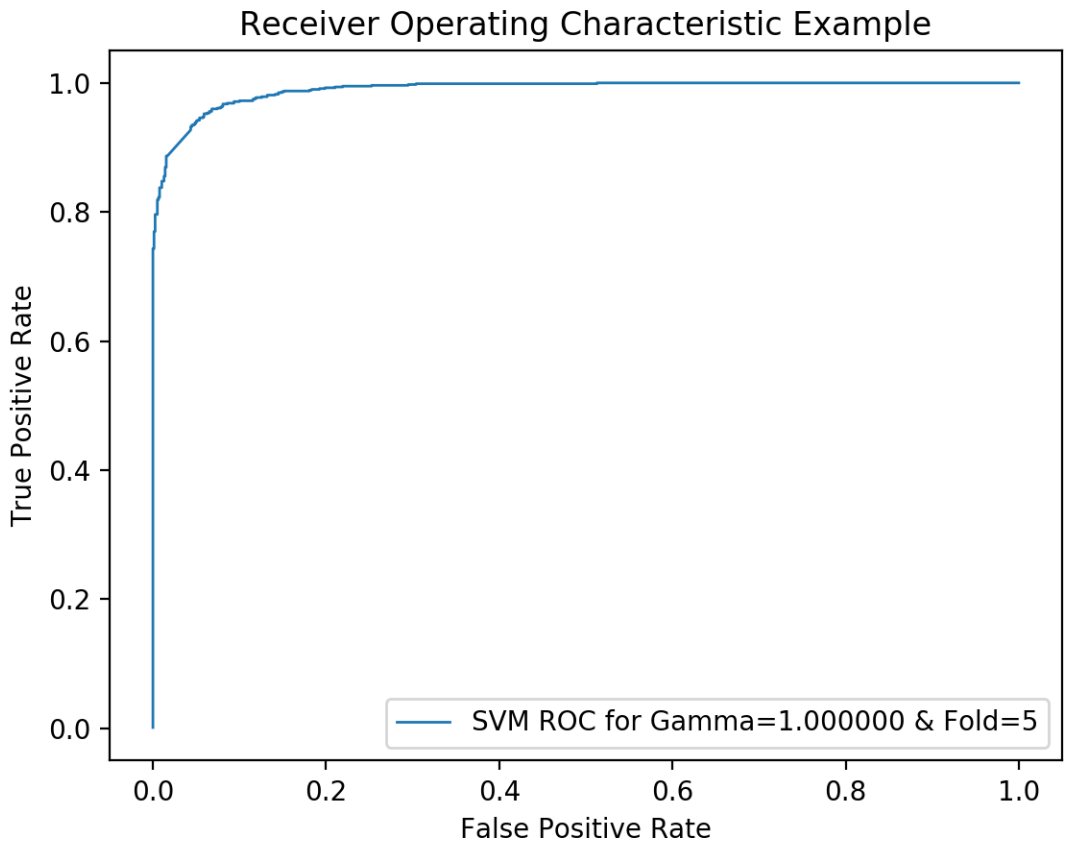
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
[33 762]]
Current ROC curve



****
The value of Gamma is 10
****

Fold Number: 1

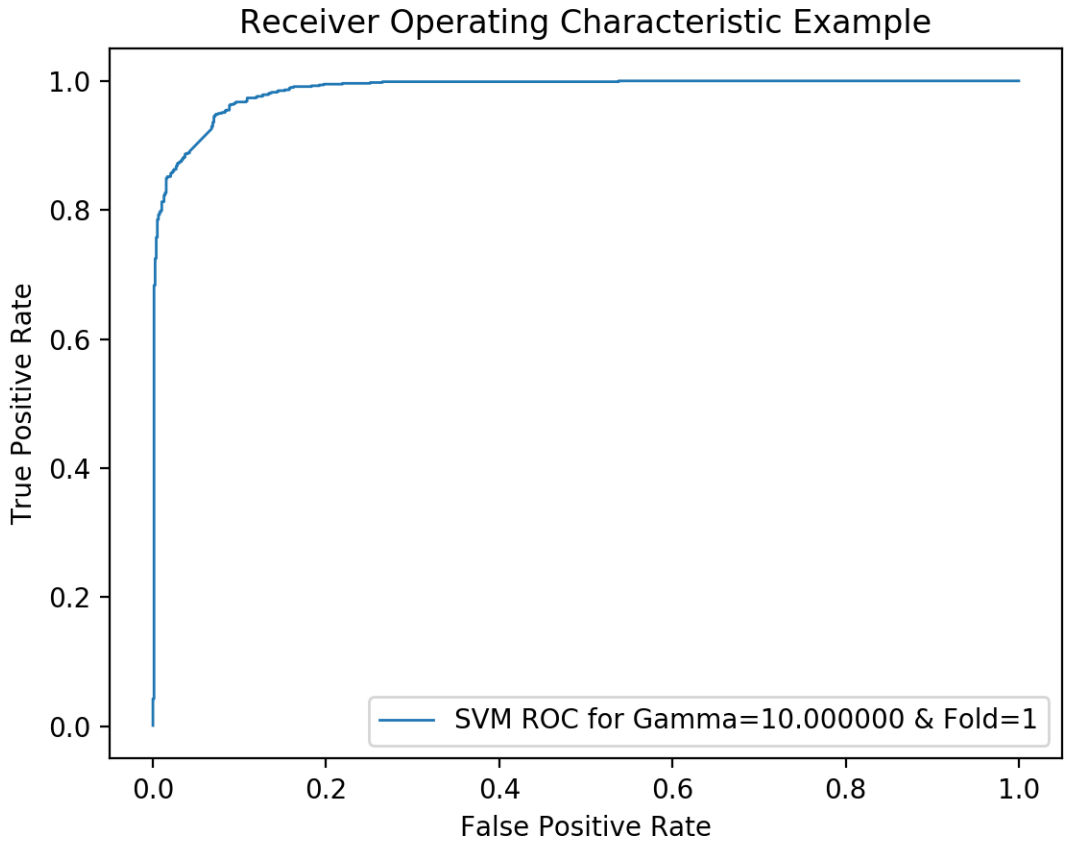
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

Current model Accuracy 0.944831959417

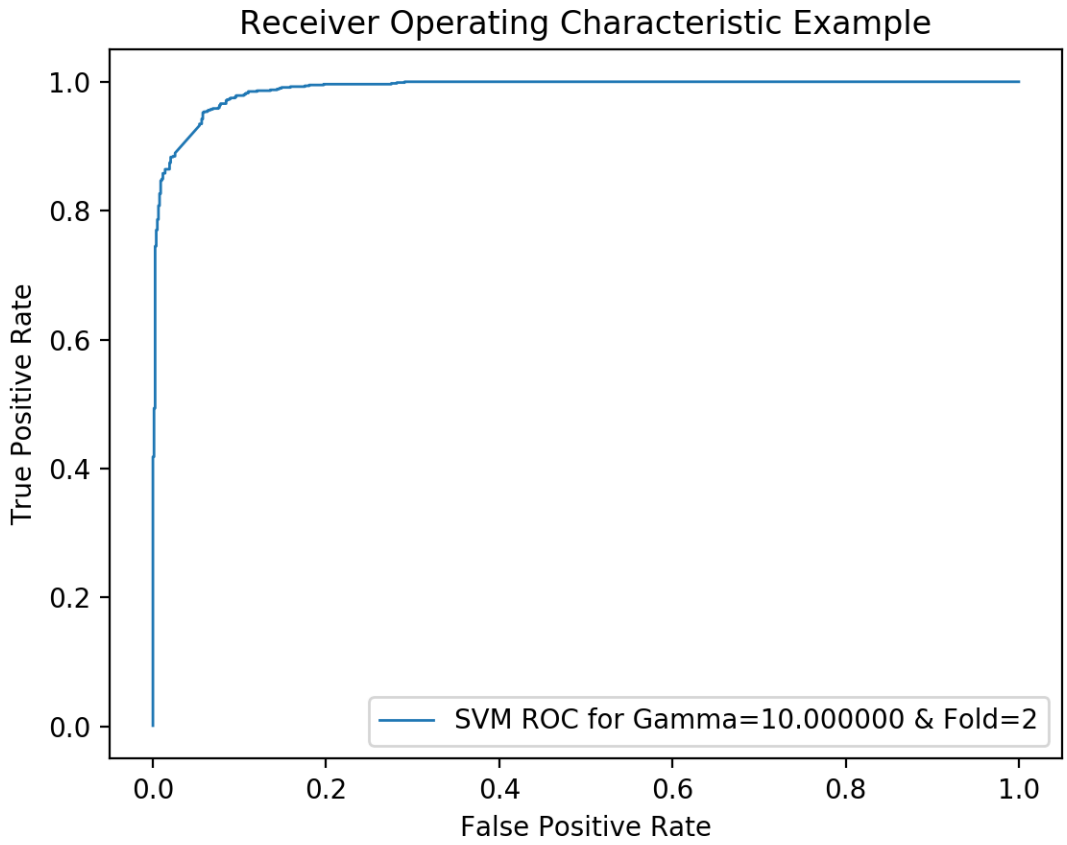
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-
[[727 54]
 [33 763]]

Current ROC curve



##
##################

Fold Number: 3

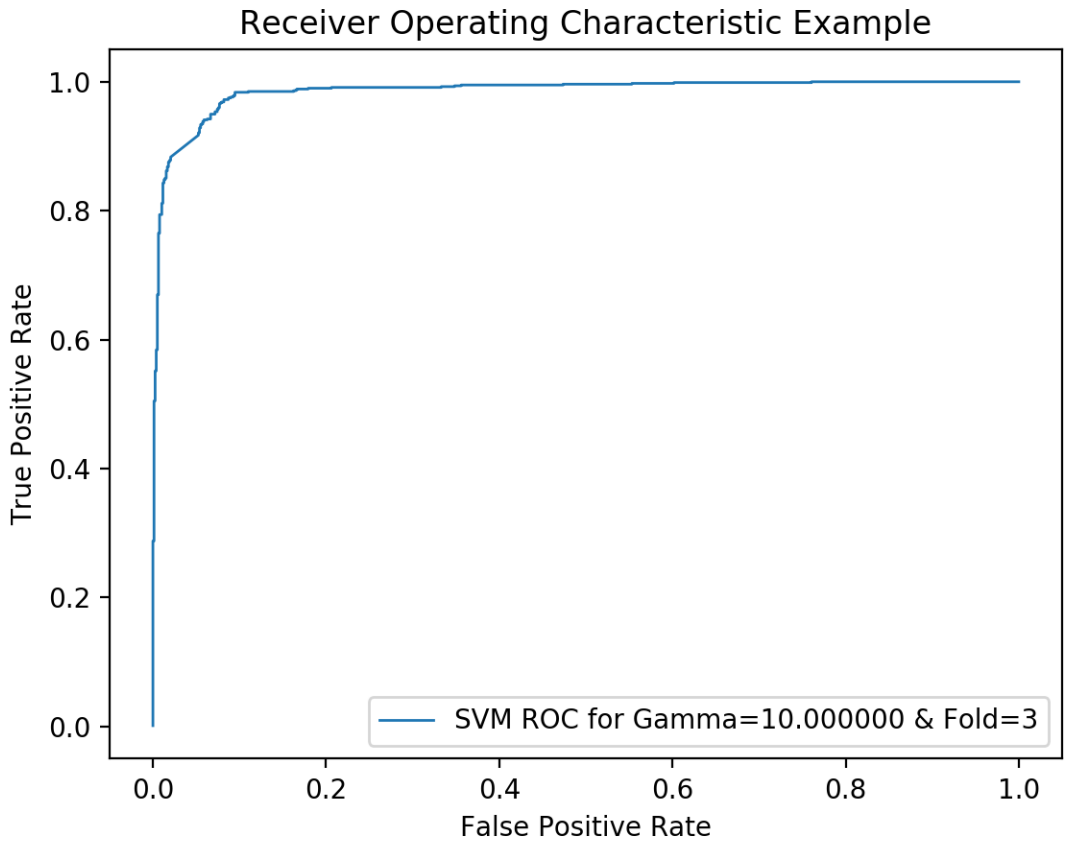
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
 [36 760]]
Current ROC curve



##
##################

Fold Number: 4

Current model Accuracy 0.94923857868

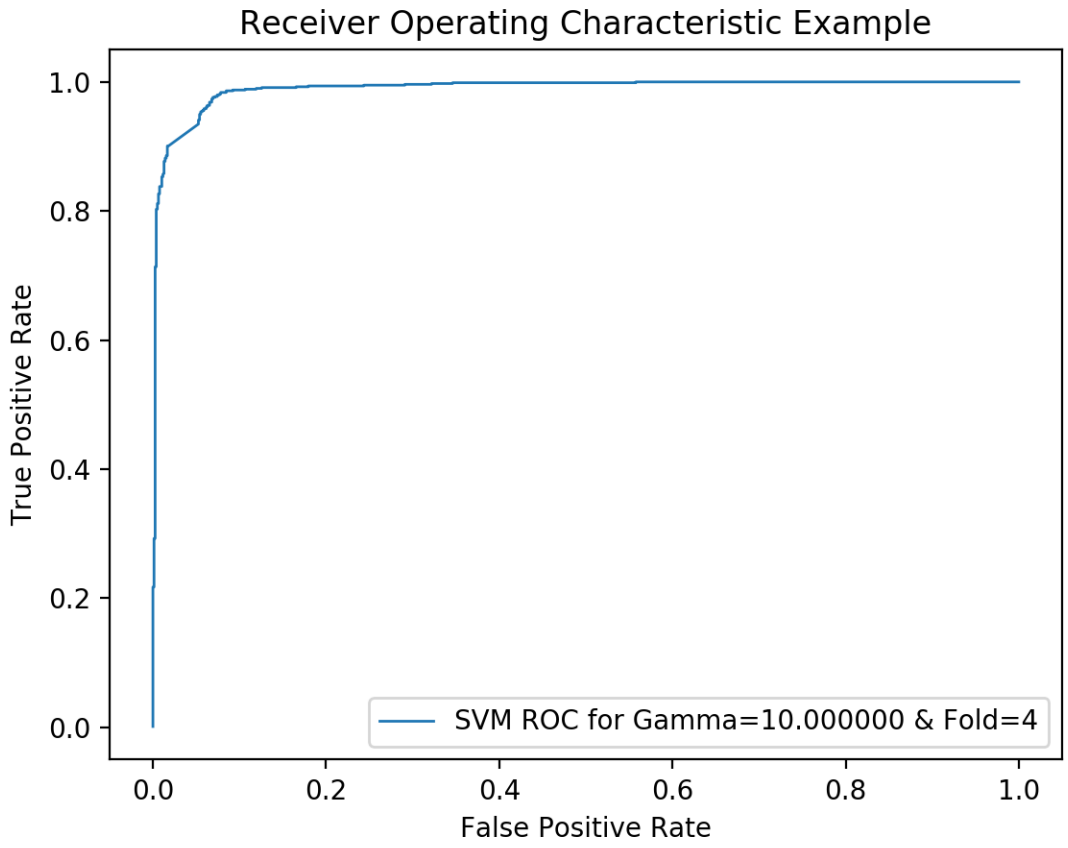
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[731 49]
[31 765]]
Current ROC curve



##
##################

Fold Number: 5

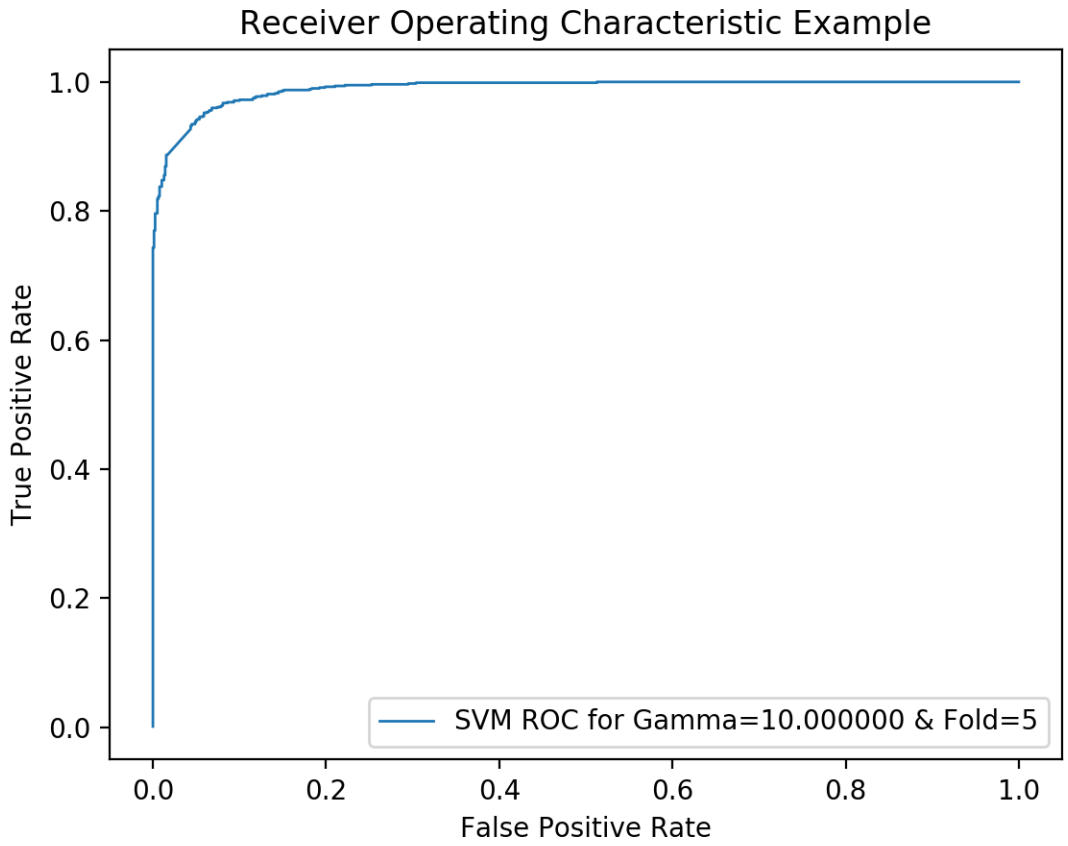
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
 [33 762]]
Current ROC curve



##
##################

****
The value of Gamma is 100
****

Fold Number: 1

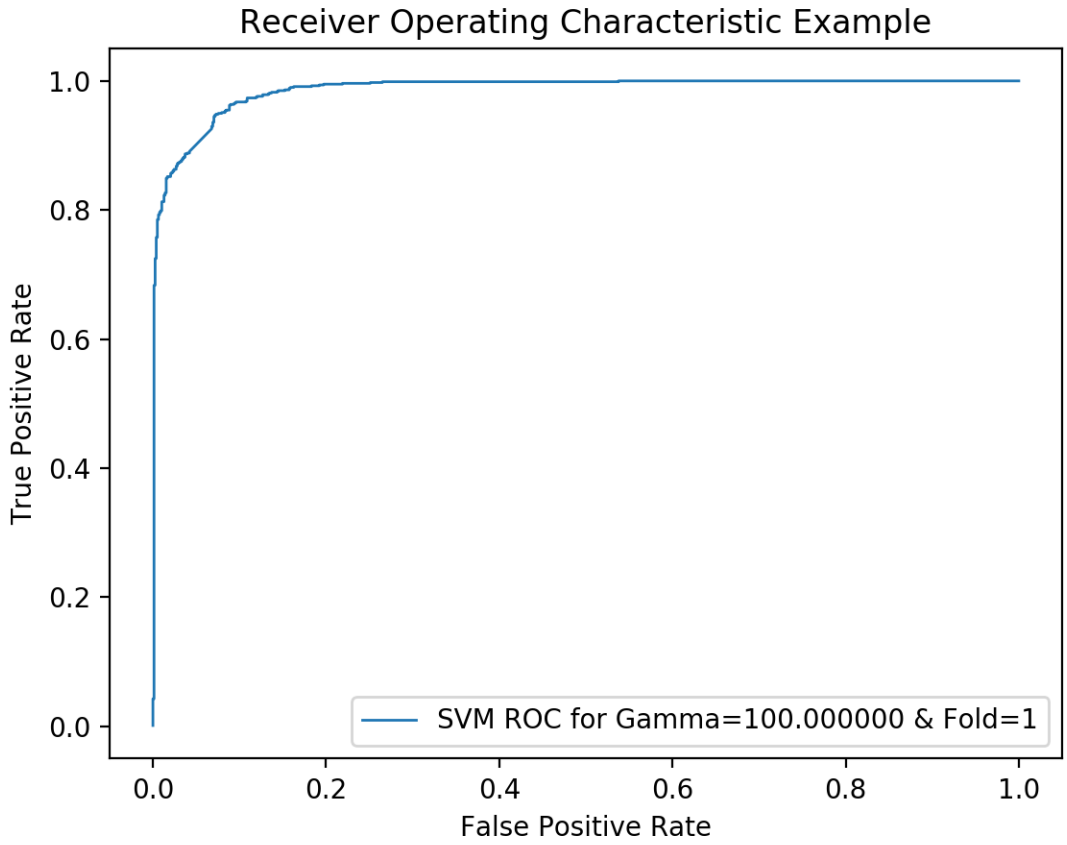
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

Current model Accuracy 0.944831959417

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

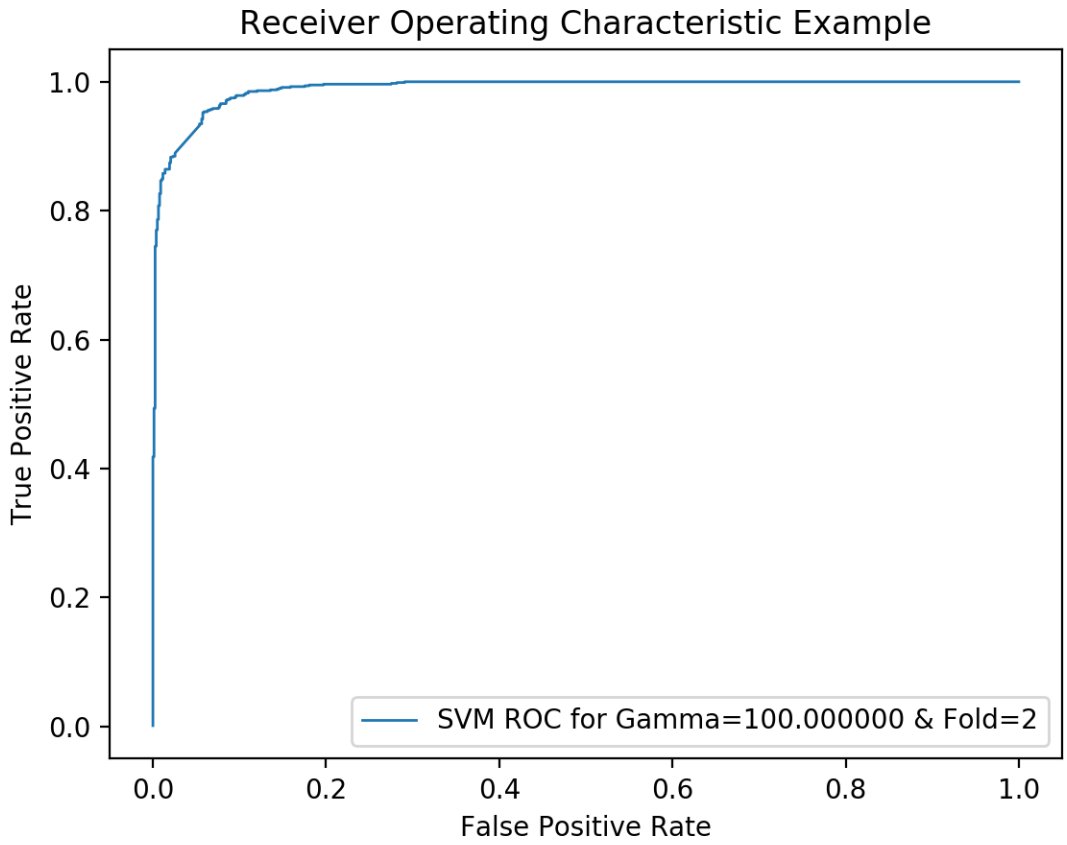
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-

```
 [[727  54]
 [ 33 763]]
```

Current ROC curve



##
##################

Fold Number: 3

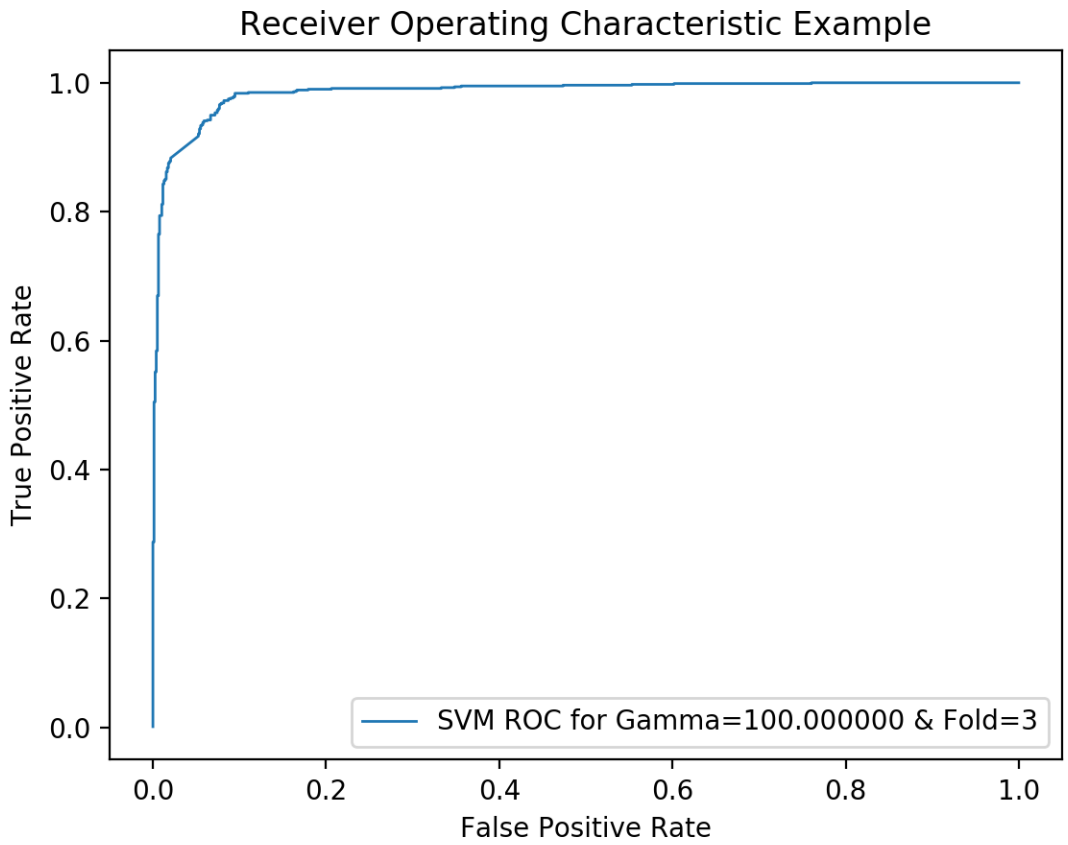
Current model Accuracy 0.940393151554

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

Precision, recall values, f1-score,support are :-

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

Current confusion matrix :-
[[723 58]
[36 760]]
Current ROC curve



##
##################

Fold Number: 4

Current model Accuracy 0.94923857868

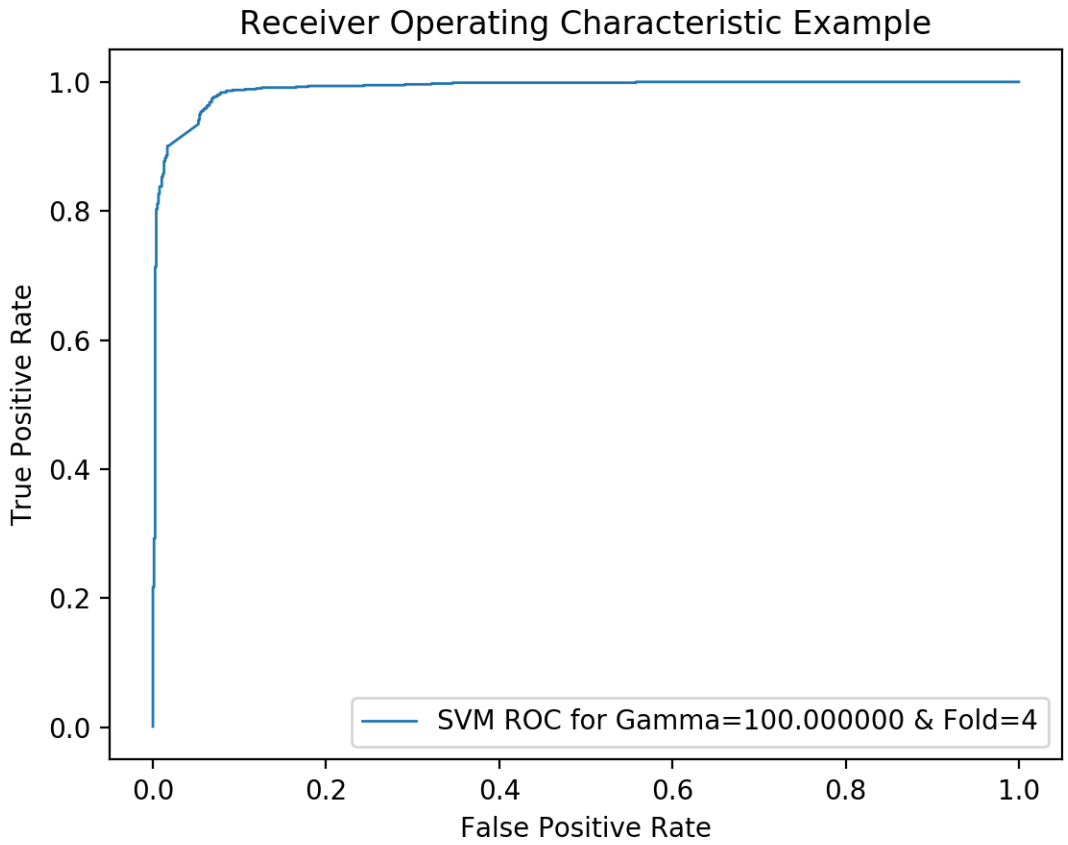
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[731 49]
[31 765]]
Current ROC curve



##
##################

Fold Number: 5

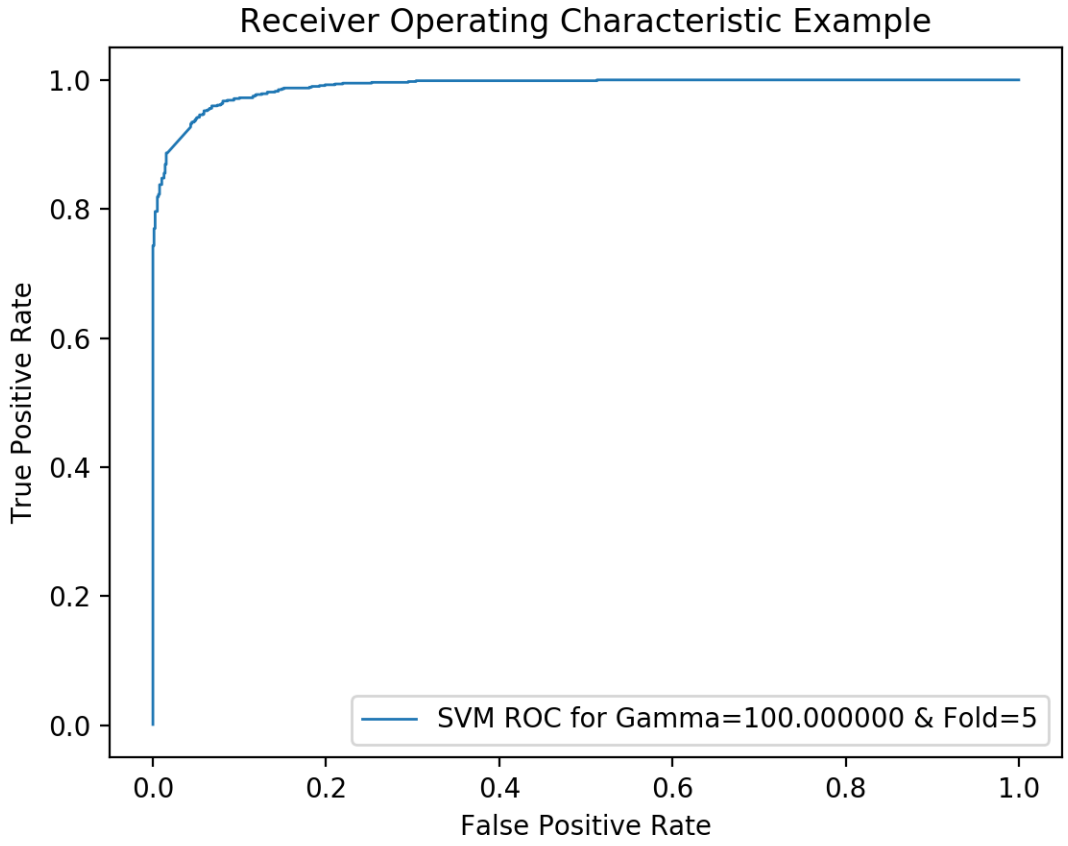
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795
avg / total	0.95	0.95	0.95	1575

Current confusion matrix :-
[[727 53]
 [33 762]]
Current ROC curve



##
##################

****
The value of Gamma is 1000
****

Fold Number: 1

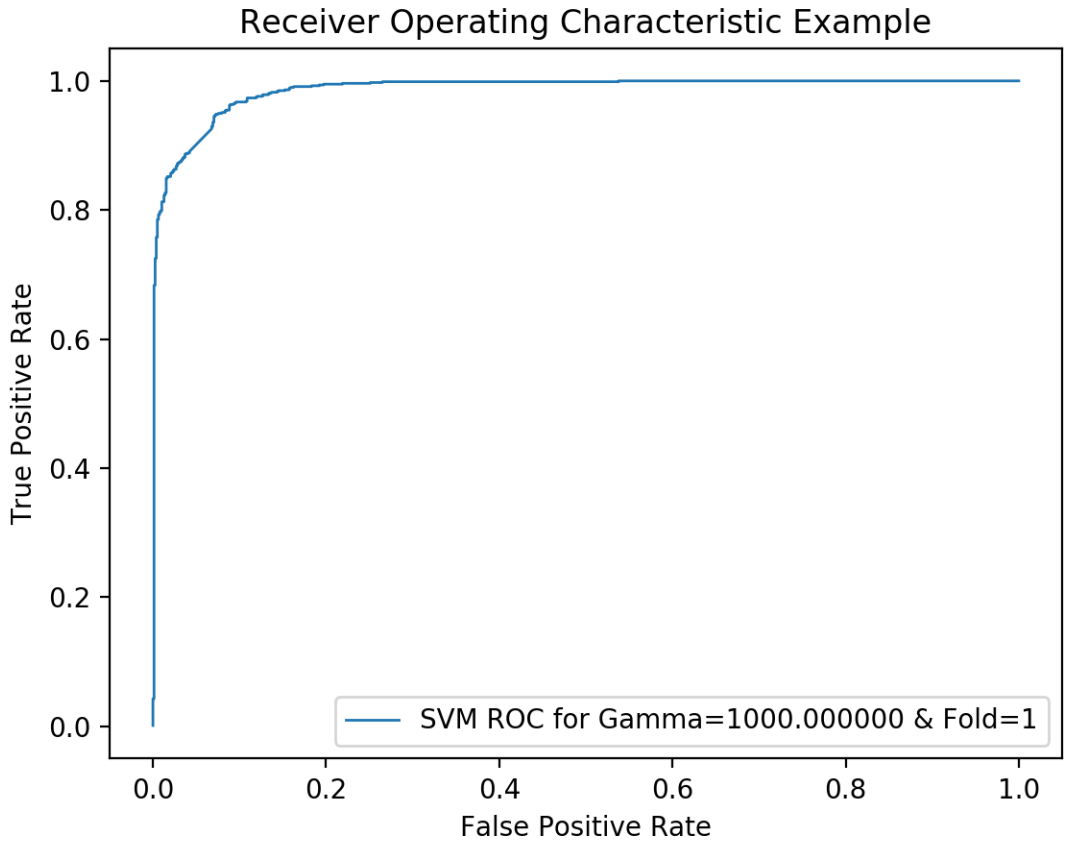
Current model Accuracy 0.9365884591

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.91	0.93	781
1	0.92	0.96	0.94	796
avg / total	0.94	0.94	0.94	1577

Current confusion matrix :-
[[710 71]
 [29 767]]
Current ROC curve



##
##################

Fold Number: 2

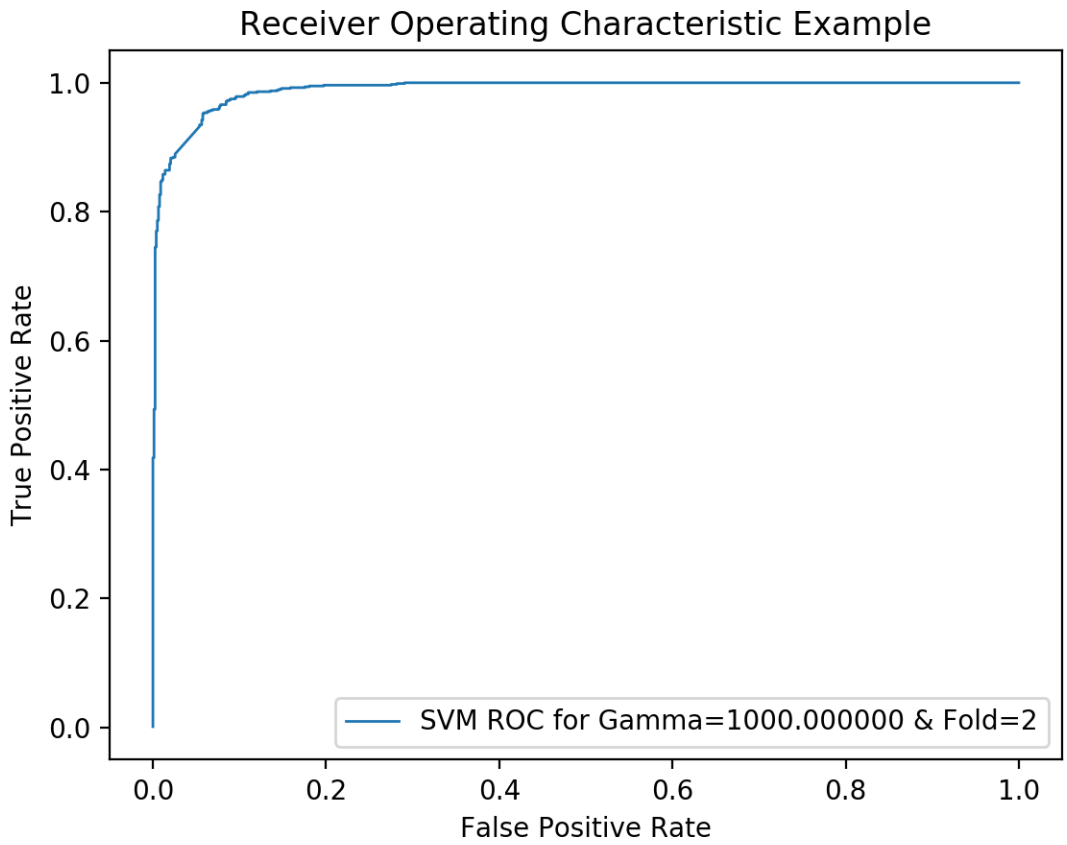
Current model Accuracy 0.944831959417

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	781
1	0.93	0.96	0.95	796
avg / total	0.95	0.94	0.94	1577

Current confusion matrix :-
[[727 54]
 [33 763]]
Current ROC curve



##
##################

Fold Number: 3

Current model Accuracy 0.940393151554

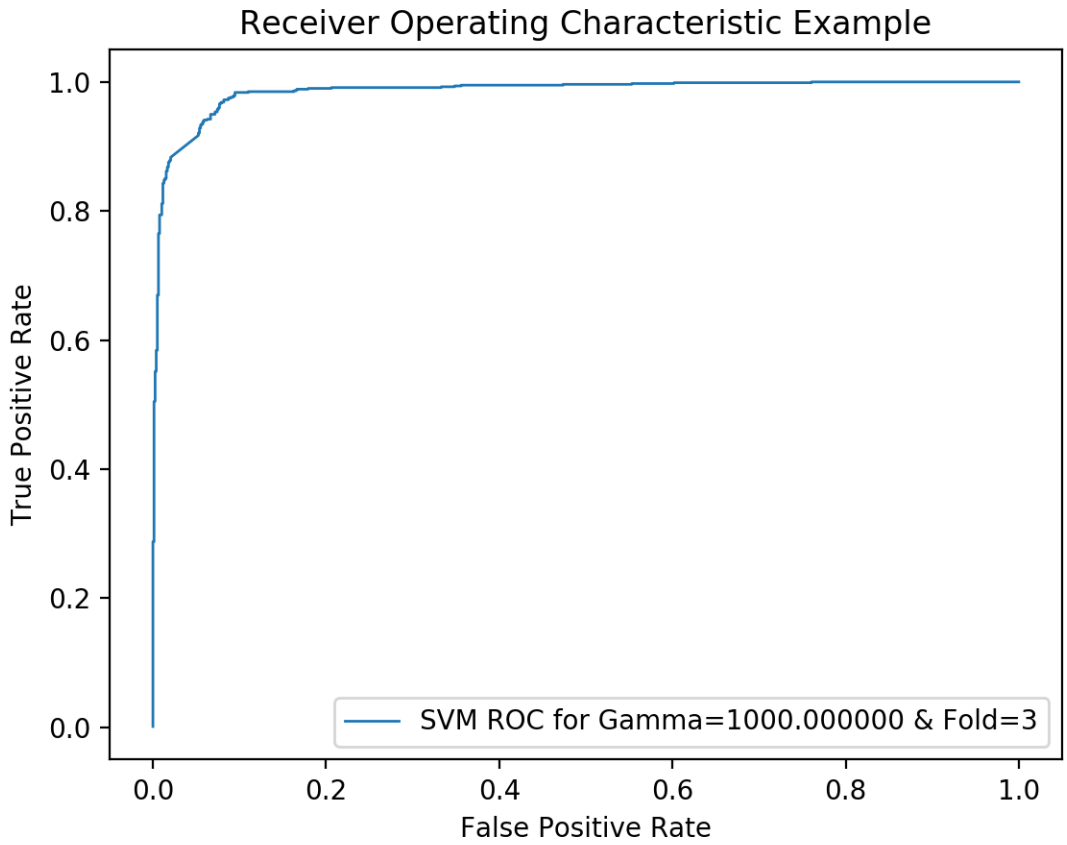
'0' is Computer Technology and '1' is Recreational Activity

Precision, recall values, f1-score,support are :-

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

avg / total 0.94 0.94 0.94 1577

Current confusion matrix :-
[[723 58]
[36 760]]
Current ROC curve



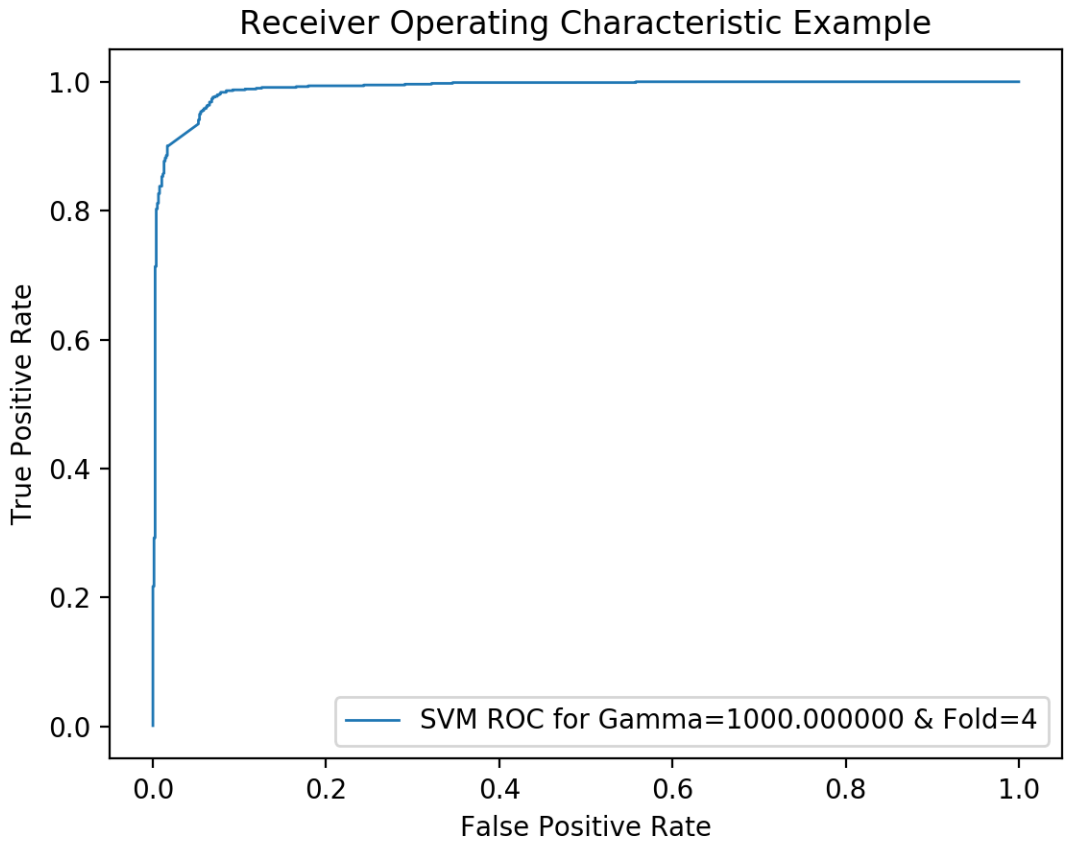
##
##################

Fold Number: 4
Current model Accuracy 0.94923857868
'0' is Computer Technology and '1' is Recreational Activity
Precision, recall values, f1-score,support 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

Current confusion matrix :- [[731 49]
[31 765]]
Current ROC curve



##
##################

Fold Number: 5

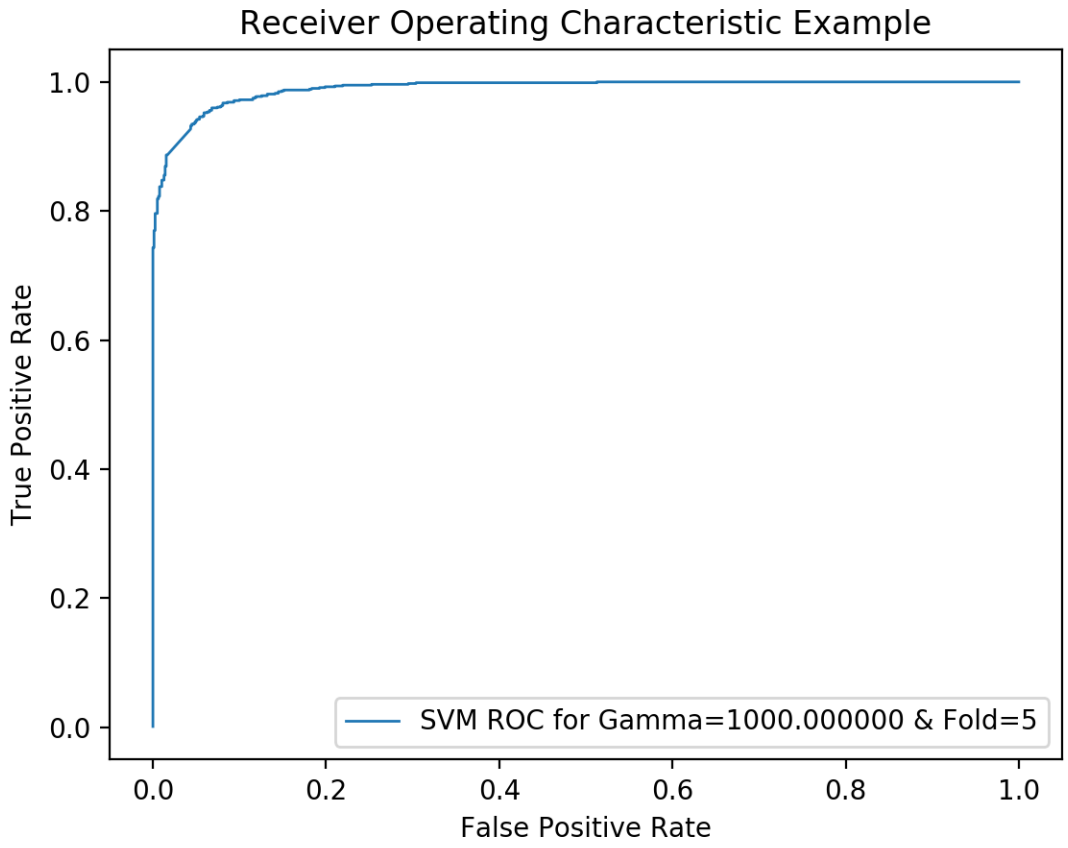
Current model Accuracy 0.945396825397

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.96	0.93	0.94	780
1	0.93	0.96	0.95	795

avg / total 0.95 0.95 0.95 1575
Current confusion matrix :-
[[727 53]
[33 762]]
Current ROC curve



##
##################

The gamma distribution over 5 folds from 0.001 to 1000 seems quite even and doesn't vary significantly. But overall it could be said that gamma is optimum at 0.01 ,0.1

Solution g :-

- Here we use naïve Bayes algorithm for the classification task.
- It outputs the probability that a document belongs to a particular class.
- We study the accuracy and other parameters of this method

Current model Accuracy :- 0.89619047619

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

Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.92	0.87	0.89	1560
1	0.88	0.92	0.90	1590
avg / total	0.90	0.90	0.90	3150

Current confusion matrix :-

```
[[1353  207]
 [ 120 1470]]
```

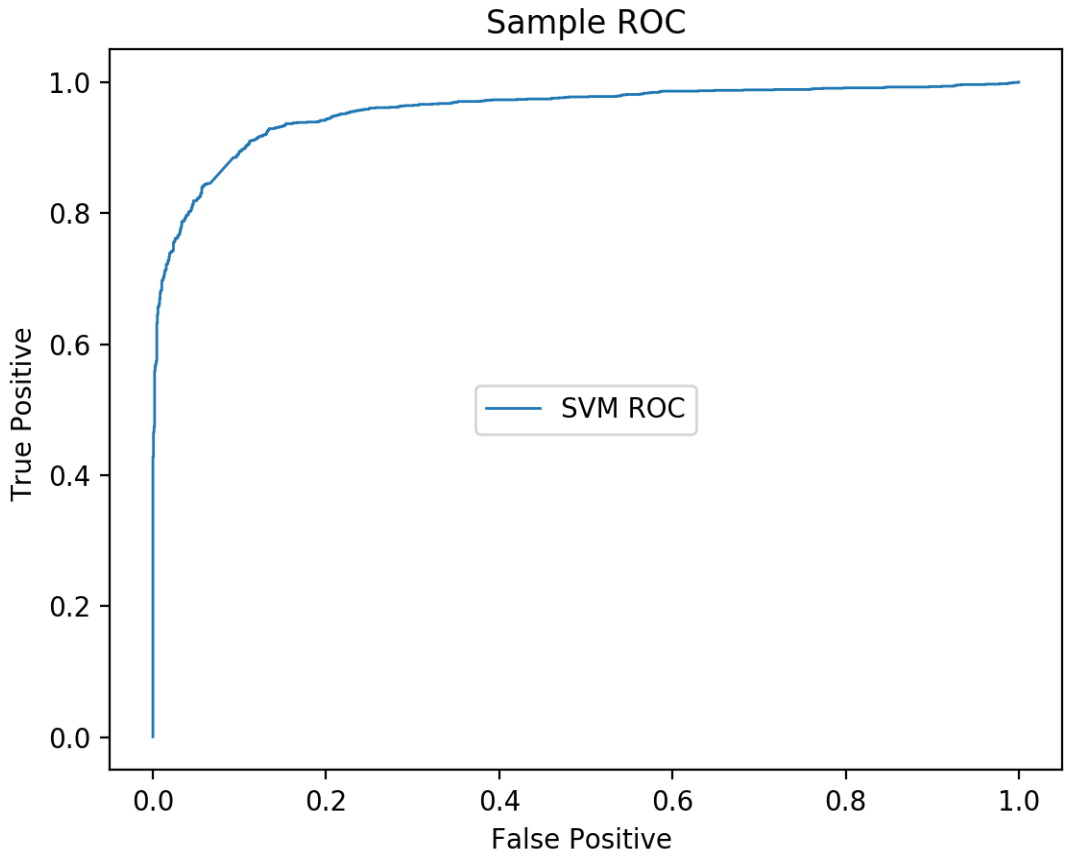
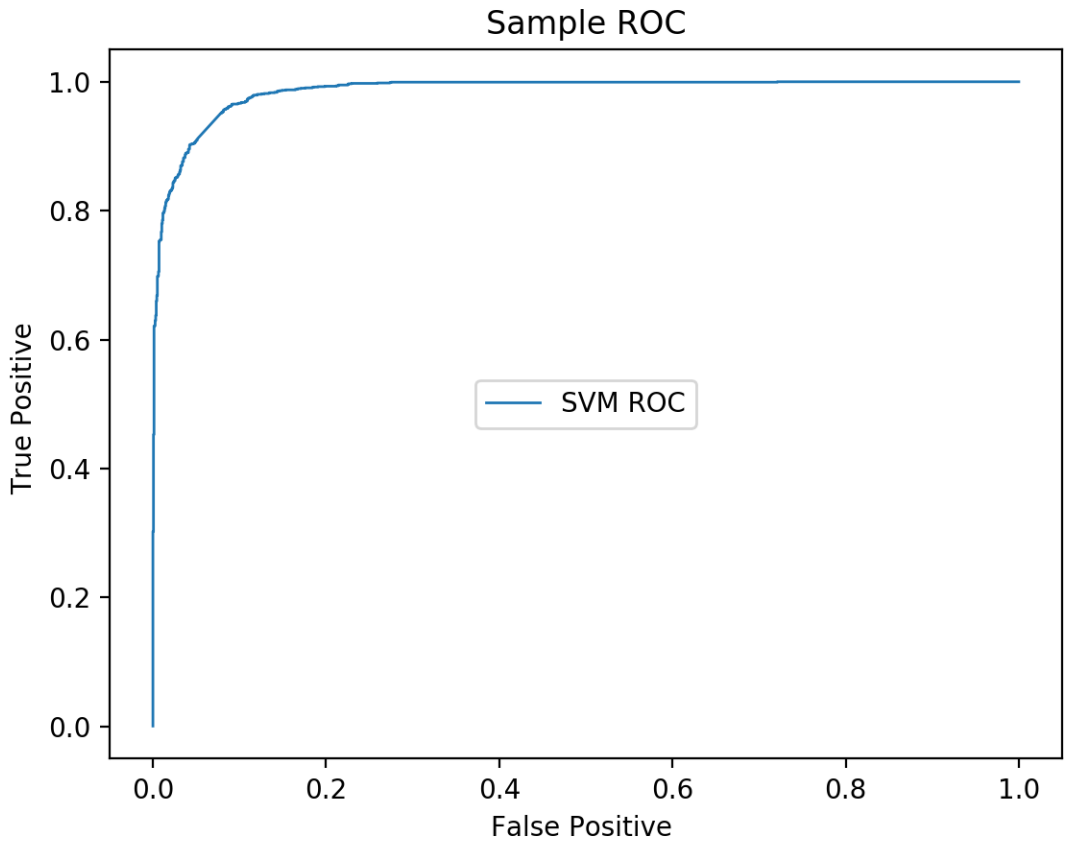


Fig :- Niave Bayes ROC

Solution h :-

```
Current model Accuracy :-   0.936507936508

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

Precision, recall values, f1-score,support 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

Current confusion matrix :-
[[1434  126]
 [  74 1516]]
```



Solution i :-

Here we perform Naïve Bayes classification and multiclass SVM classification (with both One VS One and One VS the rest methods described above) and report the confusion matrix and calculate the accuracy, recall and precision of your classifiers.

Multiclass SVM with One vs Rest Method:

Current accuracy is :- 0.819808306709
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.76	0.76	0.76	392
1	0.72	0.77	0.74	385
2	0.86	0.82	0.84	390
3	0.95	0.93	0.94	398
avg / total	0.82	0.82	0.82	1565

Current confusion matrix
[[297 69 22 4]
[58 297 23 7]
[33 32 318 7]
[2 17 8 371]]
****
Multiclass SVM with One vs One Method:

Current accuracy is :- 0.817252396166
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.75	0.76	0.75	392
1	0.72	0.77	0.74	385
2	0.84	0.81	0.83	390
3	0.96	0.93	0.95	398
avg / total	0.82	0.82	0.82	1565

Current confusion matrix
[[297 65 26 4]
[60 296 24 5]
[35 33 316 6]
[3 16 9 370]]

****
Multi class Naive Bayes Model with OneVsRest

Current accuracy is :- 0.74249201278
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.61	0.72	0.66	392
1	0.69	0.60	0.64	385
2	0.79	0.70	0.74	390
3	0.89	0.94	0.92	398
avg / total	0.75	0.74	0.74	1565

Current confusion matrix
[[282 73 29 8]
[101 231 29 24]
[75 27 273 15]
[3 5 14 376]]
****
Multi class Naive Bayes Model with OneVsOneClassifier

Current accuracy is :- 0.741853035144
Precision, recall values, f1-score,support are :-

	precision	recall	f1-score	support
0	0.60	0.72	0.66	392
1	0.69	0.61	0.65	385
2	0.81	0.69	0.74	390
3	0.90	0.94	0.92	398
avg / total	0.75	0.74	0.74	1565

Current confusion matrix
[[284 78 22 8]
[102 235 26 22]
[82 26 269 13]
[4 4 17 373]]