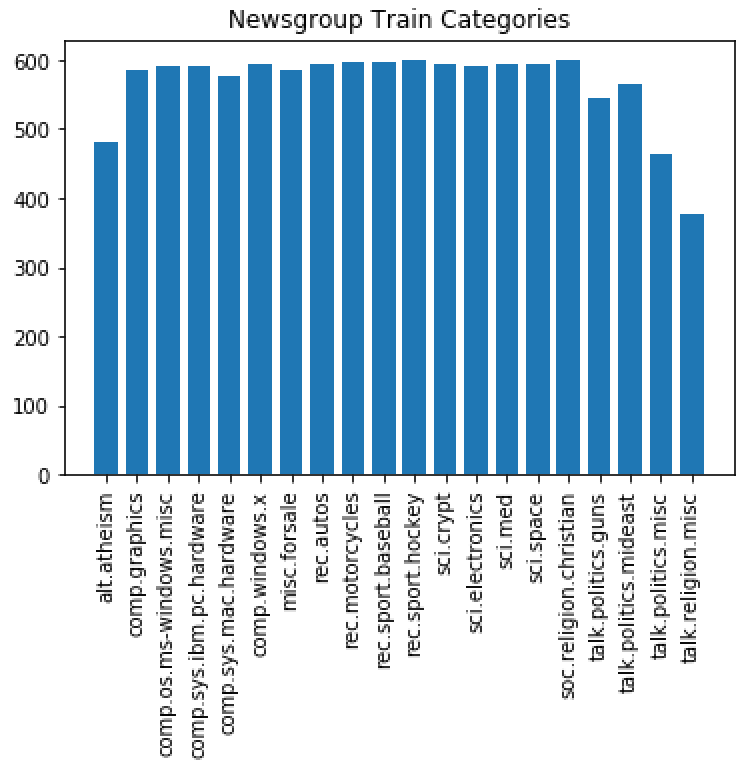
Project 1: Classification Analysis on Textual Data

Jake Turelli | 23 April 2018

Question 1

Plot a histogram of the number of training documents per category to check if they are evenly distributed.



Question 2

Report the shape of the TF_IDF matrices of the train and test subsets respectively.

Train: (4732, 16319)
Test: (3150, 11243)

Question 3

Which is larger: FrobNorm(X - W*H) or FrobNorm(X-U*Sig*V^T)?

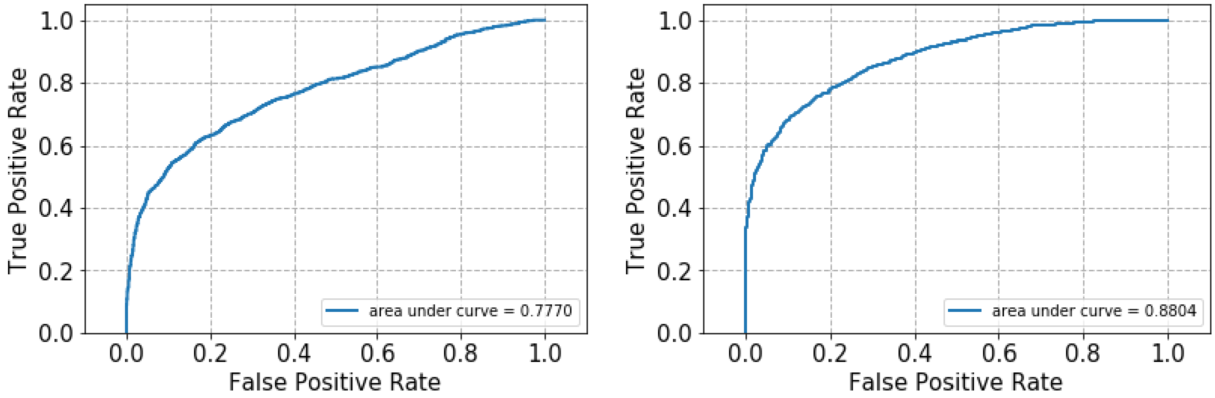
LSI:	
FrobNorm(X-U*Sig*V^T) =	64.0505837928
NMF:	
FrobNorm(X - W*H) =	64.3674139304

NMF is larger.

Question 4

Hard vs. Soft Margins

Left: Hard Margin (gamma = 1000)
Right: Hard Margin (gamma = .0001)



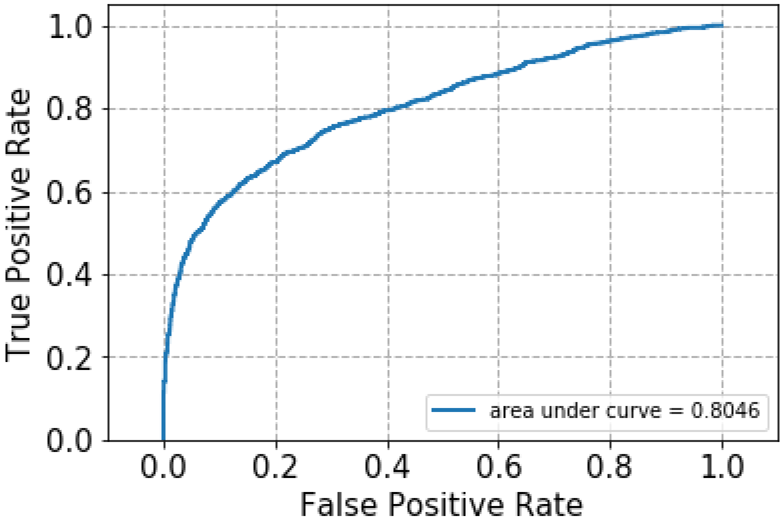
```
confusion matrix, hard =
[[1365  195]
 [ 705  885]]
confusion matrix, soft =
[[   0 1560]
 [   0 1590]]
accuracy, hard = 0.714286
accuracy, soft  = 0.504762
recall, hard       = 0.556604
recall, soft        = 1.000000
precision, hard = 0.819444
precision, soft  = 0.504762
f1 score, hard   = 0.662921
f1 score, soft    = 0.670886
```

It's hard to say which one “performs better” depending on how one determines how each classifier performs, but, in general, it is safe to say the hard margin classifier performed better than the soft margin.

The soft margin SVM never classifies true negatives nor false negatives because gamma is so small.

Cross Validation

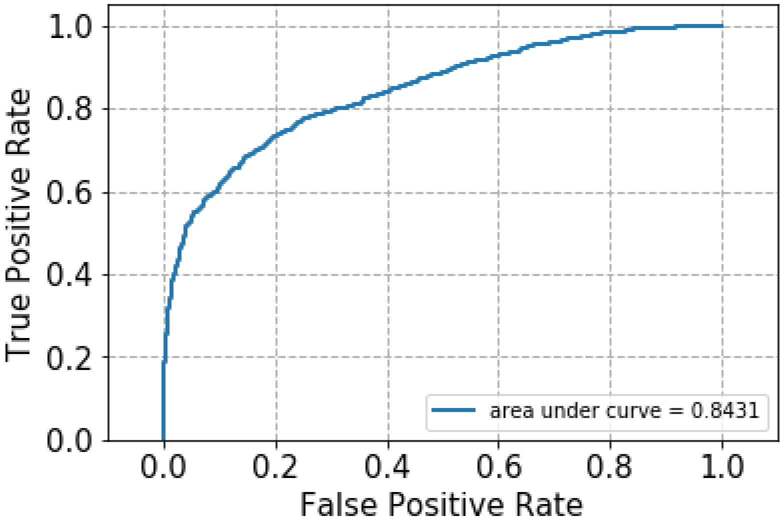
It is found that gamma = 10 produces the best results in 5-fold cross-validation.



```
confusion matrix, gamma = 10:
[[1402  158]
 [ 678  912]]
accuracy,  gamma = 10:          0.734603
recall,    gamma = 10:          0.573585
precision, gamma = 10:          0.852336
f1 score,  gamma = 10:          0.685714
```

Question 5

Logistic Classifier

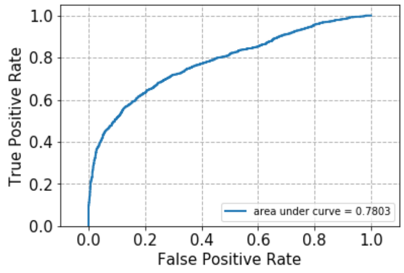


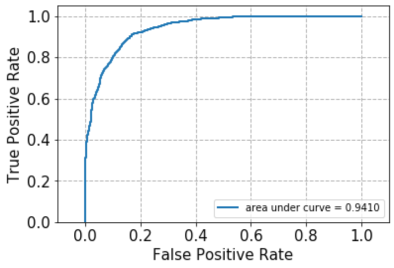
```
confusion matrix, Log Reg:
[[1391  169]
 [ 592  998]]
Accuracy, Log Reg:     0.758413
Recall, Log Reg:       0.627673
Precision, Log Reg:    0.855184
f1 score, Log Reg:     0.723975
```

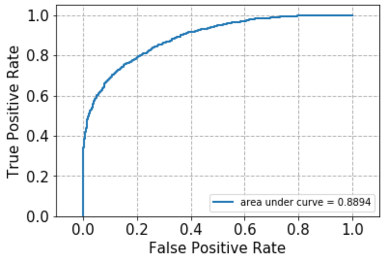
Regularization

It is found the best regularization strengths for L1 and L2 regularization are 10 and 100, respectively.

Of the 3 classifiers: no regularization, L1 regularization, and L2 regularization:





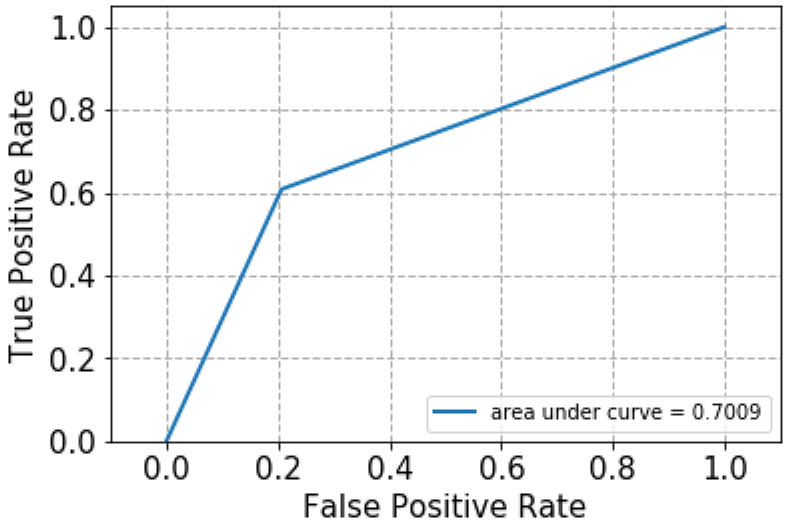


```
confusion matrix, Log Reg, no reg:
[[1364  196]
 [ 700  890]]
confusion matrix, Log Reg, L1:
[[1478   82]
 [ 522 1068]]
confusion matrix, Log Reg, L2:
[[1364  196]
 [ 446 1144]]
--
accuracy, Log Reg, no reg:      0.715556
accuracy, Log Reg,     L1:      0.808254
accuracy, Log Reg,     L2:      0.796190
--
recall, Log Reg, no reg:        0.559748
recall, Log Reg,     L1:        0.671698
recall, Log Reg,     L2:        0.719497
--
precision, Log Reg, no reg:     0.819521
precision, Log Reg,     L1:     0.928696
precision, Log Reg,     L2:     0.853731
--
f1 score, Log Reg, no reg:      0.665172
f1 score, Log Reg,     L1:      0.779562
f1 score, Log Reg,     L2:      0.780887
```

As can be seen from the ROC curves above, one can see regularization has a significant effect on test error, reducing error significantly. It is especially reduced with the L1 regularization. Learnt coefficients (accuracy, recall, precision, f1 score) are all increased for the better with regularization. One might be interested in L1 regularization if they are interested in a robust function or multiple solutions, whereas one might be interested in L2 regularization for a stable solution or a single solution.

Question 6

Naive Bayes Classifier:



```
confusion matrix, GNB:
[[1238  322]
 [ 623  967]]
accuracy,  GNB:         0.700000
recall,    GNB:         0.608176
precision, GNB:         0.750194
f1 score,  GNB:         0.671761
```

Question 7

Grid search using Pipeline, find the best combination:
… still waiting for my code to finish :(…

Question 8

Multiclass SVM Classification:

```
confusion matrix, ovo:
[[175  17 200   0]
 [176  57 152   0]
 [286   8  96   0]
 [ 15   4  40 339]]
confusion matrix, ovr:
[[175  17 200   0]
 [176  57 152   0]
 [286   8  96   0]
 [ 15   4  40 339]]

accuracy,  ovo: 0.426198
accuracy,  ovr: 0.426198

recall,    ovo:    0.426198
recall,    ovr:    0.426198

precision, ovo: 0.533617
precision, ovr:  0.533617

f1 score,  ovo:  0.431965
f1 score,  ovr:  0.431965
```

One vs. One and One vs. Rest provide the same results.