CS229: Meeting Notes 4

October 15, 2013

Emaad Ahmed Manzoor

1 SVM-LDA

I am evaluating an SVM classifier after reducing dimensionality using LDA. Because the full dataset takes a very long time, I've run this experiment on the smaller dataset provided for the CS340 class homeworks.

The design of the evaluation experiment is:

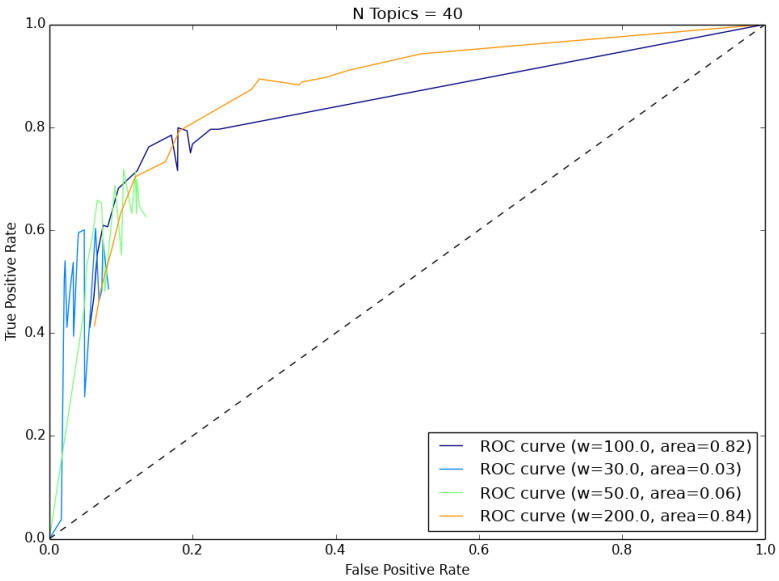
LDA For each number of LDA topics = 30, 40, 45, 50, 60, 70

- Train an LDA model on the non-malware samples
- Infer the distribution of non-malware samples over the topics
- Infer the distribution of malware samples over the topics
- Combine the two distributions for SVM classification

SVM On each of these files containing the entire data reduced to the LDA topics space, train SVM classifiers with an RBF kernel with configurations ($malware\ weight$, c, $gamma$), where:

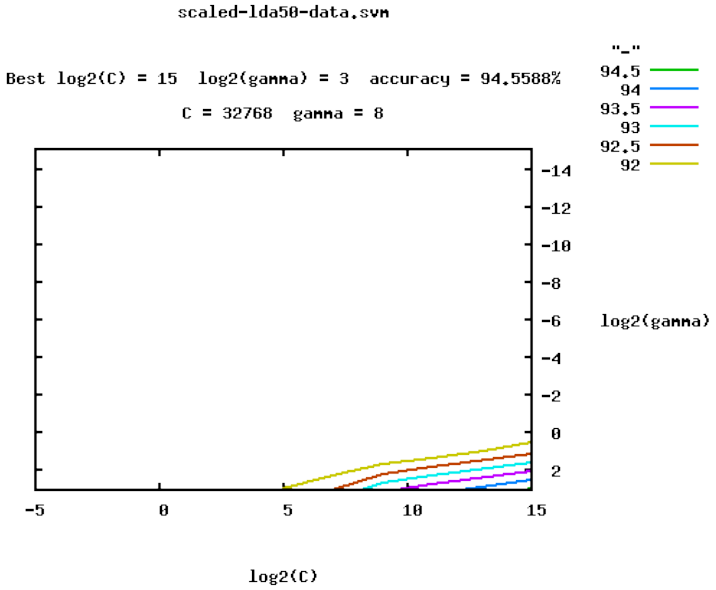
- $malware\ weight = 30, 50, 100, 200$
- $\log_2 c = -5\ to\ 15,\ step\ by\ 5$
- $\log_2 gamma = 3\ to\ -15,\ step\ by\ 6$
- $weights = malware\ weight : 1$

Results For each number of LDA topics, I plotted an ROC curve with each point representing a configuration (c, $gamma$), and each curve representing a setting for the $malware\ weight$. I have currently completed this only for 40 topics.



Observations With low malware misclassification cost, the results are extremely bad, with almost everything classified as non-malware. As the cost increases above 50, the results are a lot better. But with this greater misclassification cost, the training time is a lot more.

Other Results Prior to the above experiment, I ran a grid-search on the reduced dataset for fixed malware weights and plotted accuracy varying with c and *gamma*. Since accuracy isn't a good metric here, this experiment was not useful.


scaled-lda50-data.svm
Best log2(C) = 15 log2(gamma) = 3 accuracy = 94.5588%
C = 32768 gamma = 8
"-"
94.5
94
93.5
93
92.5
92
-14
-12
-10
-8
-6
-4
-2
0
2
log2(gamma)
-5
0
5
10
15
log2(C)