CS229: Meeting Notes Week 2

(originally sent via email)

Observations on the data

There are only ~32,000 unique rows in our permission matrix. If we group apps that have exactly the same permissions together, only 37 of these groups have a mixture of malware and normal apps. The other groups are pure.

Observations on the correlation between every pair of permissions

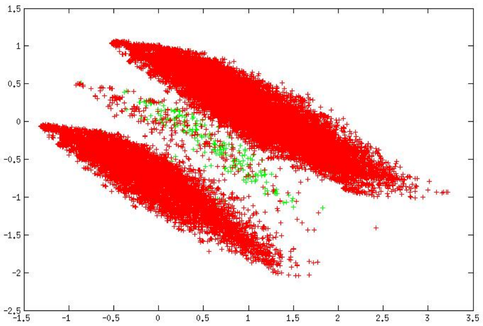
A table of pairwise feature correlation coefficients:

https://docs.google.com/a/kaust.edu.sa/spreadsheet/ccc?key=0AtiJ3begshxbdGVKMGg3YmVCaVlvekVhVGF6QzRacGc#gid=0

It seems like all the "system" features (i.e., features that are unavailable to 3rd party applications) are highly correlated. Can we replace all these features with a single "system" feature?

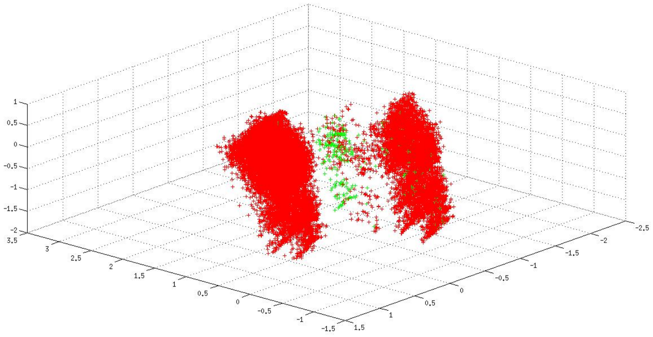
Observations on PCA

I applied PCA, but I'm not sure it makes sense with binary features. I went ahead and plotted the data in the space of the first 2 principal components and 3 principal components. The cumulative variance for the first 73 principal components is in cumvarpca.txt (attached).



Observations on Latent Dirichlet Allocation

To replace PCA, I tried reducing dimensionality with this technique that's usually used to discover "topics" of documents. Each topic is a mixture of features. Each document is represented by a binary vector indicating the words in



the document, and after LDA, each document is represented as a mixture of n topics, where n is set before running LDA.

An example of the "topics" I found is here:
https://docs.google.com/a/kaust.edu.sa/document/d/1fJDwb5TfSCQZ9JQzoqY180qNf4O34GUyXIF4MV_2Hbo/edit

I calculated the AUC with Naive Bayes and 5-fold cross-validation for different values of n:
https://docs.google.com/a/kaust.edu.sa/document/d/1tY8eJa4-UXAHnhF3IRN0GeMPNoBUDGB2QVWbNXcjpTU/edit

It seems that the AUC flattens out at ~50 topics. Around the same number of principal components are what cover 95% of the variance in the data.

Plans for this week

The Bayes Network tutorial mentioned in class talks about learning network structures. I have a candidate structure in mind. I will first evaluate this candidate structure. I will then read about learning better structures using the algorithms mentioned in the tutorial, and implement them to see if we can obtain better structures.