CS229: Meeting Notes 3

Emaad Ahmed Manzoor

1 Essential Permissions For Each Category

1.1 Permission Likelihood, Given Category)

If I use the likelihood of a permission given a category as a metric, the only permission that scores above 0.9 is INTERNET, in any category.

1.2 TF-IDF

I considered a concatenation of all the apps in the same category as my document; so I had 30 such documents. Then I computed the TF-IDF scores for each permission in a category.

I have attached an Excel sheet with the sorted TF-IDF scores for each category, and the permissions sorted by TF-IDF score for each category.

The maximum TF-IDF scores were significantly high for only 10 categories, the other 20 categories have low TF-IDF scores for all permissions.

This may be a brittle metric, because a permission will have a TF-IDF score 0 if it is present in at least 1 app in each of the categories. So even a single outlier in each category having a certain permission will cause the TF-IDF score of that permission to become 0 in all categories.

1.3 Decision Tree

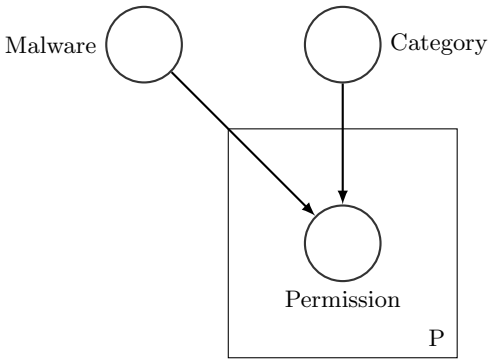
I constructed a binary decision tree with 10-fold cross validation on the data with permissions as features and the category as the response. I haven't yet evaluated this.

1.4 Observations

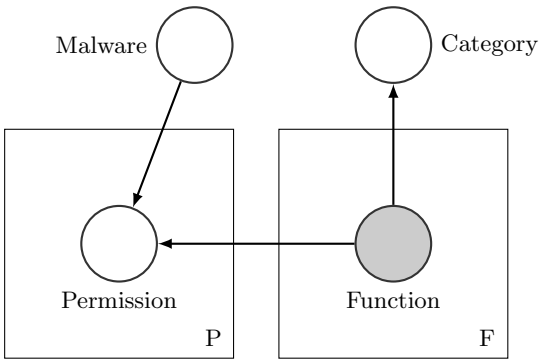
I think the app's category does not influence its permissions. The app's "functions" (photo, music, etc.) would directly influence its permissions, and a given category would contain apps with different functions. This is why a category would contain a mix of permissions. So we need to infer an app's functions from its permissions.

2 Bayes Network

I started creating the following network model, all nodes being fully observed.



But I realised we do not have the category for malware apps. Also, since the category appears to indicate a group of apps with different functions, the following network seems more appropriate.



The functions are unobserved nodes. Functions and permissions are assumed to be mutually independent from other functions and permissions, respectively. The category is partially observed (since it is available only for non-malware apps).

3 Next Steps

3.1 Infer Functions From Permissions

One approach is to go back to the topic models and observe if each topic encodes an app function. If each app can be represented as a mixture of topics (functions),

check if there is some set of features that will discriminate malware apps in this representation. The quality of the topic model itself will also have to be evaluated.

3.2 Bayes Network Learning

The network proposed above has a known structure, but unobserved and partially observed nodes. I will have to learn the parameters of this network and evaluate its quality.