2.15a. 13.73%

2.15b. 17.40%

2.15c. 25.97%

15.1a.

When we use n-fold cross validation we iterate through the data separating as a majority for the training set and a small subsample as validation data for testing the model. So we have M/n since M is the data and n is the subsamples, but we iterate this over the whole dataset making each section a small subsample meaning we do this n * M/n times which is equal to M.

15.1b.

Total of N/n * n = N no match scores

15.1c.

Computational Inexpensive using 5-fold vs 10-fold but the advantage of 10-fold is that we have a better estimate of our model making it more generalizable.

15.7.

In Zip

15.10a.
.99995

15.10b.
.0868

15.10c.

If we change the threshold to reduce the false positive rate the false negatives would rise but we would be destroying less files.Doing a secondary test would be the best since we could almost ensure that the files we are looking at are malware but it requires more resources and time.

Also, train an HMM with N=2 hidden states on English text, then use the model in a generative mode to generate 100 characters of fake English text. Repeat with N=12 hidden states. Is the fake English text based on N=12 hidden states any better than that based on N=2 hidden states?

The 12 hidden states work better than the 2 hidden states because more hidden states allow the model to capture patterns in the data. This ends up producing better fake English text however it still isn't good, 26 hidden states would work the best.