Ivan Miller - Problem Set 7 for Adversarial AI. Tuesday, Nov. 22, 2022

1. Suppose you have a training set containing 1000 spam emails and 1000 ham emails, all correctly labeled. Suppose that most individual words occur with roughly equal frequencies in the spam and ham emails. (e.g. words that appear in 50 of the spam emails occur in approximately 50 of the ham emails as well.) What else might you try in order to build a model to classify spam vs. ham based on this training data?

To distinguish between spam and ham emails when dealing with two equal sets of words we'll have to look at unique combinations of those words instead of analyzing them individually (we could probably start with pairs or triplets of words in each combination). Additionally, we could include more data points in the decision making process and, not only analyze the body of each email message, but also add some weight to the factors such as whether a given email contains URLs, has attachments or includes media elements (e.g. images).

2. Describe a strategy you might employ to try to craft spam emails to get around a statistically based filter. How might you develop and test your strategy if given access to the filter as a blackbox, but with no access to the training data or process used to create the filter?

Since in this scenario we have no influence on the training process, the only option is to observe the decisions the classifier is making (so-called exploratory attack). We could do that by first "feeding" email(s) to the filter in order to check whether a given email is being classified as ham or spam, then we could focus on emails we know were classified as ham. Using one such email we could start making minor incremental changes (in the general direction of the message we are trying to send) to the original message and testing the result after each change to see if the modified version still passes the filter as ham.

By repeating the process many times with a large number of different ham emails we could make assumptions around words or combinations of words, punctuation symbols, overall length of the message etc. that we should avoid in order to make sure that our modified message is not being detected by the filter.