­­­­­­LinkedIn AI Academy, FY18 Q1

Homework 1: Spam Detection and Cross Validation

In class, we built a naïve-Bayes spam detector. We introduced smoothing and the key metrics: *Precision, Recall, F-Score* and *Accuracy*. The data provided in class was divided into three groups: Ham, Spam and Test.

For this assignment:

1. Aggregate the data into one labeled data set.
2. Store the labels in a separate file with the following on a separate line for each email file:

EMAIL\_ID LABEL (with a tab in between)

1. Divide the data into 10 “folds” (or subsets)
2. Iteratively hold out one of the folds as test data. Train on the other 9.
3. Record the results in a master results file. The form should be

EMAIL\_ID LABEL CLASSIFIED\_AS (with tabs in between)

1. By the end of all 10 experiments, all of the email files will have been part of a test set, and your master results file will contain results for the whole data set
2. *Compute Precision, Recall, F-score* and *Accuracy* for the complete experiment.
3. Repeat the 10-fold experiment again, but don’t do smoothing. Instead, just ignore any unseen words. That is, just don’t add them to your running total of log-probs. Compute all stats. Does smoothing matter for this problem or not?

If you have any questions, thoughts, things to discuss, let’s please do it in Teams so everyone can benefit.