Hi everyone, my name is Xinyue Liu and I am major in Statistics and Computer Science. Today my topic is “Can We Determine Whether an Email is SPAM?”

This semester, under Michael’s guidance, I learned an introductory-level online course “Statistical Learning”. This course is provided by Stanford and focus on data analysis and machine learning. More specifically, I learned to apply different methods to make prediction on the given data sets.

In this spam detection project, I got the data from 4601 emails sent to an individual named George at Hewlett Packard Lab, and each email has been hand labeled as either being spam or good email (we call it “ham”). My goal is to try to classify spam from ham based on the frequencies of words in the email and by comparing the mean error rate of the two prediction methods, which I will talk later, decide which method is a better fit of this data set so we can adopt it for future spam detection. Here I picked a row from the data sets, as we can see, there are 4601 rows, and each row represents an email. There are 58 columns. Among them, each of the first 57 columns is a number between 0 and 1, which represents the frequency of some words or symbols (number of times the WORD appears in the e-mail) / total number of words in e-mail) and the last column tells whether the email is a spam email. For example, this is the 1813th email, it has the word “make” appear 0.31 percent of time, and the word “all” did not appear. In this data set, there are 1813 spam and 2788 ham. Here is a table of the mean word frequency of some of the words among spam emails, the first row, and ham emails, the second row. An interesting finding of this data set is, null of the spam emails contain the name of the receiver, George, but on average good emails contain receiver’s name, at an average rate of 1.27 percent per email. This may because this data set was generated 10 years ago, and at that time people did not have the technology to automatically generate receivers’ names as email titles to make it sophisticated and natural. Finally, because this data set is so large, I chose R to program it.

Now let me introduce the two methods I used for this project. The first one is Generalized Linear Model logistic regression, GLM for short, which is the combination of variations of regressions. Recall Linear Regression. In simple linear regression, we predict variable y from another variable x, so y = ax + b. In more complicated linear regression, we predict one variable from some other variables. For example in our project, we predict the last variable, classification from the previous 57 word frequencies. The variables we are basing our predictions on are called predictor variables, we assign weights to the predictors to find the best-fitting line (regression line) to minimize the error, which is the distance from the points to the regression line. In this project, we want to predict the classification of the emails, which is either 0 or 1. Therefore, we use the method logistic regression. Logistic regression is a variation of linear regression, which transforms the predictions to ensure the value of predictions is between 0 and 1. Since the values are bounded by 0 and 1, we can treat this value as the probability. For example, if the point of an observation is here, has the value 0.1, we say it has a 10% chances classified as spam.

The other method is Linear Discriminant Analysis, LDA for short. The approach is to model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain the conditional probability of Y given X. Let’s first talk about Bayes theorem. Bayes theorem says that the probability of the variable Y equals k given the variable X equals x can be write as probability that X is x given Y equals k – that’s the first piece on the top there – multiplied by the marginal probability or prior probability that Y is k and then divided by the marginal probability that X equals x. Bayes theorem is the basis for discriminant analysis. We assume X and Y follow normal distributions. Then we can use Bayes theorem and our data to update the distributions and make it more accurate. In the case of discriminant analysis, things become slightly different. The probability y equals k is written as pi k. So if there are 2 classes, spam and ham, there are going to be two values for pi, pi spam is equal to 1813 over 4601 and pi ham is equal to 2788 over 4601, the probability for each of the classes. And the probability that X is x given Y equals k, can be written as a probability density function for X in class k. To estimate f k, we assume the value of X follows normal distribution. And then the marginal probability is summing over all the classes. If we only have one predictor, then probability of spam is the green curve and ham is the purple curve. Generally, discriminant analysis method try to separate the points into 2 groups, and minimize the number of points which classified wrong. Linear Discriminant analysis method separates the points with straight lines. The difference of GLM and LDA is --GLM try to fit all the points to the regression line but LDA try to separate the points using a line.

Now let’s talk about the k-fold cross validation, which is the way we used to estimate the error rate of the prediction made by both of the methods. We randomly divided the data set into 10 equal-sized parts, marked as 1 to 10. We leave out part 1 as testing data set, fit the model to the other 9 parts combined as the training data set, and then obtain predictions for the testing data set. This is called cross validation. Then we repeat this process for each of the other parts. We do cross-validation 10 times by switching around testing data set from part 1 to part 10 and treat the rest of the data sets as training data sets. Every time we record the error rate, and after we finished 10 times cross validation, we calculate the mean error rate. We can do four or five or 10-fold cross validation to avoid extreme cases. The larger k we use, the better result we can have. For example, there are four points in our data set, 7, 22, 13 and 91, if we chose 7, 22 and 13 as our training set and fit our model to it, the error will be big because 91 is too large.

Finally, after we do the 10-fold cross validation on both methods 100 times, we throw out the 2 smallest error rates and 2 largest error rates so we got the 96% CI. This process is called bootstrap. (sampling from data, do the estimation many times) The 96% CI for logistic regression is … and LDA is …. Therefore, we can conclude that LDA is a better fit for spam detection and by this method; we can filter about 90% of spam emails.