Due: Friday, January 18, by 5pm

Note: For handwritten questions, please scan (or photo-scan) and incorporate into your answer.

1. Install the "gclus" library and load the "bank" data set in R. There is a variable called `status` which indicates whether the observation (cash bill) is legal tender or counterfeit. For all models discussed below, provide proper estimates of long-run error using cross-validation!

   (a) Fit linear discriminant analysis for this data set with 'status' as the response, using all predictors except "Bottom". Provide a classification table and other metrics we have covered in lecture/lab.

   (b) Fit quadratic discriminant analysis for this data set with 'status' as the response, using all predictors except "Bottom". Provide a classification table and other metrics we have covered in lecture/lab.

   (c) Fit k-nearest neighbours ($K = 3$) for this data set with 'status' as the response, using all predictors except "Bottom". Provide a classification table and other metrics we have covered in lecture/lab.

   (d) Which of the above models would you suggest is 'best' for this data? Why?

   (e) Explore this data further. Suppose you were tasked to provide a retail chain (like Walmart) a simple test that cashiers could carry out using a highly sensitive measuring device — and note that there is a time cost associated with multiple measurements. What suggestion would you make? How did you come to this conclusion?

2. Find the "okcupidprofiles" data set on github. This is a (real and great) data set comprised of OkCupid dating profiles from the San Francisco region. Mine it for association rules and comment on which you think are interesting.
   Some notes: this is a fairly large data set, and has had minimal cleaning done to move forward with analysis. For example, there are lots of missing values depending on the variable. I will leave it fairly open-ended in terms of directions (like thresholds to use, variables to discretize or throw out, etc) to provide realistic practice.

3. Can be handwritten if preferred: Suppose that we wish to predict whether a given stock will issue a dividend this year ('Yes' or 'No') based on X, last year's percent profit. We examine a large number of companies and discover that $\bar{X} = 12$ for dividend providing companies, while the mean for those that didn't was $\bar{X} = 1.5$. In addition, the variance of X for both sets of companies was $\sigma^2 = 36$. Finally, 75% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

4. Handwritten question: Consider the model for univariate discriminant analysis discussed in class. Suppose we assume that the two groups have equal variances (and standard deviations) such that $\sigma_1 = \sigma_2 = \sigma$, but the proportion of observations (or prior probabilities) are assumed different ($\pi_1 \neq \pi_2$). It can be shown that for a new value x_0, we would predict group 1 if

$$x_0 > \frac{\mu_1 + \mu_2}{2} + c \qquad \text{if } \mu_1 - \mu_2 > 0$$

   for some c. Find c.