**Chapter 4: Classification**

* Any time a straight line is fit to a binary response that is coded as 0 or 1, in principles we can always predict p(X) <0 for some values of x and P(x) > 0 for others. To avoid this problem, we must model p(x) using a function that gives outputs between 0 and 1 for all values of X. In logistic regression, we use the logistic function.
* The quantity p(X)[1-p(X)] is called the *odds*
* By logging the odds, we get the *logodds* or *logit*
* Because the relationship between p and X is not a straight line in logistic regression, b1 (slope) does not correspond to the change in p(X) associated with a one-unit increase in X. But regardless of the value of X, a positive B1 is associated with increasing p(X), and a negative B1 is associated with a decreasing p(X).
* In terms of fitting coeffcients to the logistic regression model, we could use a number of measures – one of the more popular ones is *maximum likliehood*.
* The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for B1 and B0 (coefficients) such that the predicted probability of p(x) of default for each individual corresponds as closely as possible to an individual’s observed status. In other words, we try and find b0 and b1 such that plugging these estimates into the model for p(X) yields a number close to one for all individuals who defaulted anda number close to 0 for all individuals who did not.
* We can measure the accuracy of the coefficient estimates by computing their standard errors
* There are dangers associated with performing regressions with only a single predictor when others might also be relevant. In general, this is known as *confounding*.
* Logistic regression is primary used for response variables that have only two classes (e.g. it is binary)

*Linear discriminant analysis*

* In linear discriminant analysis, we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Baye’s therom to flip these around into estimates.
* Advantages of approaches other than logistic regression:
  + When the classes are well separated, the parameter estimates for the logistic regression model are surprising unstable
  + If the *n* is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model
  + Linear discriminant analysis is more popular when the response variable has more than 2 classes
* If pi(k) represents the overall or *prior* probability that a randomly chosen observation comes from the kth classs, then we can plug in estimates of pi(k) and f(X) into Y from the population to compute the observations that belong to the kth class.
* WE know that the Basyes classifier, which classifies an observation to the class for the which pk(X) is largest, has the lowest possible error rate out of all classifiers. Therefore, if we can find a way to estaimte fk(X), then we can develop a classifier that approximates the Bayes classifier.
* We refer to pk(x) as the *posterior* probability that an observation X = x belongs to the kth class. That is, it is the probability tat the observation belongs to the *k*th class, *given* the predictor value for that observation.
* The word linear in the classifier’s name stems from the fact that the discriminant functions are linear functions of x (as opposed to a more complex function of x).
* LDA assumes that the observations from within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes.
* In its vanilla form, the LDA does not make a distinction between the different prediction states (e.g. “Yes”, “No”). This is because the Bayes classifer works by assigning an observation to the class for which the posterior probability p(X) is greatest. E.g. in the two class case, this means a p of 0.5 or more.
* If we are more concerned about classifying some classes rather than others (e.g. we really care about identifying people at risk of default), however, we can manually adjust this threshold, e.g. set the p at 0.2 or more for one group. Note that this will tend to increase the error rate of the model overall, but the error rate for the group in particular will go down.
* The ROC curve is a popular graphic for simultaneously displaying the two types of errors (prosterior probability error and overall error rate) for all possible thresholds.
* An ideal ROC curve will hug the top left corner.
* In summation, varying the classifier threshold changes its true positive and false positive rate. These are also called the sensitivity and one minus specitificty of our classifier.

*Quadratic discriminant analysis*

* Unlike LDA, QDA assumes that each class has its own covariance matrix
* By assuming that the K classes share a common covariance matrix, the LDA model becomes linear in x, which emans that there are Kp linear coefficients to estimate. Consequently, LDA is a much less flexible classifer than QDA, and so has substantially lower variance. But if LDA”s assumption that the K classes share a common covariance matrix is badly off, than LDA can suffer from high bias. Roughly speaking, LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial. In contrast, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern.
* Since QDA assumes a quadratic decision boundary, it can accurately model a wider range of problems than can the linear methods. Though not as flexible as KNN, QDA can perform better in the presence of a limited number of training observations because it does make some assumptons about the form of the decision boundary.
* When the true decision boundaries are linear, than the LDA and logistic regression approaches will tend to preform well. When the boundaries are moderately non-linear, QDA may give better results. Finally, for much more complicated decision boundaries, a non-parametric approach such as KNN can be superior.