Gabriel West (notes for chapter 4)

Chapter 4

Lecture 1 - Classification intro

- Classification deals with discrete, unordered responses
- sometimes with some probability
- Box plots are useful for displaying quartile data, or summarizing the spread of data
- Can't just use linear regression with an encoded 0/1 response variable (because you will sometimes get values outside of [0,1]
- That's with two classes, if there are more than two classes, it gets even harrier (starts to impose order to the classes)

Lecture 2 - Logistic regression

- Logistic regression model is guaranteed to give a value between [0, 1], and is a transformation of a linear model
- log odds transformation gives the name
- Maximum likelihood is just the probability of observing the data (assuming observations are linear)
    - Maximizing this function gives you your parameters

- We can get p values, std error, and z statistics just like linear regression
- Just like with linear regression, we can encode discrete predictors
- We also get confounding effects when doing multiple regression, so it can be hard to interpret coefficients.
- scatterplot matrix is very useful for teasing out information in a dataset
- collinearity in the predictors causes variables to act as surrogates for each other (this is a recurring theme. :))
- Case control sampling results in skewed probabilities, giving us a bad constant term in the model, but this can be fixed with a transformation (using some baysian magic)
    - this wouldn't be necessary in a prospective study

- for multiple classes (more than two), just give each class its own linear model,

then weigh them agains eachother/normalize (softmax)
- This is called multinomial regression

Lecture 3 - Discriminant Analysis

- DA is informed largely by Baye's Thm.
- Classify to the highest density
- DA does better (is more stable) than log reg when the classes are well separated (huh)
- DA is more stable in the case of small n
- DA useful when >2 response *classes* because it provides low-dimensional view of data
    - Q. What does a low-dimensional view of the data mean in this case? Why is it helpful
    - A. see Fischer's d plot

- If the data is really normal, then baye's gives the best possible results.
- Plugging in the normal density function gives a simpler reduced form
- with real data you have to estimate the parameters for your density function, this can be done fairly easily for a normal
    - Q. Isn't pi_k supposed to be p(Y=k|X=x)? n_k/n is just an estimate of p(Y=y)
    - A. requires covariance matrix estimation (n^2 space complexity)

- When there are multiple predictors, X becomes a vector, and some linear algebra gets us a fairly simple equation for the discriminate equation
- Discriminant eq. is used to pick which predictor to use to classify. (pick the biggest one)
- Q. What are the drawbacks of LDA? It seems like it's just the best.
- LDA and Fischer's d plot allows us to look at the result in a k–1 dimensional space
- Easy to get probabilities out of LDA
- Q. LDA on credit data: always guessing NO gives a misclassification rate of 3.3%, so….What's going on with this dataset?
- A. Ha! They bring it up later. This is called the 'null rate', and should be compared using precision/recall
- I love confusion matrices
- you can change the threshold to find the best precision/recall
- Enter the ROC curve, which shows the two error rates (true positive/negative and false positive/negative) in one plot
- *Naive Bayes* is when you assume that the predictors are conditionally

independent (Pairwise?)

- quadratic LDA is very bad for large p
- Can do combinations and explicit quadratic fit by creating new predictors