Stat 435
Arnav Palkhiwala

1. Summary:

This week we continued with classification, focusing on conducting inference for beta with confidence intervals and significance tests. We also looked into understanding multinomial logistic regression when we have more than 2 classes to predict. From there we went into discriminant analysis and how linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) can be used for classification when logistic regression does not work. This is in cases when certain assumptions are met and proven which allows for better accuracy in some cases.

2. Concepts:

Classification

- To estimate parameters using data {xi, yi, Epsilon i} use maximum likelihood
- After derivation you get (Beta hat 0, Beta hat 1) = argmax from sum i = 1 to n, y(i) log q(x(i))/Beta + (1 - y(i)) log(q - q beta (x i))
- This can then be solved using gradient descent
- Our classifier takes a value of 1 if the value of q(x) > 1/2 , or otherwise 0
- For inference for our parameter Beta and taking the argmax of the sum of likelihood we get the fact that as n increases to infinity, Beta hat takes a Normal distribution (Beta, Fisher information matrix) using Hessian second derivative
- Then, to standardize it you can say that (beta hat - beta star)/sqrt(Fisher Information) takes a normal distribution with Mean 0 and Variance 1
- We can use this information for inference to test whether beta star = 0 and to also build a confidence interval using the traditional confidence interval formula we have used
- We used the default or not example to demonstrate this testing/calculations
- Multinomial logistic regression means there are multiple classes to pick from (1…k)
- P(Y = i | x) = e^(beta0 + beta 1 * x)/sum from j=1 to k of E^(beta 0 + beta 1*x) which is has the normalization in the denominator

Discriminant Analysis:

- Classifier(x) = argmax k P(Y=k | x) = argmax (f_k(x) pi_k)/P(x)
- We know P(x) is a constant so we are just maximizing f_k(x) pi_k
- The idea behind discriminant analysis comes from Bayes Rule and finding the proper probability
- The goal is picking the k that maximizes P(Y=k | x) and comes from making an assumption on the distribution of X|Y=k which is a normal distribution

- In Linear Discriminant Analysis LDA, you model f_k(x) as a normal distribution for every k with different means for each k but same variance for all of them
- And P(X=x | Y=k) is then the pdf of the normal distribution
- Using a log transformation we know that sigma is a constant since variance is the same so the discriminant function is argmax k -mu_k^2/2sigma^2 + x*muk/sigma^2 + log(pi_k)
- Because we only have one term of x (and it is not squared) we can see a linear relationship and there is only one single transition point x that satisfies the equation so the barrier is thus a linear line between the two sides
- For QDA you follow the same process but in this case you assume the variances are different so thus the constant is removed and at the end you have an x^2 term and thus you have two transition points/boundary points
- In practice we can estimate all of the terms used such as mu_k, sigma_k, pi_k
- Discriminant analysis is useful when classes are well separated and distribution of predictors are normal and you have more than 2 response classes

3. Uncertainties:

I feel like I understand logistic regression, LDA, and QDA very well in terms of what they are actually doing and how they work. However, I am more curious and unsure about when each one is the right choice and we did go over that in class but I was confused by how in lab logistic regression still had better results. In some of my research on the subject I found that LDA and QDA are used less than logistic regression for the most part in the real world so I am curious why LDA/QDA seem to be more stable measures. How often can we assume that our predictors are normally distributed and in the real world is this a far fetched assumption? I am also uncertain on multinomial logistic regression, I am a little confused on how exactly that works and how the probabilities are chosen since we don't have the barrier of ½. So do we just calculate all probabilities and see which is the highest?