Exam 2

Please write your answers below each question. You should not have access nor use any materials during this exam.

A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

“I have neither given nor received unauthorized aid on this test or assignment.”

1. In doing a classification task, we discussed the idea of classification and the idea of discrimination. What are these and what is the difference between the two?

These two tasks often overlap, but not always.

Classification is about developing a rule to predict which class or category an observation falls into.

Discrimination is about understanding which features separate our data into the classes.

A classification rule is often based on the features that separate the groups!

1. Suppose we have a categorical response with four levels. We could label those four levels with numeric values, say or . Explain the implications of treating our problem as a regression task with these values for . Could it ever make sense to do this?

If we did this, it would imply an ordering of the levels of the response and a magnitude for their differences. We model the mean response. This does not make sense if we have a nominal variable. This is a categorical variable with no ordering to the categories.

This could be applied to an ordinal categorical response variable. However, this has big implications. This would imply that the difference between level and is the same as the difference between and , etc. Similarly, the difference between and would need to be the same as and , etc. In a case where that is reasonable, the regression model may do a ok job.

1. Select true or false for each classification method.
   1. We can never use the Bayes classifier in a real scenario. (TRUE)
   2. LDA is a special case of QDA (TRUE)
   3. Logistic Regression provides a discriminant for classifying our observations (TRUE)
   4. Binary logistic regression generally requires a larger sample size than multinomial logistic regression. (FALSE)
2. We discussed the idea of the Bayes’ error rate. Can we ever do better than this rate? Explain.

No, the Bayes’ error rate is the optimal rate we could have. As the data has randomness, we have no way of finding a classifier that will have 100% accuracy. The (true) conditional distributions can be used to determine the ‘best’ classification but even this will produce misclassifications. The Bayes’ error rate is this ‘best’ rate.

1. One measure of the quality of a classification model is accuracy. Define the no information rate and describe how interepreting the accuracy of a model is related.

The no information rate is the accuracy we’d obtain by simply using the most prevalent class as our classification. Any model we would consider useful should have accuracy higher than the no information rate.

1. Define the terms sensitivity and specificity.

Sensitivity is the true positive rate. That is, the number of positive cases correctly classified divided by the total number of positive cases.

Specificity is the true negative rate. That is, the number of negative cases correctly classified divided by the total number of negative cases.

1. When using a generative model for classification, we need to estimate the *prior probabilities* for each class. What is the most basic way we discussed for estimating these probabilities?

Using the prevlance of each class in the data to estimate these. That is, we use the proportion of each class as the estimated probability.

1. Suppose we have a categorical response with categories and a single predictor variable . When fitting an LDA model, we use normal distributions. What quantities do we model with a Normal distribution? Are those normal distributions related in anyway?

We model the distribution of , , …, each with normal distributions. The means are assumed to be different but the variances are all assumed to be the same.

1. When trying to use LDA or QDA with predictors, we can note that LDA is a special case of QDA. Why might we still prefer LDA to QDA even though QDA is more general?

In using QDA we have to estimate a variance/covariance matrix for each class. This is a large number of parameters! As such, we would need a lot of data to have reasonably stable estimates. LDA assumes a common variance/covariance matrix and would therefore not require as many observations to have a reasonably stable fit.

1. We discussed the Naive Bayes classifier. This is a generative model. What simplifying assumption do we make when using the Naive Bayes classifier?

We assume that the (joint) distribution of our ’s given our response is the product of (marginal) individual distributions of each given the response (i.e. we assume conditional independence). That is, we simply model each distribution separately and multiply them together to get our distribution of .

1. What is the difference between a cubic spline model and a natural cubic spline model?

A natural cubic spline model ensure that the boundaries of our model are linear rather than cubic. The two extra degrees of freedom from each boundary region are then added as additional knots. This tends to produce models with more stable estimates at the boundaries.

1. Suppose we have data on whether or not someone has heart disease (No = 0, Yes = 1) and a number of predictors such as Age (quantitative), ExerciseAngina (Y or N), and Cholesterol (quantitative). We fit a logistic regression model with ‘main effects’ for each of these predictors. Relevant output is given below.

|  | Estimate | Std. Error | z value | Pr(>|z|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -4.4039 | 0.6501 | -6.7742 | 0.0000 |
| Age | 0.0530 | 0.0100 | 5.2905 | 0.0000 |
| ExerciseAnginaY | 2.4644 | 0.1925 | 12.8046 | 0.0000 |
| Cholesterol | 0.0024 | 0.0015 | 1.6052 | 0.1085 |

1. What is the fitted equation for those without Exercise Angina? Be careful how you write the left hand side of the model! No need to simplify.
2. How would we use this fitted equation to find a decision boundary for those without exercise angina? This isn’t something you can solve! Just write down how you would use the equation to find the boundary for values of Age and Cholesterol.

As a 0 log-odds corresponds to the probability of success and probability of failure being the same we can set this estimated log odds to 0 and form an equation for the decision boundary in terms of Age and Cholestrol.

1. How do we interpret the meaning of the intercept coefficient for this model? Be sure to use the context of the data.

The intercept of -4.4039 represents the log odds of having heart disease for those with an Age of 0, no exercise angina, and a cholesterol of 0. Very unlikely to have heart disease but also not meaningful in the context of this problem!

1. How do we interpret the meaning of the age slope coefficient for this model? Be sure to use the context of the data.

The Age slope of 0.0530 represents the change in log odds of heart disease for increasing age by one unit, while holding cholesterol and exercise angina status constant. This implies that your odds of heart disease increases as you get older.

1. How do we interpret the meaning of the ExerciseAnginaY coefficient for this model? Be sure to use the context of the data.

The exercise angina coefficient of 2.4644 represents the change in log odds of heart disease as compared to someone without exercise angina, while holding Age and Cholesterol constant. This positive value indicates that those with exercise angina are much more likely to have heart disease.