ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Zach Ginder

Date of Exam: Thursday, March 6, 2025 - Friday, March 7, 2025
Time Limit: 75 minutes
Allowed Materials: None (closed book & closed notes)

Student – NC State University Pack Pledge

I, Zach Ginder have neither given nor received unauthorized aid on this exam or assignment. I have read the instructions and acknowledge that this is the correct exam.

Student's Printed Name

Student Signature 03/06/2025 *Date*

Exam must be turned in by: 9:30 *Exam End Time* ZG *Student's Initial Agreement*

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

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? (8 pts)

Classification involves determing a rule to place observations into groups/classes based on predictors. The classes are the response variable.

Discrimination involves determining the predictors that best classify the data. Many times these overlap but discrimination is more of the variable selection portion of classifying a dataset

2. Suppose we have a categorical response with four levels. We could label those four levels with numeric values, say $Y = 1,2,3,$ or 4. Explain the implications of treating our problem as a regression task with these values for Y. Could it ever make sense to do this? (6 pts)

The problem when treating this as a regression task is that in regression we are modeling the mean response. This can take values less than zero and greater than one. Since in a classification setting we want the $p(Y=j|X)$ this value should be between zero and 1, which is not guaranteed in regression. It normally does not make sense to do this

It is normally unlikely to have the Bayes classifier in a real scenario because we would need $P(Y|X)$ which we normally do not know

3. Select true or false for each classification method. (3 pts each)
   a) We can never use the Bayes classifier in a real scenario. True
   b) LDA is a special case of QDA. True
   c) Logistic Regression provides a discriminant for classifying our observations. True
   d) Binary logistic regression generally requires a larger sample size than multinomial logistic regression. False
4. We discussed the idea of the Bayes' error rate. Can we ever do better than this rate? Explain. (5 pts)

the Bayes error rate is the theoretical error rate even in the Bayes classifier scenario (overall variability). we normally cannot do better than this rate as this is the error rate we would have if we had the Bayes classifier. This is what we are trying to determine (get as close as possible too) so our model will include the Bayes error rate combined with our model error rate

5. One measure of the quality of a classification model is accuracy. Define the no information rate and describe how interpreting the accuracy of a model is related. (6 pts)

The no information rate is the rate at which we just classify all new observations into the most prevalent class. So if class A has 80% of observations/prevalence. If we placed all new obs into class A we likely only placed 80% correctly (in theory) thus we try to develop a model that improves on this rate to prove the classification rule is better than the no information rate

6. Define the terms sensitivity and specificity. (6 pts)

Sensitivity and specificity are used widely in things like pharmaceutical testing (think covid testing)

Sensitivity is the probability of a test showing positive or "success" when it is a positive/success

Specificity is the probability of a test showing negative or "failure" when it is a negative/failure

7. 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? (6 pts)

The most basic way to do this is take the number of existing observation classified in each class and divide by the total number of observations. For example the prior probability for class 1: $P(Y=1) = \frac{\#\text{ of observations in class 1}}{\text{total \# of observations}}$

8. Suppose we have a categorical response with m categories and a single predictor variable X. 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? (6 pts)

The quantities that we model with the Normal distribution are the densities $f_{X|Y}(x|y)$ which are each normal distributions with their own mean (for each class $Y=j$) but in LDA they are assumed to have equal variances.

9. When trying to use LDA or QDA with $p = 10$ 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? (6 pts)

We might prefer to use LDA if we have a small number of observations in our training dataset.

10. We discussed the Naive Bayes classifier. This is a generative model. What simplifying assumption do we make when using the Naive Bayes classifier? (6 pts)

We assume that the predictors X_i's are independent within each class. More simply:

$$f_{X|Y}(x|y) = f_{X_1|Y}(x_1|y) \cdot f_{X_2|Y}(x_2|y) \cdot \ldots \cdot f_{X_p|Y}(x_p|y)$$

11. What is the difference between a cubic spline model and a natural cubic spline model? (6 pts)

A natural cubic spline forces the regions on the perimeter/ends to be fitted using a linear function as opposed to every knot having a cubic function. This is because the ends tend to have more variability so a natural cubic spline tries to account for this by making the ends linear.

12. 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

a) 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. (6 pts)

$$\ln\left(\frac{P(\text{no exercise Angina})}{1-P(\text{no exercise Angina})}\right) = -4.4039 + 0.0530(\text{Age}) + 0.0024(\text{cholesteral})$$

b) 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. (6 pts)

For each predictor (Age and cholesteral) we would want to determine where P(no exercise Angina) is greater than 0.5. This is where we would have the boundary and would classify as "no exercise Angina". When P(no exercise Angina) < 0.5 we would classify as "exercise Angina".

c) How do we interpret the meaning of the intercept coefficient for this model? Be sure to use the context of the data. (5 pts)

the log odds of having no exercise Angina for someone of Age=0 and cholesteral=0 is

-4.4039

d) How do we interpret the meaning of the age slope coefficient for this model? Be sure to use the context of the data. (5 pts)

For every one unit increase in Age, the log odds for someone not having exercise Angina increases by 0.0530

e) How do we interpret the meaning of the ExerciseAnginaY coefficient for this model? Be sure to use the context of the data. (5 pts)

This is an indicator variable, therefore if someone does have exercise Angina the intercept for the log odds function of having exercise Angina changes from -4.4039 to

$-4.4039+2.4044$.