ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Jarrett Glass

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, Jarrett Glass 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 06Mar2025 *Date*

Exam must be turned in by: _____

Exam End Time *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 is the task of predicting/assigning the category that an item goes into based on other values it has.
Discrimination is the determination of where those decision boundaries fall, i.e. order to perform the classification.

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)

with categorical responses, the intended result is a probability (or odds, log-odds, etc) that a response would go into a specific level. Any result would need to fall between $[0, 1]$.

With a numeric regression, the response could be negative or exceed 1. It could make sense to do this for a strictly binary case – or at least, it may provide some results that appear meaningful. But generally, treating the problem as a categorical task is needed.

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 - the true distribution unknown.
   b) LDA is a special case of QDA. True
   c) Logistic Regression provides a discriminant for classifying our observations. False
   d) Binary logistic regression generally requires a larger sample size than multinomial logistic regression. False - More parameters requires more data.

4. We discussed the idea of the Bayes' error rate. Can we ever do better than this rate? Explain. (5 pts)

No, because this is an error rate that is inherent to the data and not to the model.

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 NIR describes how much you'd be right if no data were known — example, if you had 4 options but picked the same one every time, you could be correct about 0.25 of the time; so the NIR there is 0.25. The accuracy of a model should be better than the NIR.

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

Sensitivity is a measure of how prone a model is to noise (eg a measure of variability)

Specificity is a measure of how prone a model is to accuracy (or, a measure of bias).

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)

To assume the probability of each class is Normal distributed, and to find the mean & s.d. for each class.

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)

What is modeled as Normal are the posterior probabilities inherent in each category of the response. They are related only in that the sum of their probability should add to one.

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)

The QDA is more complex than the LDA, and its flexibility will only increase further as the number of predictors increases. Increased flexibility also means increased chance of bias due to overfitting, so the LDA may indicate lower test error at $p=10$.

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)

That we have/know the true distributions for the parameters & components that we are working with, and use Normal priors.

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

A natural cubic spline is continuous - the graph generated makes contact between each section, with no large jumps or discontinuities. A cubic spline model does not, and there may be such gaps.

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)

$$\ln\left(\frac{P(Y=1 \mid \text{ExerciseAngina}=N)}{1-P(Y=1 \mid \text{ExerciseAngina}=N)}\right) = -4.4039 + 0.053\,\text{Age} + 0.0024\,\text{Cholesterol}$$

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)

The decision boundaries exist where the probability is even,

$$P(Y=1 \mid \text{Exercise} = \text{no}) = P(Y=0 \mid \text{Exercise} = \text{no}) = 0.50.$$

When $P = 0.50$, the odds $= 1$, and log-odds $= 0$.

So finding this decision boundary involves setting the above equation to zero,

$$-4.4039 + 0.053\,\text{Age} + 0.0024\,\text{Cholesterol} = 0,$$

then finding the values of Age & Cholesterol that make this statement true.

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 intercept -4.4035 means that if age and cholesterol were both zero (not realistic), and no exercise angina occurred, the log-odds of having heart disease would be -4.4035. That's theoretical since age & cholesterol wouldn't ever really be zero, but that's the interpretation of it.

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)

The age slope coefficient 0.053 means that for every unit increase of a patient's age, their log-odds of having heart disease increases by 0.053 - or that their odds of it increases by a factor of 0.053.

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)

The Exercise Angina coefficient, 2.4644, means that If a patient does have Exercise Angina, ~~the~~ the log-odds of them also having heart disease increases by 2.4644, or that the odds of it increases by a factor of 2.4644.