ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Naman Pujani

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, Naman Pujani 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 *DATE* 3/7/25

Exam must be turned in by: 1:15 NP

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 when we develop some sort of rule to allocate new objects in a class of the response. Discrimination is when we find predictors that split our data into groups.

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)

It won't make sense to do this because in a regression setting, the response regression might just be equal to the mean. What this will do is provide us with probabilities that are below 0 or above 1, which isn't useful for any analysis.

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. False
   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)

   Bayes error rate is the misclassification error rate when we use a Bayes classifier. Minimizing the test error is the best way we can classify our observations or we can try to maximize the test accuracy.

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 is the % of observations in the most prevalent class. They're related because accuracy helps us find out the number of correctly classified observations in total. The NIR gives us the percent in each class.

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

   Sensitivity & specificity are basic measures of our classifications. Sensitivity is the # of correctly classified obs ($\frac{TP}{TP+FN}$) so truly positive over truly pos. plus falsely negative. Specificity is the incorrectly classified observations ($\frac{FN}{TP+FN}$) so falsely negative over falsely negative plus truly positive. We can get these measures from the diagonals and off diagonals of a confusion matrix.

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) We talked about simply taking the proportion: $\frac{n_k}{n}$, where n_k is the number of observations in a given class k.

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) We model $X|Y$ for different categories of m using a normal distribution. These normal distributions have different means for the different distributions but the same variances across the normal distributions.

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) LDA isn't as flexible as QDA and due to that it will have lower variance but a higher bias. This is preffered because our goal is to reduce variance.

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) The assumption we make is the joint probability distribution is the product of the marginal distributions. This assumes that the marginal distributions are all independent of each other.

11. What is the difference between a cubic spline model and a natural cubic spline model? (6 pts) A cubic spline model fits the data with a cubic and ensures smoothness everywhere except past the boundaries, where it's erratic. A natural cubic spline model imposes the condition that the model be linear past the boundaries to account for the erratic behavior.

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)

$$\log(\pi(x_i)) = \frac{e^{-4.4039 + .0530\beta_1 + 2.4644\beta_2 + .0024\beta_3}}{1 + e^{-4.4039 + .0530\beta_1 + 2.4644\beta_2 + .0024\beta_3}}$$

where $\beta_2 \rightarrow \begin{cases} 1 & \text{if exercise angina is present} \\ 0 & \text{o.w.} \end{cases}$

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)

Since ExerciseAngina is an indicator variable, to find the decision boundary for those without exercise angina, we'd plug in a value of 0 into the model which would indicate a value of No for 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)

β_0 represents the log odds for someone who is 0 years old, no exercise angina, and has a cholesterol level of 0, which isn't useful for analysis.

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 slope coefficient age represents the increase in log odds for a unit increase in X holding the other predictors constant.

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 coefficient for exercise angina represents the log odds when a person does have exercise angina. Otherwise, the value would be 0 since it's an indicator variable