# ST 563 601 – SPRING 2025 – POST
# Exam #2

**Student's Name:** Sanith Rao

**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, Sanith Rao 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*

Sanith Rao

*Student Signature* *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)

In ML, we basically have regression tasks and classification tasks. Classification tasks are those where we are given a qualitative variable and predictors and we want to determine what class it belongs. Discrimination on the other hand is basically used to figure out how we place data points into each class. This function is usually optimized to fit as many data points into their correct class as possible.

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)

Unfortunately, this problem cannot be treated like a regression task. This is because all the 4 values are qualitative and not quantitative in that they are essentially just labels and that we don't care about them arithmetically. Regression basically find's the expected value and so it will assume the difference between 1 and 2 is the same as the difference between 2 and 3 and so on which we have no evidence of being true. One case where regression could be used is for binary classification since we don't deal with multiple levels and the task becomes much simpler.

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)

Bayes Error Rate is a theoretical value that is the minimum error rate that a particular model can have. It requires knowing all the conditional probabilities $Y|X$ from the data which is not possible in a real life problem. Thus this is the most optimized model's error and is impossible to beat.

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)

No information rate is basically the percentage of the number of rightly classified observations. given that we has no prior information about the data. This essentially tells us how good our model is at classifying our data

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

$$\text{Sensitivity} = \frac{\text{True Positive Rate}}{\text{True Positive Rate} + \text{False Negative Rate}}$$

$$\text{Specificity: } \frac{\text{True Negative Rate}}{\text{True negative rate} + \text{False positive Rate}}$$

Sensitivity basically tells us out of all the true predicted values, how many were actually true, specificity tells us out of all the false predicted values, how many were actually false

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 we can do this is by using the pdf and then finding the probability at a particular for all the data points in our given data.

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)

when fitting an LDA model, we use a normal distribution where we try to find the mean (for each class) and variance (same for all classes). Here basically model conditional probabilities $X|y=1, X|y=2 \ldots X|y=m$. All distribution have the same variance.

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)

QDA is more general and what this means is that is is more flexible. When a model is more flexible, it makes it more prone to overfitting. This is especially true when we don't have sufficient data points for the problem. In situations like, LDA is definitely preferred.

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)

Naive Bayes classifier essentially uses the Bayes theorem to find the conditional probabilities of an observation belonging to a particular class. The main difference in assumption we make here is that all the features are independent of each other for every class. This eliminates the need for a covariance matrix which is used in LDA and QDA.

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

Cubic spline and natural cubic spline are both piecewise regression models that ensure smoothness at the ends. Only difference is that the natural cubic spline requires the second derivative at the end points to be 0 to ensure there are no weird curves. No such restriction for cubic spline. Cubic is more flexible than natural.

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)

$$\log\left(\frac{P(\text{heart disease})}{1 - P(\text{heart disease})}\right) = -4.4039 + 0.053 x_1 + 0.0024 x_2 + 2.4644 x_3$$

$x_1$ - Age

$x_2$ - Cholesterol $x_3$ - Exercise angina -

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 equation basically finds the log odds that for heart disease given they don't have exercise angina. We basically want to set this equation to 0 to find our decision boundary because that is exactly the point where the log odds will be equal for both classes. This is used to draw the decision boundary and separate different classes. We basically do this using the equation above since here also we have no 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)

Basically the intercept can be interpreted as log odds of success (in this case heart disease) when all the parameters - age, exercise angina, and cholesterol values are 0.

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 coefficient is basically the change in log odds of success for a 1 unit change in the age keeping all 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 exercise coefficient of ExerciseAnginaY is basically the change in the log odds of success when a person has exercise angina and the person does not have exercise angina keeping all the other factors constant.