44

ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Constantino Raptis

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, Constantino Raptis 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* 3/7/2025

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."

does both

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)

-6 not clear, not a general answer

For example when discussing about classification we can analyze the logistic regression which starts from computing the log-odds and then convert to probability and then if $P(Y=1) > 0.5$ it classifies as 1 (success) otherwise 0 (Failure). Idea of discrimination we use LDA / QDA where we use linear or quadratic function giving us a linear or quadratic boundary, we assign the highest $\hat{\delta}_k x$ for each class.

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)

No because a categorical response should be in a binary (0 or 1). Having 4 levels in a regression task would be very difficult to classify those as categorical response.

-5

-11

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 -3
   d) Binary logistic regression generally requires a larger sample size than multinomial logistic regression. True -3

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

No, in practice we want to find the lowest Bayes' error rate for the model. -2

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) -2

Accuracy goes for $\frac{\#\text{ of correct classifications}}{\text{Total }\#}$

NIR provides the rate of classifications that we don't know. -3

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

Sensitivity and specificity is about the true positives, False negatives, False positives and True negatives, and How well the model performs with those two metrics. I don't remember the formulas. -2

-15

$$p = \frac{1+\text{odds}}{\text{odds}} \qquad \text{odds} = \frac{1+p}{p}$$

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)

-6

On generative models we have (LDA, QDA and Naive Bayes). it (log-odds > 0, it classifies as 1 (success) otherwise 0 Failure, for Probability > 0.5, it classifies as 1 (success), otherwise Failure 0.

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)

Model X|y uses normal distribution and for y=m each normal distribution has a mean and variance. — same

different

-2

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 still use LDA due to the fact that it will fit better the data with linear boundary. QDA used for complex data will give a more flexible quadratic boundary, and in this case will not represent well the data with p=10.

maybe.

-5

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)

When using the Naive Bayes classifier we use the assumption that the predictors are independent.

X|y are

-1

-14

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

This will touch the way that knots and degrees of freedom are defined. -4

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)

Y(whether someone has heart disease) = -4.4039 + 0.0530(Age) + 0.0024(Cholesterol) -2

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)

We would use LDA to find a linear boundary using a linear function. Then we would assign each observation to the class with the highest $\delta_k x$. -6

-12

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)

$\hat{\beta}_0$ -> is the case that all indicators are 0, which in our case is -4.4039. This means that the chance of having a heart disease is very small. ok

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)

$\hat{\beta}_1 = 0.0530$. This value of the age slope coefficient will increase the age meaning that a higher age will increase the chance of having a heart disease. -1 with others held constant -1

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)

$\hat{\beta}_2 = 2.4644$. This value of Exercise Angina coefficient works in the sense that if the individual Exercise Angina (yes = 1) will increase the chance of having a heart disease. otherwise if the person doesn't Exercise this value will be 0 and the chance of having a heart disease will not increase related to Exercise Angina. -1 log odds by 2.464 -1 others held constant -1

-4