87

ST 563 601 – SPRING 2025 – POST
Exam #2

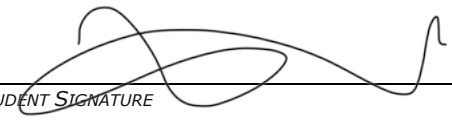
Student's Name: Julia Fish

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, Julia Fish 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 3/7/25 *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 focuses mainly on producing a value (or class) to a new observation based on fixed predictors. Discrimination provides a density function as well as a classification (i.e. a formula in which to classify that is estimated using Bayes' Theorem).

almost −2

Log Reg also does discriminants

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)

There could be a scenario in which doing this isn't detrimental, however there are 2 main issues with this. The model that treats these values numerically is estimating an average, long-run prediction instead of as one new observation. This means that $\hat{y}$ values in between these discrete values is very probable when the desire to predict the specific class of this observation should not generate them. In addition, values below 1 and above 6 could be possible in the model when it is not possible for the data itself.

oh ok

ok

other implications − 1

when could it make sense? − 1

−4

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. False

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

   No. The Bayes' error rate is essentially equivalent to the irreducible error rate. We cannot change anything in our model to systematically reduce this by definition. ok

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 accuracy when all observations are predicted to be in the class with the highest proportion of data no matter what. Interpreting accuracy of a model is related to this because all models with more specific classification "rules" should perform better than this baseline.

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

   Sensitivity and specificity are different metrics in which to capture model performance.

   The sensitivity of a certain class is the rate of true positives versus total observed for that class. ?

   The specificity of a certain class is the rate of true negatives versus total negatives for that class.

   oh I see

-3

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 For LDA and QDA, ~~we~~ discussed assuming normality for these densities.

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 model $X|Y=1$, $X|Y=2$, ..., $X|Y=m$. These models are assumed to have different means but all share the same variance-~~covariance~~ matrix. one pred only

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 requires much less estimation of parameters. Because of that, the sample size does not need to be huge in order to estimate all of those. With p=10 (not small), that is especially the case.

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 make the simplifying assumption that the joint conditional density is ~~additive~~ in the marginal conditional densities.

multiplicative ok

-6

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

A cubic spline is a fit that is continuous at the knots. ← df = M+4
A natural cubic spline is a cubic spline that also takes the property that the model is linear beyond the knots on either side. ↖ df = M

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)

$$\ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = -4.4039 + 0.0530(\text{age}) + 0.0024(\text{cholesterol})$$

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 boundary is when the log odds $\ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = 0$, or when $P(Y=1|X) = P(Y=0|X) = .5$.

Solve for the decision boundary by setting the right hand side of the equation in part a) to 0.

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 coefficient for this model is the estimated log odds of having heart disease for someone of someone without exercise angina with age 0 and cholesterol 0 (not meaningful/attainable in value).

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 estimated log odds of having heart disease are estimated to increase by around 0.0530 for any one year increase in age for a fixed level of cholesterol as well as a fixed exercise angina status.

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 estimated log odds of having heart disease are expected to increase by around 2.4644 for someone with a fixed age and cholesterol level that is diagnosed with exercise angina.