



DELTA Testing Services

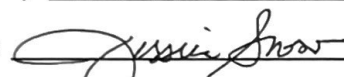
Student Name: Alex Devoid Date: 3/6/25

Student's NCSU Email Address: adevoid@ncsu.edu

Course: ST 563 601 Exam #: 2

Start Time: 10:20 am End Time: 11:35 am

Proctor's Name (Print): Jessica Snow

Proctor's Signature: 

Institution: Southwestern Community College

PLEASE SIGN & DATE THIS SHEET AND RETURN ALONG WITH THE EXAM

Proctoring Guidelines

If you are unable to comply with the following, please destroy the exam and have the student submit the name of another proctor for approval.

1. Please ask student for their photo ID.
2. **Have the student put their name on the exam and exam answer sheet.**
3. The test should be conducted in an atmosphere conducive to good concentration (quiet, good lighting, etc.).
4. The student must take the exam without outside help. Have the students leave all materials (except blank paper, pen or pencil, or calculator, as needed) outside the testing room. This includes notes, books, calculators, phones, etc. (excluding materials required for the exam).
5. Close and constant supervision must be provided.
6. Please scan and email the proctoring form, completed exam, and any formula sheets permitted for the assessment to delta-testing@ncsu.edu or fax to 919-515-7180.
7. Not including exams that permit all notes or textbooks, students should not be permitted to leave the testing room with formula sheets or scrap paper unless explicitly stated.
8. **DO NOT GIVE THE EXAM TO THE STUDENT TO MAIL BACK**

If you have any questions, please contact DELTA Testing Services at our main Venture IV location via phone: (919)-515-1560 or e-mail: delta-testing@ncsu.edu.

Thank you for assisting our students.

DELTA Testing Services

NC State University

ST 563 601 – SPRING 2025 – POST Exam #2

Student's Name:

Alexander Devold

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,

Alexander Devold

STUDENT'S PRINTED NAME

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 SIGNATURE

[Handwritten Signature]

DATE

3-6-25

Exam must be turned in by:

11:35
EXAM END TIME

A D

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)

discrimination: Knn and logistic regression practice discrimination directly modeling $P(Y|X)$.

classification: QDA and LDA on the other hand model the probability distributions of each class first and then compute the probability of Y given X using the bayes theorem $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

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)

We want to model the probability of each class. These probability values ^{must} fall between 0 and 1. If we treated our problem as a regression task we may end up with values greater than 1 or less than zero

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) *No*

Bayes minimizes the probability of misclassification. It is the optimal classification probability

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) Interpreting the accuracy of a model can be misleading. Say, for example, fraud happens in only 2% of transactions. If our model predicted no fraud 100% of the time, it would be accurate 98% of the time. But it would be useless for detecting fraud. This is why we look at the positives and true negatives in a confusion matrix.

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

In the bias variance trade off

Sensitivity relates to variance, How much noise the model will pick up.

Specificity relates to bias, How rigid the model is

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)

Compute using your existing data =
$$\frac{\text{class 1}}{\text{class 1} + \text{class 2}}$$

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 the conditional distributions. LDA assumes each of these normal distributions has the same variance but different means.

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 assumes that the distribution of each ^{conditional on} distribution is normal and that each have different variance and means. If you know the classes have shared covariance, then you will want to use LDA.

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)

All predictors are independent of each other, which is usually unrealistic in practice.

- 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)

This is the baseline log-odds if all predictors are zero.

The baseline log odds when all x_s are zero is -4.4039

- 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)

This is the change in log-odds with one unit change in age.

log-odds change by 0.0530 each year in age.

- 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)

log-odds of 2.46 is given to the result if a person does have Exercise Angina

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

wild things can happen at the edges of cubic splines, while natural cubic splines normalize or smooth out these transitions

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)

$$P(Y|X) = B_0 + x e^{B_1} \dots x e^{B_3} + \epsilon$$

$$P(\text{heart disease} | X) = -4.4039 + x e^{0.0530} + x e^{2.4644} + x e^{0.0024} + \epsilon$$

- 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 use a discriminant function to directly model $P(Y|X)$ for those without exercise angina. $\text{ExerciseAnginaY} = 0$ for those without it.

$$P(\text{heart disease} | X) = -4.4039 + x e^{0.0530} + 0 e^{2.4644} + x e^{0.0024}$$