DELTA Testing Services

go.ncsu.edu/testing

Campus Box 7555
1730 Varsity Dr.
Venture IV, Suite 236
Raleigh, NC 27695-7113

919.515.1560 phone
919.515.7180 fax

delta-testing@ncsu.edu

NC STATE
UNIVERSITY

DELTA Testing Services

Student Name: Matthieu Cartron **Date:** 3/6/25

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

Course: ST 563 601 **Exam #:** 2

Start Time: 1131 am **End Time:** 12:46 pm

Proctor's Name (Print): Billha Lucero

Proctor's Signature: Billha Lucero

Institution: University of New Mexico Testing + Training Ctr.

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

Updated March 2022

ST 563 601 – SPRING 2025 – POST

Exam #2

Student's Name: Matthieu Cartron

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, Matthieu Cartron 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

3/6/2025
DATE

Exam must be turned in by: 11:46am
EXAM END TIME

MC
STUDENT'S INITIAL AGREEMENT

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

Matthiew Carton

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:

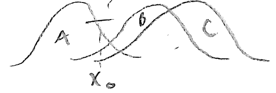
"I have neither given nor received unauthorized aid on this test or assignment."

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)

With classification, we model class membership probabilities and seek to assign the most probable class to a new observation (in the case of prediction)

With discrimination, however, we derive discriminant functions for each class and assign class membership based on the class whose discriminant function is largest.

linear combination of predictor likelihoods and priors



eg point where discrim function for class A is larger than clas B&C for a given X_0.

$P(C_1 | x_1, \dots, x_p)$ multivar $N(\mu_m \Sigma_m)$

$P(C_m | x_1, \dots, x_p)$ Multivar $N(\mu_m, \Sigma_m)$

for m classes and p preds

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)

Generally this approach does not make sense becouse, with regression, we are modeling the mean response $E(Y|X)$ when, with classification, we are modeling class probabilities. Furthermore, treating this problem as a regression task may lead to extrapolation (negative values for Y, etc.)

This approach might not be useless if we have a response w/ categories with quantitative information. For example, maybe we have 10 different weight groups, each with observations having a certain equal weight interval (eg 101-110, 111-120, etc.). Here the mean response would make more sense, and we would not be as worried about extrapolation.

as opposed to an indicator variable, zip code, etc.

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)

The Bayes' error rate is the classification analog to regression's irreducible error. Even if we knew $P(Y|X)$, e.g the full conditional distribution of Y given X (which we can never know in practice), there will still be some inherent classification error (that inheres in the data generating process — Y and X are random variables, after all).

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 baseline accuracy that we would get by just assuming that all observations are assigned to the most prevalent class. Accuracy is the number of correctly classified observations over the total number of classified observations.
(1 - misclassification rate)

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

Sensitivity - The true positive rate, the number of Correctly guessed "successes" over the number of predicted successes 50/52 from the example Confusion matrix

True positive

	1	2
Actual	50	2
Not actual	2	30

False negative

False positive

True negative

Specificity - The true negative rate, the number of correctly classified failures over the number of predicted failures.

30/32 from the example 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)

For the prior probabilities, we can simply use the sample proportions in our data. For example, if a class has 30 members and $n = 100$, our prior for this class can be 0.3.

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 means and variances with these normal distributions. In the case of LDA, we assume that these normal distributions have the same variance.

(same Σ across $P(Y|X_1, \dots P(Y|X_p))$

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 have to estimate the additional variances

QDA does not handle high-dimensionality as well as LDA because we have more parameters to estimate (Covariance Σ matrices are allowed to vary across $P(Y|X_1), \dots, P(Y|X_p)$) and thus need a much lager sample size for the model to explore the predictor space.

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)

We assume that all of our predictors are independent

=> We can model the joint distribution of X by the product of the marginals, which simplifies (greatly) our model.

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

A natural cubic spline model is a cubic spline model but with linear fits at the boundaries to provide more stable estimates in these regions (and has the effect of reducing df by 2).

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)

$$\log\left(\frac{p}{1-p}\right) = -4.40 + 0.05X_{Age} + 2.464\,I(X)_{\text{Exercise Angina}} + 0.0024\,X_{\text{cholestrol}}$$

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)

$$\longrightarrow \quad 0 = -4.40 + 0.05X_{Age} + \underset{\substack{\uparrow \\ \text{no exercise angina}}}{0} + 0.0024X_{\text{cholestrol}}$$

When log-odds is set to zero, we get the point where P(success), in this case heart disease, equals 50%. This is our estimated decision boundary - above 50% yields a classification of heart disease, and below 50% (negative log odds) yields a no heart disease classification.

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)

when age = 0, cholesterol = 0, and there is no exercise angina, the log-odds of a person having heart disease is -4.4039, which is well below a 50% probability of having heart disease.

We can't read too much into the intercept alone. For example, in practice, age and cholesterol will never equal zero.

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)

Holding cholesterol and exercise angina constant, a 1-year (unit) change in age will produce a 0.0530 change in the log-odds of having heart disease.

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)

This is an indicator variable. Holding age and cholesterol constant, if a person has exercise angina, this results in an increase in the log-odds of having heart disease by 2.4644 vs those who do not have exercise angina.