ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Lilian Mgonadi

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, Lilian Mgonadi, 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*

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 classification task we develop a rule that assign new observation to the class of the response variable while discrimination find predictor that split the data into groups.

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)

it is expected that the probabilities should be between 0 and 1 but when treated as a regression task we can get probabilities that are below 0 or more than 1 and this is inappropriate if doesn't make sense to do this because it gives us a result

that is not reliable and cannot be used to draw reasonable and appropriate conclusions.

(a) The difference between the two is that we are using a classification rule while the discrimination isn't developing any rule but actually splitting the data into groups.

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 misclassification error. We can do better than this rate if we have a higher accuracy.

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 proportion of observations that has the most prevalent cases. Check other page

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

Sensitivity is the number of positive cases correctly classified / Total no of positive cases ie

$$\text{Sensitivity} = \frac{\text{Total no of positive cases correctly classified}}{\text{Total no of positive cases}}$$

Specificity is the number of negative cases Correctly classified / Total no of negative cases

$$\text{Specificity} = \frac{\text{Total no of negative cases Correctly classified}}{\text{Total no of negative cases}}$$

5 Cont'd

Accuracy is the total number of observations Correctly classified over the total number of observations. If we have eg 80% accuracy in the model that means the NIR will be 80% since it has the higher prevalent rate in the model.

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)

We can estimate the prior probabilities via the relative frequency $\frac{n_k}{n}$ No of observation in k class / no of observation.

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 distribution of $X|Y$ and then obtain the mean and variance. Here we have 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)

We might prefer LDA to QDA because LDA can still work even when the distribution is not normal. It will give a better fit than the QDA

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 assume that the conditional probability distribution is independent

8 Contd

Yes, they are related only with the same variance but they have different means

$$x \mid y = 1 \sim N(\mu_1, \delta^2)$$

$$\vdots$$

$$x \mid y = m \sim N(\mu_m, \delta^2)$$

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

A cubic spline model is more smooth compared to a natural cubic spline model. The spline is descontinous

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)

The fitted equation is given in the next page

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 will assign new observations to class with the higher probability values eg if $P(Y=1|x) > P(Y=0|x)$ we will assign new observation to class with $P(Y=1|x)$

12a) for those without exercise angina

$$P(Y=1|X) = \frac{e^{-4.4039 + 0.0530\,Age + 0.0024\,Cholest}}{1 + e^{-4.4039 + 0.0530\,Age + 0.0024\,Cholest}}$$

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)

it is the log odds that someone has a heart disease when all the predictor variables are 0. That is age, exercise angina and cholesterol

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)

it is the change in the log odds that someone has a heart disease when there is a unit increase 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)

it is the change in the log odds that someone has a heart disease when there is increase in the exercise AnginaY.
That is when exercise Angina takes the value 1.