ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Nirmal Timilsina

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, Nirmal Timilsina 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

NT 3/7/2025

Student Signature *Date*

Exam must be turned in by: 75 min NT

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)

Discrimination is finding features (rules) for classification. (Example finding the features that seperate apple and oranges).
Classification is classifying the (new) observation using the rules / features to the most probable 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)

→ When we treat this problem as regression task, it will model a straight line function. It will also predict the values beyond 0 and 1. for probability, it is not possible to have these values. So, fitting a regression will not be a proper analysis. Instead, we could use logistic regression to fit the model. It will use sigmoid function & the probability remains between 0 & 1. Also, the regression will fit a function with sharp increase, whereas logistic regression fits S shaped curve

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)

⇒ Bayes error rate is the minimum error rate possible. It is like "irreducible error" in regression. Bayes classifier tries to classify the observation in most probable one. So, it is not possible to get values or error rate less than this. This is a kind of benchmark for other models performance.

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)

⇒ Accuracy is the total correct classification divided by total number of classification. No information rate is the measure obtained when the observation is classified as most occuring class. So, any model with higher accuracy than no information rate is a better model. If not, the model built is not a good model.

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

⇒ Sensitivity $= \frac{TP}{TP + FN}$; Specificity $= \frac{TN}{TN + FP}$

Sensitivity, or recall, is the models measure to correctly classify the observation as true positive out of True positive and false negative classifications

Specificity is the model measure to accurately classify the observation as negative out of the true Negative & false positive classification.

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)

⇒ The prior probabilities is denoted by P(A) for class A. The prior probabilities are estimated by the occurance of that class divided by total observation. P(A) = occurance of that class / total occurances

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 distributions mean & variance when fitting an LDA model. LDA model assumes same normal distribution for all class of predictor, unlike QDA. so, the normal distribution are same.

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)

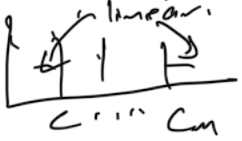
⇒ We might still prefer LDA to QDA because:
- LDA works well when there are less observation than QDA,
- there are less parameters to be estimated in LDA than QDA, so LDA is computationally more efficient.

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)

⇒ Naive bayes adds a constraint that simplifies the model. The assumption is that the classes within predictors are independent of each other. This is like QDA where the off diagonals of the matrix 0 i.e no correlations.

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

⇒ Natural cubic splines adds more constraint to cubic spline. The cubic splines are not stable at the boundary. So, natural cubic splines adds the constraint of linearity at boundary knots.

fig: 

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)

with / without

$P(Y=1|x) = \frac{e^{\beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3}}$; β_0 = intercept, β_2 = Exercise (Y or N), β_1 = Age, β_3 = cholesterol

$P(\hat{Y} = \text{heart disease}(1) \mid x) = \frac{e^{-4.4039 + 0.0530 * (\text{age}) + 0.0024 * (\text{cholesterol})}}{1 + e^{-4.4039 + 0.0530 * \text{age} + 0.0024 * \text{cholesterol}}}$

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)

⇒ In the above equation, we can use the link function to get the values. The values provided by logit function can be used to create the decision boundary. for values of Age & Cholesterol.

$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$; β_0 = intercept, β_1 = age, β_3 = cholesterol

$Y=1 \Rightarrow$ success or having heart disease

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)

⇒ β_0 ⇒ Intercept coefficient = -4.4039 ⇒ This means that the log odds of sucess (heartdisease=yes) decreases by 4.4039 for those without exercise angina with zero age and zero cholesterol (that is, when other predictors are 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)

⇒ β_1 ⇒ slope for age = 0.0530 ⇒ This means that for an unit increase in age (year) the log odds of sucess (having heart disease=yes) increases by 0.0530. when exersice angina & cholesterol are held constant.

β_3 ⇒ Slope for cholesterol = 0.0024 ⇒ This means for an unit increase in cholesterol increases log odds of sucess (having heart disease=yes) by 0.0024. when age & exersize angina are held constant.

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)

⇒ β_2 ⇒ Exercise AnginaY coefficient = 2.4644 ⇒ This means that the log odds of sucess (heart disease=yes) increases by 2.4644 for someone with exersie AnginaY compared to those without when age & cholesterol are held constant.