B-13

ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Jaswinder Kaur

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, _____ 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

Jaswinder Kaur — March 7 '2025

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)

Classification
classification task involve classifying observations of a data in predefined groups/ classes in response variable eg:- observation being apple classified in Y = fruit group

Discrimination.
This task involve finding relationships among the variable & based on the relationship & information can be used to make specific classes / groups to which obs. can be assigned to.

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)

With treating these as regression task take form of numeic ranges instead of factor & major problem would be that their probabilities can now even be less than 0 & more than 1, which doesn't make sense & should not be true. as $0 \leq P \leq 1$

In binary cases, we can do regression as done in class but that also with limited data. But this won't be generalized for multiple classes.

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, this error rate comes from randomness of data. It is equivalent to 'irreducible error rate' we can't do better in it or improve it.

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)

No information rate (NIR) is a rate at which even if nothing is applied the model will perform well. Foreg: if we have 3 classes with 5, 15, 55 observation each respectively than even under all condition model will predict all 55 right because of prevelance. Accuracy of model should be higher than NIR than only we can say its better model we should beat NIR

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

Sensitivity = $\frac{\#\text{ of positive predicted positive}}{\text{Total no. of observation}}$

Specificity → $\frac{\#\text{ of negative being predicted negative}}{\text{Total no. of observation}}$

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)

Using Bayes theorem, predicting posterior conditional probabilities followed by back flipping them & using them to get prior conditional probabilities

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 on density function & normal distribution of individual predictors Through normal distribution, we also model based on sample variance & sample mean for class K for each of predictor.
no The ND is not related anyway

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)

Even though QDA is more flexible, but the downside to it is it require alot of sample training data, alot of predictors, Therefore in data with less predictors LDA would be better along with when small training data is present

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 joint distribution of a model are product of individual marginal distribution of predictors variables. And all x are independent of each other & don't show any relationship or interaction

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

Natural cubic splnie, ends of the splnies are left as such. But in cubic splnie model the, tails are linear when x is smaller or extremly langer than the regnis in knot.

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)

logit function

$$\ln\left[\frac{P(Y=1\mid X)}{P(Y=0\mid X)}\right] = -4.4039 + 0.05 \times \text{Age} + 0 + 0.0024 \times \text{Choles}$$

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 can solve the equation to find conditional probability of without exercise angina. These conditional probability will set the rule for classification. Further applying glm model boundaries for this can be obtained & plotted.

Similar can be made & repeated for Age & cholesterol in respect to 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)

$\beta_0 = -4.4039$, would be the 'μ' when all the other estimated predictors are '0'

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 log loss would be increased by 0.05, when age of patient increases by one year (for eg 23 to 24) keeping all other variables were kept 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)

The log loss would be increased by 2.46, of with Exercise Angina, with respect to without Exercise Angina, given all other variable were kept constant