ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: VIDYUL JAIN.

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, VIDYUL JAIN 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

Vidyul . 03/07/2025.

Student Signature *Date*

Exam must be turned in by: _____ VJ.

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: Refers to classifying based on certain variables a binary / discrete outcome referred to as classes. Here we predict which class a particular data point belongs to based on the features we have.

Discrimination: Involves computing posterior probabilities and classifying the datapoint to the class with the highest posterior probability.

Main difference between the two is how we classify data points. Directly learning from features & classifying eg. KNN or using / computing probabilities to classify eg. LDA, QDA, naive Bayes

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)

Regression in general predicts the mean value of Y for given X. In this case it would predict the average of Y for all data points in the eg.

Eg. Data : $Y = 1, 2, 3, 1, 1, 2, 3, 4,)$

$\Rightarrow \hat{Y} = \frac{18}{9} = 2.$

Another implication is that this may predict values more than 1 which in the classification case doesn't make sense since we are trying to compute probabilities which is bounded by 0 & 1, i.e. $0 \leq p \leq 1$.

- It's never reasonable to use regression in a classification setting.

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)

Bayes' error rate is the best error rate we can achieve. We cannot do better than this error rate. It tries to model $Y|X$ using $X|Y$ which may not be possible in a real world situation.

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 of a classification model might be misleading as it only tells the proportion of datapoints it correctly classified. We may also want to know more about the type I and type II error which may be an important metric under different use cases and shows how the model actually performs on the wrongly classified datapoints.

No information rate →

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

Sensitivity - It's a metric to access the performance of a classification model. It measures how sensitive the model is to new data points.

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 assume that these probabilities follow a normal distribution & we the density function to compute these probabilities.

$X_1|Y=x \sim N(\mu_1, \sigma_1^2)$, - - - , $X_p|Y=x \sim N(\mu_p, \sigma_p^2)$.

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 probabilities, i.e.

$X|Y=m \sim N(\mu, \sigma^2)$

Under LDA model, we assume the variance-covariance matrix to be same for all variables 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 still prefer LDA over QDA in cases where the number of datapoints is small and hence restricting the variance becomes important and LDA helps in reducing model flexibility thereby controlling variance.

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 the data is independent & identically distributed, i.e. we take a naive assumption that the data belonging to each class is independent of other. Hence in naive Bayes classifier, joint distribution can be written as the product of marginal densities.

$f_{X|Y}(x|y) = f_{X_1|Y}(x_1, y) \times \cdots \times f_{X_n|Y}(x_n, y)$

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

Cubic spline

degrees of freedom: $d + M + 1$

Tails: may not be appropriate for the data; gets too wiggly at the tails

Natural cubic spline

M

linear tail helps in avoiding wiggliness in tails.

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)

$$\log\left[\frac{P(X|Y=1)}{1-P(X|Y=1)}\right] = -4.4039 + 0.0530(\text{Age}) + 2.4644(0) + 0.0024(\text{Cholesterol})$$

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)

We use this to compute the probabilities by passing it through the logistic function and if the probability is greater than some threshold value, we classify it to the point for which the probability exceeds the threshold value. This way we compute the decision boundary.

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 measures the log of odds of having heart disease given age = 0, ExerciseAnginaY = No = 0 and cholestrol equal 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)

It measures the amount of change in log of odds of having a heart disease for a unit change in the age variable with other variables constant/0 and ExerciseAnginaY=0.

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 measures the change in log of odds of having a heart disease when ExerciseAnginaY = Yes =1 & other variables 0.