ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Ali Shashaani

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, Ali Shashaani 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

3/7/2025

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 → defining a rule for the classification

discrimination → finding the predictors X_i that associate with the classification

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)

(everything is estimate)

we don't use regular reg. due to extrapolation

Yes. this is a logistic regression task that we do to obtain classification rule.

Generally $E(Y|X) = P(Y|X) = \frac{e^t}{1+e^t} \Rightarrow t = \log \frac{P}{1-P}$ (success / fail)

where $t = \beta_0 + \beta_1 X_1 + \dots$

for four levels: $P_1 + P_2 + P_3 + P_4 = 1$

we use one P as a base probability

so $\frac{P_1}{P_k}, \frac{P_2}{P_k}, \frac{P_3}{P_k}, P_k = P_{\text{for instance}} = \frac{1}{1+\exp(t_k)}$

both are special case of naive bayes

3. Select true or false for each classification method. (3 pts each)
   a) We can never use the Bayes classifier in a real scenario. T conditional discrete
   b) LDA is a special case of QDA. T
   c) Logistic Regression provides a discriminant for classifying our observations. F
   d) Binary logistic regression generally requires a larger sample size than multinomial logistic regression. F
4. We discussed the idea of the Bayes' error rate. Can we ever do better than this rate? Explain. (5 pts)

no - It's similar to irreducable error in regression task. $E\{I(Y \neq \hat{Y})\}$ $\begin{cases} 1 \text{ if } Y \neq \hat{Y} \\ 0 \text{ if } Y = \hat{Y} \end{cases}$

$= 1 - \max P(Y=j \mid X=x_0)$ → This is the optimal prob.

→ we don't know this

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)

assume we have defined a classifier → Accuracy = $\frac{\#\text{correct prediction}}{\#\text{total prediction}}$

NIR → is where don't use any classification method

we evaluate our model by comparing model metrics with NIR

rule of thumb

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

sensitivity → rate of positive correct prediction to total

specifity → rate of negative correct to total

we use confusion matrix to obtain these metrics

- correct prediction on diag
- incorrect " off diag

also we can compare values of $\hat{p}_k f$

$$\hat{P}(Y=k \mid X=x_0) = \frac{\hat{P}_k \hat{f}_{x_i|y_i}}{\sum \hat{P} f}$$

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)

$\hat{f}_{x_i|y_i}(x|k)$ density function estimate

$P_y = \frac{n_k}{n}$ we compare all P based on number of classes → most prevelent decision making

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)

LDA → $N_1(\hat{\mu}_1, \hat{\delta})$ ⋮ $N_m(\hat{\mu}_1, \hat{\delta})$

different $\hat{\mu}_i$

$\hat{\delta}$ is equal for all categories → linear discriminant function → linear decision boundry → an var-cov matrix is enough

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)

$P=10$ → model complexity incveases → overfitting possibility

LDA is a special case where we assume linear combination model and equal variance

usually when all assumption for LDA and QDA are made → LDA tend to have better performance

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)

assumption:

- predictors are linearly independent

Non linear combination of X's

Var-Cov matrix similar to QDA but with zero's on off diagonal elements

↳ zero correlations

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

cubic spline → BS() → can stable at the edges tend to extrapolate on both ends

Natural cubic spline → NS() → add a constraints term to control this issue

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

→ = 0

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)

$P(Y \mid X = age, cholestral) = \frac{e^t}{1+e^t}$

$t = -4.4 + 0.05\,age + 0.002\,chal + 2.4\,[\mathbb{I}(exercise)]$

indicator variable

zero

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 conditional probability for that

if $P(Y \mid X) > 0.5$ since we have a binary response

we would assign our observation to

class YES → person has heart disease

$t = \log odds = \log \frac{p}{1-p}$ $\frac{\text{success}}{\text{fail}}$

β_0 Angina = 0

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)

what would be the value/change in the Log odds given age and chalestral are zero and no Angina

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)

each year increase in Age waul change the Log odds 0.05 given chalestral remains constant → Person has Angina / " doesn't have Angina

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)

→ yes

If the person has Exercise Angina we would expect heard disease log odds to increase 2.46 units, given age and chalestral remain constant