ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Koji Takagi

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, Koji Takagi 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 3/7/2025 *Date*

Exam must be turned in by: 2:23 KT

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)

To make a classification, we create the decision boundary, which means discrimination. After that we will decide how to classify the observation. For example LDA, using Bayes theorem, we use prior probability and fitted normal distribution this is the discrimination, after that, taking account for the result we can classify using the new observation.

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)

In classification, Y is estimated as mean response, we want to know the probability of classification, the range should be 0 – 1.

So, in this case, we might get above 1 or negative values when the regression task was performed. So, we should use those values as factor values and logistic model instead.

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. . Bayes error rate is the minimum rate of the error.

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)

$NIR = \frac{\text{High rated class account}}{\text{Total}}$

if NIR > Accuracy this prediction model is not great.

	Positive	negative	
Positive	40	10	50 ← high rate class
negative	5	40	45

$NIR = \frac{50}{95}$ in this case,

$Accuracy = \frac{80}{95}$ Accuracy > NIR so this model is not bad.

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

Actual \ Predict	Positive	negative
Positive	TP	FN
negative	FP	TN

$Sensitivity = \frac{TP}{TP + FN}$

$Specificity = \frac{TN}{FP + TN}$

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 need fitted normal distribution.

$$P(Y=1) = \frac{P(X, Y=1) \times f(\text{mean } X, \sigma^2)}{P(X, Y=1) f(\bar{X}, \sigma^2) + P(X, Y=0) \times f(\bar{x}, \sigma^2)}$$

$$P(Y=0) = \frac{P(X, Y=0) \times f(\bar{x}, \sigma^2)}{P(X, Y=1) f(\bar{x}, \sigma^2) + P(X, Y=0) f(\bar{x}, \sigma^2)}$$

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)

mean and variance.

we suppose ~~that~~ covariance matrix is same among the classes. Example is as follows. $X|Y=1 \sim N(\mu_1, \sigma^2)$

$X|Y=2 \sim N(\mu_2, \sigma^2)$

$\mu_1, \mu_2, \mu_3, \ldots$ are different but σ^2 is same. $X|Y=3 \sim N(\mu_3, \sigma^2)$

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)

Generally speaking, it is curst of dimention.

Even though QDA model is more flexible, due to the high number of predictor, it causes overfitting.

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)

In the class, we suppose the observations are independent.

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

Although both spline are continuous, natural cubic spline are more likely linear at the end of the connection.

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)

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -4.4039 + 0.053 \times \text{Age} + 0.0024 \times \text{Cholestero}l$$

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)

if log odds > 0, we can classify this as success (having heart disease) at the specific values of Age and cholesterol.

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)

The log odds of success (Y=1, having heartdisease) when Age = 0, cholesterol = 0, and no excersise 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)

The log odds of success (having heart disease) for a unit change in age (1 year increase) when colesterol value and excersise Angina are 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 logodds of success (having heart disease) for having exercise Angia when colestrol and Age are constant.