ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Dezhong Xu

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, Dezhong Xu 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

Dezhong Xu. Mar. 7th 2025

Student Signature *Date*

Exam must be turned in by: _____ DX.

Exam End Time *Student's Initial Agreement*

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

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 ← hard to find true conditional distribution
   b) LDA is a special case of QDA. T
   c) Logistic Regression provides a discriminant for classifying our observations. T → direct estimate conditional distribution
   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. Baye's error rate refer to the natural error rate existed in the observation. We can't improve it, otherwise, we may overfit the data.

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 refers to the True negative errors where the model suppose to make positive predictions but failed. It interprets how well the model is fitting the true predictions, how well the model performed on TP predictions.

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

Sensitivity is the true positive rate. it defines how well the model performance in predicting the true condition observations.

Specificity is the false negative rate. it define how well the model performance in predicting the false condition observations.

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 is distingwish or assigning the qualitative predicted response for given predictors. It focus on assigning the data points to different groups that separated by the baye's equilibrium line. It focus on the probabilities of different response levels at the point of interest.

Discrimination is focusing on the bayes equilibrium. It is trying to find the lines or standard that would provide information for 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)

it depends on what regression model we use, for example:

It does not make sense when we use a linear regression model in this case, because the linear regression may output some counterintuitive values such as negative value or value outside the levels range ($Y<1$ or $Y>4$)

It would make sense if we use multinomial log regression, it would directly estimate the conditional probability of each level of Y given a set of X. ($Pr(Y=k|X=x)$) And we could classify the levels of Y for the point of interest based on the probability estimation result.

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 could use the number of observations in class that we interested and divide it by the number of total observations.

$$\pi_k = \frac{n_k}{n}$$

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 need just 1, because LDA assumes all class have same normal distribution

Yes, they are share the same normal distribution with same mean & same variance, they are related.

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)

LDA has lower computation cost by assuming all classes share same normal distribution.

Meanwhile, LDA would give more stable output, the p value is not small, QDA might overfit the data.

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)

the Naive Bayes assume that the conditional distribution (f_k) for each classes follow a normal & Gaussian distribution, but also, all predictors' distribution are independent from each other in each class.

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

The main difference would be the fitting line outside the boundary (when x is smaller than the smallest cutpoint and bigger than the biggest cutpoint). Natural cubic spline has linear fitting line outside the boundary which gives more stable predictions.

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)

$$\log\left(\frac{P(x)}{1-P(x)}\right) = -4.4039 + 0.0531 X_{age} + 2.4644 X_{exerciseAngina} + 0.0024 X_{cholesterol}$$

$$P(x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}$$ could convert to get this P(x)

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)

(1) we could set the $X_{exerciseAngina}$ to 0 first, so we could focus on age & Cholesterol.

(2) we got the log odds equation with X_{age} & $X_{cholesterol}$:

$$\log\left(\frac{P(x)}{1-P(x)}\right) = -4.4039 + 0.0531 + 0.0024 X_{cholesterol}.$$

(3) we then set $\log\left(\frac{P(x)}{1-P(x)}\right)$ to 0 which means, we want to find a line that any X_{age}, $X_{cholesterol}$ has a 50%, 50% probability

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 means the expected log-odds when all predictors are 0,

in our context, the natural probability / average probability / baseline for any human to have a heart disease is : $\frac{e^{-4.4039}}{1+e^{-4.4039}}$

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 age slope is 0.0530, it means for one unit increase in X_{age}, the log-odds will increase 0.053. In the data context, since the coefficient is positive, it means with the increase in age, the probability of have heart disease will increase. With 1 year increase in age, probability of having heart disease will increase $e^{0.053}$.

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 exercise Angina Y is 2.4644, it means that if we compare the log odds with exercise Angina (Yes) to exercise angina (No), the log-odds of exercise Angina (YES) would have a 2.4644 discrepancy compared to the log-odds of no.

In the data context, Since the coefficient is positive, it means that people with exercise angina would have more chance to get a heart disease compared to the people without. The probability discrepancy would be $e^{2.4644}$.