ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Kevin Kronk

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, Kevin Kronk 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

Kevin Kronk 3/7/2025

Student Signature *Date*

Exam must be turned in by: 4pm KK

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 is when we estimate the class of a response variable based on predictors. Discrimination is when we determine which predictors are useful in estimating the class of the response. In a classification task we are always doing classification, but not necessarily discrimination. Logistic regression does the first, LDA does both.

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)

Our regression would be fitting a line to the data in which we would get an estimate of y from 1 to 4. In this case the continuous response would represent some sort of closeness to a given class. One problem is that if we extrapolate beyond the X values in the data set we could end up with a value lower than 1 or greater than 4. What would it mean to get a 5? Additionally, this setup assumes that class 4 is farther away from class 1, than class 2 is from 1. But if the classes are what color someone would choose, this doesn't make sense. At best it might make sense with ordinal data like satisfaction with a service, but even then, we would be better off using something like logistic regression that gives the direct estimated probability of each class.

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)

No. The Bayes error rate comes from predicting classes based on the actual conditional distribution of the data: $P(y|x)$. Any error in this model is from random error that is irreducible.

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)

The no information rate is the proportion of the class that occurs most in a sample. Say class 1 occurs 60/100, and class 2 occurs 40/100 times. A model that simply predicts class 1 each time will get an accuracy of 60/100 or 0.60. This doesn't actually tell us anything about the data, hence the name no information rate.

so the NIR is 0.60

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

$\text{Sensitivity} = \frac{TP}{TP+FN}$ The number of times you correctly guessed the positive class, out of the total positive classes.

$\text{Specificity} = \frac{TN}{TN+FP}$ The number of times you correctly guessed the negative class, out of the total negative classes.

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)

Calculate the number of times each occurs divided by the total number of observations (Class A = $\frac{300}{1000}$ = .3 Class B = $\frac{200}{1000}$ = .2 Class C = $\frac{500}{1000}$ = .5)
$\hat{P}_K(A)$ $\hat{P}_K(B)$ $\hat{P}_K(C)$

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)

Used in the calc of prob of each class. → $\hat{P}_Y(k)\,\hat{f}_{X|Y}(x|Y=m)$ ← A conditional density is calculated for all m categories. In LDA these are assumed to be normal distributions, each with their own μ, but all of them having the same variance. So they're related in the sense that they have the same variance, but each are based on their respective $X|Y=m$.

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)

If we have limited observations - n, then there will be more variance from the training to test set. An LDA model is less flexible so it will be less likely to overfit the training set, meaning that it may perform better on the test set. Also if the classes themselves are separated linearly, then a linear decision boundary will likely perform better.

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)

$\hat{P}_Y(k)\,\hat{f}_{X_i|Y}(x_i|Y=m)$ In Naive Bayes we assume the predictors, X_i, are independent. ← This allows us to multiply the marginal densities in order to get the full conditional density. $\hat{f}_{X|Y}(x|Y=m) = \hat{f}_{X_1|Y}(x_1|Y=m) \times \hat{f}_{X_2|Y}(x_1|Y=m)$ With the assumption these are Normal distributions, then we don't have to worry about covariance, and there are less parameters to estimate.

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

The natural cubic spline model adds the restriction that the splines at the boundaries must be linear. This reduces the degrees of freedom and helps prevent a fit at the boundaries that is too "wiggly", making it less able to fit the data well and extrapolate

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)

$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

$$\ln\left(\frac{\hat{P}(Y=1)}{\hat{P}(Y=0)}\right) = -4.4039 + 0.0530(\text{age}) + 0.0024(\text{cholesterol})$$

log odds

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)

For values of age and cholesterol we calculate the log odds. When this is >0, we predict they have heart disease, <0, we predict they don't have heart disease. So when the log odds of heart disease $=0$ we get 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)

The intercept is the log odds of a person having heart disease given they are age=0, don't have exerciseAnginaY, and have cholesterol=0. Clearly these are impossible, so there isn't much practical meaning.

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 coefficient represents the change in the log odds of having heart disease with one unit (likely year) increase in age, given that ExerciseAnginaY and Cholesterol are held 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 ExerciseAnginaY coefficient represents the change in the intercept of the model. So the change in the log odds of having heart disease given that you have ExerciseAnginaY, holding age and cholesterol fixed.