ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Henry van Eijk

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, Henry van Eijk 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 *Date* 3/7/25

Exam must be turned in by: HV

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:

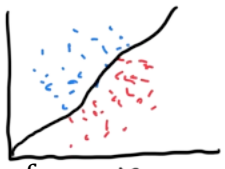
"I have neither given nor received unauthorized aid on this test or assignment."

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 predicting a categorical response (e.g. Spam vs not Spam)

Discrimination such as LDA/QDA, is about finding the discriminant functions in order to determine the boundaries of classes $1, \ldots, m$. In a binary case, this is when $P(Y=1|X) > 0.5$. We can visualize this through plotting the decision boundary.

Classification relates to just predicting a categorical response where as discrimination refers to finding the discriminant functions to separate classes $1, \ldots, M$.



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)

This doesn't make sense since y's are categories. For example, if treated as numeric category 1 is further from category 4 than categories 1 and 2. Instead we want to model these categories in terms of probabilities since our response is either a 0 or a 1. If we were to model these categories as numeric values, we could get probabilities outside the range 0 and 1 which does not make sense.

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 we cannot. Even if we knew the true distribution $P(Y=k|X)$, we could still never do better than the Bayes error rate. This is due to irreducible variation of the response $P(Y=k|X)$.

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 means we would always predict the class with the highest proportion, not even considering X. E.g., if our data has classes 0 and 1, with 90 o/o of responses in class 1 and 10% in class 0, the NIR is 90 o/o. Thus our goal is to create a model with accuracy > NIR otherwise our model is useless.

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

A confusion matrix is a step further than accuracy where we look at true negative, true positive, false negative, false positive. Sensitivity and specificity are two other metrics composed of TN, TP, FP, FN that take a more granular approach than accuracy. This could be important in applications like medicine where we want to avoid FPs. E.g., predicting a patient does not have a disease when they actually do.

$\hat{Y}$ \ Y	0	1
0	TN	FP
1	FN	TP

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)

Just look at the proportion of $P(Y=k)$. E.g., if we have six 1's and four 0's then our prior probabilities are

$P(Y=1) = 0.6$ and $P(Y=0) = 0.4$

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 $P(X|Y=1), \ldots, P(X|Y=m)$ with normal distributions.
Yes, since we are doing LDA, each normal has the same variance thus we have $N(\mu_1, \sigma^2), \ldots, N(\mu_m, \sigma^2)$.

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)

We may prefer LDA since it is a simpler model due to the equal variance across each normal distribution. Unlike QDA which each normal has its own σ^2.

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 ignore the marginal distribution $P(X)$. Bayes formula is

$$P(Y=k|X) = \frac{P(X|Y=k)\,P(Y=k)}{P(X)}$$ thus we consider $P(Y=k|X) \propto P(X|Y=k)\,P(Y=k)$

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

Cubic spline is not continious where as a natural cubic spline is continious.

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
x_1	Age	0.0530	0.0100	5.2905	0.0000
x_2	ExerciseAnginaY	2.4644	0.1925	12.8046	0.0000
x_3	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)

$$\text{log-odds}(\hat{p}) = -4.4 + 0.05\,x_1 + 0.002\,x_3 \quad \text{(log-odds)}$$

$$\hat{p} = \frac{1}{1 + \exp(-4.4 + 0.05\,x_1 + 0.002\,x_3)} \quad \text{(probability)}$$

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 this quanity $\hat{p} > 0.5$, we predict yes = 1 else No = 0. Or if $\text{log-odds}(\hat{y}) > 0$ we predict yes = 1 else NO = 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)

The intercept $\hat{\beta}_0$ is the log odds of $y=1$ (i.e., (having heart disease) when $x_1, x_2, x_3 = 0$ (i.e., age, exercise angina, and cholestrol are all 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)

A one unit increase in age results in a 0.053 increase to log odds of $y=1$ (i.e. having heart disease)

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)

If someone has exercise angina, the log odds of having heart disease is 2.46 larger than someone without this.