97 Well done!

ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Nick Zehnle

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, Nick Zehnle 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

Nick Zehnle 3/7/25

Student Signature *Date*

Exam must be turned in by: 11:40 NZ

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)

Discrimination is assigning probabilities to possible classes, i.e. discriminating amongst them, and Classification is selecting a class for given inputs. For example, logistic regression produces a discriminant function that can then be used to classify. LDA and QDA also use discriminant functions. KNN simply classifies because it is non-parametric, on the other hand.

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)

-2

No, even if you were to say something along the lines of $\hat{y} \in (1.5, 2.5) \Rightarrow \hat{y} = 2$, representing class 2, there would still be values of X that can generate $\hat{y} < 1$ and $\hat{y} > 4$. Thus, it would not be generalizable.

-2

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
   b) LDA is a special case of QDA. T (equivalent cov matrix ∀ classes)
   c) Logistic Regression provides a discriminant for classifying our observations. T
   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)

note: when classes are similar the irreducible error is larger

No, Bayes' error rate is analogous to the irreducible error since it is theoretically $1 - E[\underset{K}{\operatorname{argmax}}\, P(Y=K|X)]$, i.e. selects the highest posterior distribution based on the true conditional distribution $Y|X$ (Can't be known)

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 rate of success for only choosing the most prevalent class. Thus, it serves as a benchmark against your model accuracy. If your model is less accurate than if you were to just choose the more likely class every time then it is ineffective.

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

$$\text{Sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{Specificity} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}}$$

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)

Using the probability each class appears in the dataset; assumes sample is representative of population

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)

Fit normal distributions for $X|Y=1, \ldots, X|Y=m$.

For LDA variance is the same for all these normal distributions, but mean varies.

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)

also note curse of dimensionality for p=10 requiring an even larger sample for QDA to avoid overfitting

QDA is more flexible such that when n is small it will likely overfit. So in these cases we would prefer LDA since it will likely generalize 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)

The conditional distributions $X|Y=1, \ldots, X|Y=m$ are independent / not related

0

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

Natural cubic spline is linear at endpoints (ensures at endpoints $f''(x)=0$) and is a special case of cubic spline.

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)

log odds of success

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -4.4039 + .053X_1 + .0024X_3$$

where $P = P(Y \mid X_1, X_2, X_3) = 0$ in this case

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)

Set the log odds of success equal to 0. This is equivalent to setting $\hat{P} = .5$ such that every set of inputs that generates $\hat{P} > .5$ will be classified as heart disease ($Y=1$) and any that generates $\hat{P} < .5$ will be classified as no heart disease ($Y=0$).

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 log odds of success when $X_i = 0$ $\forall i = 1, 2, 3$. Basically the log odds of heart disease for an unborn person.

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 change in log odds of heart disease for an additional year of age. Heart disease log odds are expected to increase by .053 for each additional year you live.

holding other vars constant -1

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 log odds of heart disease are expected to increase by 2.4644 for a person with Exercise Angina.

-1