ST 563 601 – SPRING 2025 – POST
Exam #2

Student's Name: Dev Kewlani

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, Dev Kewlani 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

Dev Kewlani March 6th, 2025

Student Signature *Date*

Exam must be turned in by: _____

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."

we estimate $P(Y|X)$ using the data.

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 a broad class ML where we predict a response variable using a set of predictors [data]. Discrimination is a technique used in classification where we model the Joint Distribution of $P(X,Y)$ by estimating $P(X|Y)$ and $P(Y)$ and then using Bayes Theorem, we get an estimate of $P(Y|X)$

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)

The problem with treating this problem as a regression task is that we don't know whether the values are ordered or not. Treating it as a regression task assumes that the difference between level 1 & level 2 is same as level 2 and level 3 and so on. However level 1 could easily have been labelled as level 2 and vice versa and regression output would give us the same results which would be very concerning as the levels/categories of outputs have changed.

Q1 cont.: The difference ~~bfn~~ between the two is that classification is ~~one~~ ~~kind~~ a sort of an umbrella of problems where we model a qualitative response $P[Y|X]$ [conditional probability] and discrimination is a technique used to solve a classification problem where instead of directly estimating $P(Y|X)$ we first get an estimate of $P(X|Y)$ and $P(Y)$ and then use Bayes Thm to estimate $P(Y|X)$

Q2 cont.: This is the major problem we suffer with when treating classification problem in a regression setting. We could however use a regression setting in case of a binary outcome if we could constraint the support of the response because otherwise we will have predicted values / probabilities outside the $[0,1]$ range near the Boundaries ~~which~~ which cannot be interpreted since they are probabilities.

Thus it is best to use classification algorithms for categorical tasks.

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)

Bayes ER = $1 - E[I(y_i = \hat{y}_i)]$ and which assigns the class with the highest conditional probability of $P(Y|X)$. Since it estimates from the actual ~~dist~~ conditional distribution, we cannot do better than this. This error is equivalent to irreducible error in reg. setting.

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 is defined as $1 - \frac{\text{Correctly predicted}}{\text{Total Observation}}$ if NIR is 0.5, the model is no better than random chance. Accuracy of the model is just 1 − no information rate.

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

Sensitivity = correctly predicted % of +ve class. Or it is called the true positive rate, which is

$$\frac{I(\text{Predicted yes} = \text{True yes})}{(\text{Total yes})}$$

Specificity = 1 − False Positive rate = correctly predicted % of −ve class

$$= \frac{I(\text{Predicted No} = \text{True No})}{(\text{Total No})}$$

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)

most Basic way of estimating these is to use no. of observations of each class / Total obs.
or $\pi_K = n_K / n$

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 model, $P(Y=m \mid X=x) = \frac{f_m(X=x \mid Y=m) \cdot \pi(Y=m)}{\sum_{i=1} f_i(X=x \mid Y=i) \cdot \pi(Y=i)}$
in LDA, we model the $f_m(x)$ for each class and assume that it comes from a multivariate normal $(\vec{\mu}_m, \Sigma)$

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 might prefer LDA to QDA when p is large or n is small because QDA has more flexibility (more params) and generally requires large amount of data to give accurate predictions. With p=10, we may suffer from curse of dimensionality

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)

Naive Bayes model assumes that predictors in each class are independent of each other and so instead of modelling a joint distribution of $X_1 \dots X_p$ in each class, we could use
$f_m(X) = f_m(X_1 = x_1) \times f_m(X_2 = x_2) \dots \times f_m(X_p = x_p)$
for each individual density, we can assume it is Normal or use a Kernel density estimator. Even though this assumption is for convenience, NB classifier performs well when $\frac{n}{p}$ is not very high with respect to p

Q8 cont: Yes, The predictors density within each class share ~~a~~ common covariance matrix.

So we model the mean separately for each class & assume that the variances of each class is the same across all classes.

Q9 cont: and we will require a huge amount of data to accurately perform our classification without overfitting to the data woefully. Hence LDA maybe preferred w.r.t QDA.

Q11 cont:

It does so by adding two additional constraints per Boundary, so a total of 4 degrees of freedom are freed up as a result of that. A natural cubic spline is just an extension of cubic spline with the details mentioned. A cubic spline basically divides the predictor in K knots and estimates a different basis function for each knot and adds a truncated power basis constraint to ensure the continuity of the f^n across all regions.

↑ continued above

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

A natural cubic spline has 2 additional knots for each boundary $X < c_1$ & $X > c_m$ which makes the function estimate linear on boundaries and leads to more stable estimates. It does so —

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

[Do we have to discard cholesterol bc p-val > 0.05]? I'm using it since nothing mentioned.

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)

$$P(Y=1 \mid X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}$$

↓ cont.

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)

Decision Boundary occurs where (log odds) = 0

or $\log \frac{p}{1-p} = 0$ or $[P(x) \geq 0.5]$

$\Rightarrow$ Let $\exp(-4.4039 + 0.053\,\text{Age} + 0.0024\,\text{Chol.}) = Y$

$\Rightarrow \frac{Y}{1+Y} = 0.5 \Rightarrow Y = 0.5 + 0.5Y \Rightarrow 0.5Y = 0.5$

$\Rightarrow Y = 1 \Rightarrow \exp(\cdot) = 1$

$\Rightarrow$ $-4.4039 + 0.053\,\text{Age} + 0.0024\,\text{Chole.} = 0$

Q12a $P(Y=1 \mid X_1 = x_1, X_2 = 0, X_3 = x_3)$

$$= \frac{\exp(-4.4039 + 0.053 \text{Age} + 0.0024 \text{Cholestrol})}{1 + \exp(-4.4039 + 0.053 \text{Age} + 0.0024 \text{Chole}$$

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)

Intercept is basically the log odds when Age, Cholestrol ~~=~~ =0 and the person is without excise angina. $\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0$

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)

Age slope coefficient means that for every 1 unit change of Age, the log odds ~~increase~~ increase/decrease by the slope coefficient.

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)

ExerciseAnginaY's slope coefficient implies that if a person has exercise angina, their log odds increase by the slope coefficient of ExerciseAnginaY i.e., if they don't have ExerciseAnginaY, their log odds have a Baseline X. if they do, log odds become, $X + \beta_2$ + slope coeff. of ExerciseAnginaY.