Started 9:03 - 10:18

Chris Goodwin

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)

a classification task involves developing a rule to allocate new observations to a class of the response variable. A discrimination task is about finding the predictors that split our data into groups. The difference is that classification is more concerned with predicting the class of a new observation where discrimination is more concerned with finding the predictors that split our data into groups.

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)

For a categorical variable with four levels, a regression model will estimate the mean response for each category. However, since it doesn't impose constraints on the predictions, it may give probabilities < 0 or that don't add to 1.

3. Select true or false for each classification method. (3 pts each)

   a) We can never use the Bayes classifier in a real scenario. False
   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, it is like the irreducible error in regression. Its the lowest error that can be achieved by a classifier.

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 NIR is the proportion of observations belonging to the most common class. It represents the accuracy that you can get without using any model.

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

Sensitivity is the number of positive samples classified as positive divided by the total number of positive samples.

specificity is the number of negative samples classified as negative divided by the total number of negative samples.

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)

The number of observations in a class divided by the total number of observations. If class 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 the mean for each class and the variance across all classes. In LDA the variance is considered to be the same.

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 estimates one common variance-covariance matrix when QDA estimates one for each class, therefore a lot more parameters to estimate. Also, if training sample size is small, we may prefer it.

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 joint conditional distribution comes from multiplying the marginal distributions.

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

a natural cubic spline has a constraint that ensures they are linear beyond the boundary knots.

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)

$$\log\left(\frac{P(HD=1|X)}{1-P(HD=1|X)}\right) = -4.4039 + 0.0530\,Age + 0.0024\,Cholesterol$$

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)

The decision boundary is where $P(HD=1|X)=0.5$ or $\log\left(\frac{P(HD=1|X)}{1-P(HD=1|X)}\right)=0$. Set the equation in a. equal to 0 and solve for values of age and cholesterol.

$$0 = -4.4039 + 0.0530\,Age + 0.0024\,Cholesterol.$$

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)

Its the log-odds of having heart disease when age and cholesterol (and exercise angina) are all zero.

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)

Its the log-odds of having heart disease with a one year increase in age while holding the other predictors 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)

Its the log odds of having heart disease when exercise angina is yes while holding the other predictors constant.