Practice Exam 2 Questions

Your actual exam will have questions similar to those below with space below each question to give answers. I’ve tried to emulate that here but I haven’t taken the time to print the questions out and hand write the answers, which means there may be too much or too little space in this document. I’ll make sure an appropriate amount of space is given on the actual exam (which can also help you gauge how detailed of an answer to give!)

* Please note that the example questions below are not exhaustive!
* You may notice there are no programming questions below (that is, no R or python syntax at all)
* There are some pseudo code questions. Here you are writing out the logic of the process and how you would go about doing it within a programming language without worrying about the syntax or the process.
* There is very little calculation required for these questions. For most answers that involve output and reading/using it, you do not need to simplify calculations.
* I have not attempted to make the practice exam questions the same length as your actual exam (there will be fewer questions on your actual exam!)
* If you have any other questions about the content or structure of the exam, please post to the discussion forum!

1. We discussed using K-Nearest Neighbors (KNN) for classifying a response. Suppose we’ve determined an appropriate to use for a classification problem with three classes. We have three predictors, , , and . We want to use the KNN model to classify a response for a particular choice of , , and . Describe how the KNN model would go about classifying this observation.

We would first find the ‘distance’ (usually Euclidean) between our point of interest and all of the data points in the data set. We select the closest points in terms of this distance. We then classify based upon a majority vote (class with the highest proportion of occurrence in our neighbors).

1. Select true or false for each classification method.
   1. KNN provides a discriminant for classifying our observations (FALSE)
   2. The Naive Bayes provides a discriminant for classifying our observations (FALSE)
   3. Logistic Regression can include interaction terms. (TRUE)
   4. Shrinkage or penalized methods can be used with logistic regression (TRUE)
2. How does the Bayes’ classifier classify an observation? Based on your answer, why can we never apply this classifier in a real data situation.

The Bayes’ classifier uses the most probable class from the set of conditional distributions of .

We would need to know the conditional distributions of to use this! This is never known in practice.

1. Which two classification models we investigated directly modeled the conditional distributions ()?

Logistic regression, KNN

1. Which three classification models we investigated were ‘generative’ models that modeled the distribution of and used Bayes’ theorem to obtain a classifier?

LDA, QDA, Naive Bayes

1. Does the bias-variance trade-off idea exist in the classification setting? Explain.

Yes, an overly flexible model will be too flexible and train too closely to the data used to fit the model. This model will have high variability generally (training on a new data set would give very different classifications). A model that is not flexible enough will have high bias but low variance. That is, the decision boundary will not be flexible enough.

1. What is a confusion matrix? How is a confusion matrix more useful than simply looking at accuracy and misclassification rate?

A confusion matrix gives a contingency table between the predicted classes from the model and the observed classes from the data. This can help us to understand the ways that our model is performing well and how it is classifying incorrectly. The diagonals of the confusion matrix represent agreement between the model and the observed values. The off-diagonals describe counts for observations that were incorrectly classified.

1. Suppose we have a binary response variable. What major issue would come from treating the response as a numeric variable with values 0 and 1 and using a multiple linear regression model? Hint: Think about what our regression model is actually modeling, on average.

Our regression model will model the mean response. In the case of a 0-1 variable, this corresponds to a probability of success. There are no restrictions on the predictions from the regression model. This means we may predict a negative probability or a probability greater than 1.

1. Consider the (binary) logistic regression model with a single predictor variable . How do we interpret the slope coefficient associated with ?

The slope on corresponds to the change in the log odds of success for a unit change in .

1. Consider the (binary) logistic regression model with a single predictor variable . How do we interpret the intercept coefficient?

The intercept represents the log odds of success with is 0.

1. Consider the (binary) logistic regression model with a single predictor variable . How would we use the model to classify an observation with ?

We could either use the estimated log odds or the estimated probability of success to determine the classification (these are equivalent). If the log odds of success are greater than 0 at we would classify the observation as a success, otherwise we’d classify as a failure. Similarly, we could check if the estimated probability of success is greater than 0.5. If so, we classify as a success, otherwise we classify as a failure.

1. Suppose we have a categorical response with categories and a single predictor variable . When fitting a QDA model, we use the Normal distribution. What quantities do we model with a Normal distribution? Are those normal distributions related in anyway?

We model the distribution of , , …, each with their own normal distribution. That is, each Normal distribution is assumed to have its own mean and its own variance.

1. When doing LDA or QDA we end up with discriminant functions. How do these discriminant functions relate to the terms *linear discriminant analysis* and *quadratic discriminant analysis*? Note: I’m not looking for you to recall these functions. I just want you to note how they relate to the names.

The discriminant functions are functions of the estimated prior probabilities and the fitted normal distributions. We can compare them to determine which class to favor/classify.

When setting them equal, in the LDA setting we get a linear function of our predictor(s).

When setting them equal, in the QDA setting we get a quadratic function of our predictor(s).

1. For the Naive Bayes’ classifier we discussed the idea of using a *kernel density* estimator to model the distribution of a quantitative predictor given our response. That is, to approximate the distribution of each . Roughly describe what is meant by a kernel density estimator.

A kernel density estimator can be described as a smoothed histogram.

1. We discussed the use of generalized additive models (GAMs) for a regression task. Here we model

where each is of the form

and is the number of basis functions used to model . This is an *additive* model. Why do we call this an additive model?

This is an additive model because the contribution of each of our predictors is broken out by themselves, modeled, and then added together to form the estimate for the response.

1. Consider the piecewise polynomial regression model. Here we define our knots to be , …, and use the indicator functions

* in our regression equation given by

Suppose we have observations and we fit the model.

1. What is the estimate of in this model?

corresponds to the case when all the indicator functions are 0. This implies that our observation must fall in the region less than . is then the mean of all responses corresponding to these observations.

1. What is the estimate of in the model?

corresponds to the slope on the indicator that is between and . That means our observation must fall in this region. is then the mean of all responses corresponding to these observations.

1. Suppose we fit a piecewise linear model to a data set and then also fit a linear spline to the data set. What is the big difference between these two types of models?

The spline function requires that the resulting function is continuous. That is, the end of a line for one region must match up with the start of the line for the next region. This is not the case in the piecewise linear model.

1. What are two methods for choosing the number and placement of the knots in a piecewise polynomial or spline model?

There are really three major answers here. We could use cross-validation or a training and test set to determine these. We can use equally spaced quantiles of the predictor. We could use a maximal set of knots (smoothing spline).

1. When discussing smoothing splines where we minimize a penalized version of the residual sum of squares,

* we discussed that the optimal model here was a natural cubic spline with knots at the unique values of the ’s. That seems to imply a huge number of parameters! However, we discussed the idea of effective degrees of freedom. Conceptually, what is meant by effective degrees of freedom and what ‘controls’ this quantity?

Effective degrees of freedom is the measure of flexibility in the model here. When is near 0 the effective degrees of freedom are very large (near ). However, when is very large, the effective degrees of freedom decrease down to near 2 - which represents a linear fit.

1. Suppose we have data on whether or not someone has heart disease (No = 0, Yes = 1) and a number of predictor Age (quantitative). We fit an QDA model. Relevant output from the process is given below.

| HeartDiseaseFac | count |
| --- | --- |
| 0 | 390 |
| 1 | 356 |

| HeartDiseaseFac | Age\_mean | Age\_sd |
| --- | --- | --- |
| 0 | 50.2 | 9.3 |
| 1 | 55.9 | 8.8 |

1. Specify the distribution we would use for Age given someone does not have heart disease.

For the QDA model, we model the Age given no heart disease with a normal distribution. This normal distribution would have mean 50.2 and standard deviation 9.3.

1. Specify the distribution we would use for Age given someone does have heart disease.

For the QDA model, we model the Age given heart disease with a normal distribution. This normal distribution would have mean 55.9 and standard deviation 8.8.

1. Give the form of the posterior probabilities of heart disease and no heart disease using the QDA model for an arbitrary Age value. No need to simplify.

Here we want to form two numerators, one for no heart disease and one for heart disease, via prior probability times the fitted normal distribution evaluated at Age. We then divide each of those by their sum to find the probabilities.

Let be a normal pdf, the numerator of the probability of no heart disease is .

The numerator of the probability of heart disease is .

The probabilities are found by taking these numerators and dividing by their sums.