Problem Set #5

Kevin McAlister

March 11th, 2022

This is the fifth problem set for QTM 385 - Intro to Statistical Learning. This homework will cover applied exercises related to classification methods.

Please use the intro to RMarkdown posted in the Intro module and my .Rmd file as a guide for writing up your answers. You can use any language you want, but I think that a number of the computational problems are easier in R. Please post any questions about the content of this problem set or RMarkdown questions to the corresponding discussion board.

Your final deliverable should be two files: 1) a .Rmd/.ipynb file and 2) either a rendered HTML file or a PDF. Students can complete this assignment in groups of up to 3. Please identify your collaborators at the top of your document. All students should turn in a copy of the solutions, but your solutions can be identical to those of your collaborators.

This assignment is due by March 21st, 2022 at 11:59 PM EST.

Problem 1: The Multivariate Normal Distribution (20 pts.)

The multivariate normal distribution is an important distribution for the study of multivariate statistical models. Specifically, the multivariate normal distribution is one of just a few ways to parametrically specify a data generating process that encodes the covariance between pairs of random variables. A random vector $\boldsymbol{x} \in \mathbb{R}^P$ is said to follow a multivariate normal distribution if:

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) \sim \mathcal{N}_P(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{P}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]$$

Suppose we have N iid draws from an unknown multivariate normal distribution. We can specify the joint log-likelihood as:

$$\ell\ell(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = -\frac{NP}{2}\log 2\pi - \frac{N}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

Show that the maximum likelihood estimates of the parameters are:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})'$$

Some hints:

1. Let $\gamma = y'Ay$ such that γ evaluates to a scalar:

$$\frac{\partial \gamma}{\partial y} = 2Ay$$

2. Let $\gamma = y'A^{-1}y$ such that γ evaluates to a scalar:

$$\frac{\partial \gamma}{\partial A^{-1}} = yy'$$

This arises because of a trace trick rearrangement, $\gamma = y'A^{-1}y = \text{tr}(yy'A^{-1})$

3. A well-known matrix identity is:

$$\frac{\partial \log\det(A)}{\partial A^{-1}} = A$$

when A is symmetric.

Problem 2: Cancer Data (40 Points)

In a number of data sets, the predictors are a collection of ordered categorical ratings of objects. These predictors are then used to classify objects into categories. In class, we discussed plausibly continuous predictors (like income) and unordered categorical predictors (like manufacturing country). However, we did not spend time discussing ordered categorical predictors. This problem will see you work with two ordered categorical predictors - rating of uniformity of cell size (`UCellSize`: 1-10 with 10 being most irregular) and single epithelial cell size (`SECellSize`: 1-10 with 10 being largest) - trying to predict whether or not a cell is cancerous (`Malignant` is one if cancerous, 0 otherwise).

The problem is that there is no in-between: we must either treat the predictor as an unordered outcome and lose any information that comes from seeing how the predictor increases or decreases while preserving its discreteness **or** treat the predictor as continuous preserving any ordering but losing its discrete nature. Each choice comes with some downside, so we'll see how each one works on the training data set and use it to assess predictive accuracy on the test data set.

Part 1 (15 points)

If we choose to treat each predictor as a discrete input, then we can use a generative classifier built via Bayes' theorem to create predictions that preserve dependencies between the predictors. For discrete predictors X and Z, the classifier can be built using the following formula:

$$P(y = 1|X = x, Z = z) = \frac{P(X = x, Z = z|y = 1)P(y = 1)}{P(X = x, Z = z|y = 1)P(y = 1) + P(X = x, Z = z|y = 0)P(y = 0)}$$

Using the training data, create a lookup table that encodes the probability that an observation with $X = x$ **and** $Z = z$ is malignant. This should be a 10×10 table where each element corresponds to a possible $\{x, z\}$ pair. For some elements of this table, there are zero elements in the training set! These elements should be recorded as missing since we can't compute a probability using this approach.

Along with the probability lookup table, create a corresponding table that encodes the **Bayes' classifier** - for a given $\{x, z\}$ pair, which class has the highest probability of occurrence?

What is the general relationship between these predictors and the probability that a cell is cancerous? Does it appear that we've missed some information by treating this problem in an unordered discrete fashion?

Part 2 (15 points)

Now, build classifiers that treat the two predictors as continuous (and ordered, in turn). Specifically, use the training data to train 1) a logistic regression classifier and 2) a QDA classifier. Using these two classifiers, create probability and Bayes' classifier tables for each training method for each possible combination of x and z.

How do these tables differ from the ones computed in part 1? Have we lost anything by treating the predictors as continuous when they are truly discrete?

Part 3 (10 points)

Now, let's use the three models to create predictions for the test set and compare the results to the true class. For each observation in the test set, compute the probability that the cell is malignant using each of the three tables computed in parts 1 and 2. Using these values, compute the average probability of incorrect classification and the proportion of observations incorrectly classified under the Bayes' classifier. Which method performs best? Worst? Discuss which loss metric we might want to favor in this situation - think about the context of the classifier and how the predictions would likely be used.

Under what conditions might we expect the unordered discrete model to perform better than the continuous predictor one? Under what conditions might we expect the opposite to hold?

Note: There is one big weakness of the discrete Bayes' theorem approach. Briefly discuss this weakness and then skip any affected observations when computing the average loss.

Problem 3: Wine Data (40 points)

The Wine data set is a classic prototyping data set for classification methods. The data set revolves around a 13 different measurements of the chemical properties of different wines that originate from three different *cultivars* (varieties of plants - in this case, grapes - that have been produced by selective breeding). The goal of the classification task is to create a classifier for the three different *cultivars* using only the chemical properties.

The Wine data set only has 178 observations, so it is too small to split into training and test splits. Therefore, cross-validation methods are needed to approximate expected prediction error.

Part 1 (15 points)

Let's start with only 2 predictors: Color and OD280. Start by creating a plot that shows the predictor values in the training data colored by their class. Can you see approximately where the decision boundaries should be?

Using the training data, train a multinomial logistic regression classifier, a QDA classifier, and 2 Naive Bayes classifiers - one assuming normal conditional marginals and one using KDEs for the conditional marginals. For each classifier, produce a plot that shows the Bayes' classifier as a function of Color and OD280 for the **minimum bounding box** implied by the predictors (which is just a fancy way of saying predict the class for many combinations of predictors within the minimum and maximum of each predictor). How does the **decision boundary** differ between the 4 classifiers? Which one appears to do the best at capturing the true decision boundary within the data? Do any of the methods seem to overfit to the training data?

Hint: `Color` ranges from approximately 1 to 13 and `OD280` ranges from approximately 1 to 4. There are a lot of approaches to creating the decision plots, but I think that doing a grid evaluation is easiest.

Part 2 (15 points)

Now, let's work with all 13 predictors. The goal of this exercise is to build a model that best predicts out of sample wines, so use 5-fold cross validation to compute an estimate of the expected prediction error for three different classifiers - multinomial logistic regression, QDA, and naive Bayes (make a choice for the marginals given your observations in part 1). For each classifier, compute the 5-fold estimate of the average probability of incorrect classification **and** the misclassification rate with respect to the Bayes' classifier.

Which model performs best? Worst?

Part 3 (10 points)

Broadly discuss situations where Naive Bayes is likely to outperform QDA. Think about this from a loss perspective as well as a computational perspective. Specifically, consider the MLE for the multivariate normal you derive above - when do you think this computation would become time and/or computer space prohibitive?