STATS 415 – Practice Midterm

Winter 2018

1. (3 points) Circle the right answer for each question.

   (a) The bronze, silver or gold medal awarded at the Olympics is what kind of variable?

      i. Nominal
      ii. Ordinal
      iii. Interval
      iv. Ratio

      Solution: (ii)

   (b) If I have 100 values in my data (all of which are unique) and I add 0.5 to the largest 10 values, then how will this change the median?

      i. It will not change
      ii. It will become larger than the mean
      iii. It will become smaller than the mean
      iv. It will increase by some amount, but we do not have enough information to determine by how much

      Solution: (i)

   (c) If I have 100 values in my data (all of which are unique) and I add 0.5 to the largest 10 values, then how will this change the mean?

      i. It will not change
      ii. It will become larger than the median
      iii. It will increase by 0.05
      iv. It will increase by some amount, but we do not have enough information to determine by how much

      Solution: (iii)

2. (6 points) Explain whether each scenario is a classification or regression problem. Indicate whether we would be most interested in inference or prediction in each case. State the values of n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   Solution: Since CEO salary is a continuous response variable, this is a regression problem. Inference is the main goal, $n = 500$ firms, $p = 3$.

   (b) We are considering launching a new product and want to predict whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or a failure, price charged for the product, marketing budget, competition price, and ten other variables.

   Solution: Success/failure is a categorical response, so this is a classification problem. Prediction is the main goal, $n = 20$ products, $p = 13$.

3. (5 points) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

   Solution: For examples, see Fig. 2.17 on p. 42 of the book for training, test, and Bayes error, and Fig. 6.5 on p. 218 for bias, variance, and test error.

4. (4 points) What are the advantages and disadvantages of K-fold cross-validation relative to:

   (a) The validation set approach?

   (b) LOOCV?

Solution:

(a) The validation set approach is computationally cheaper, because there is only one model to fit instead of K, but it has higher variance.

(b) LOOCV is computationally more intensive (n models to fit instead of K) but it is very stable, and for a given dataset the results are not random. LOOCV has a higher bias than K-fold CV, and lower variance.

5. (6 points) Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y = c_1$ if receive an A and c_2 otherwise. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Write out the resulting logistic regression equation.
Solution:

$$\begin{aligned} \log \frac{P(Y = c_1|X)}{P(Y = c_2|X)} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \\ &= -6 + 0.05 \times \text{hours studied} + \text{undergrad GPA} \end{aligned}$$

(b) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.
Solution:

$$\begin{aligned} P(Y = c_1|X) &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}} \\ &= \frac{e^{-6 + 0.05 \times 40 + 3.5}}{1 + e^{-6 + 0.05 \times 40 + 3.5}} \\ &= \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.378 \end{aligned}$$

(c) How many hours would the student with a 3.5 GPA need to study to have a predicted 50% chance of getting an A in the class?

Solution:

$$\log \frac{0.5}{0.5} = -6 + 0.05 \times \text{hours studied} + 3.5$$

Therefore hours studied $= 2.5/0.05 = 50$.

6. (6 points) For the one dimensional data below, give the K-nearest neighbor classifier for the points $x = 1$, $x = 10$ and $x = 100$ using $K = 5$ and Euclidean distance. Show your work and write your answers below.

x	y
2	1
4	-1
6	1
8	-1
10	1
15	-1
20	1
60	-1
65	1
70	1
75	-1
80	1
85	-1
90	1
95	-1
100	1
200	-1

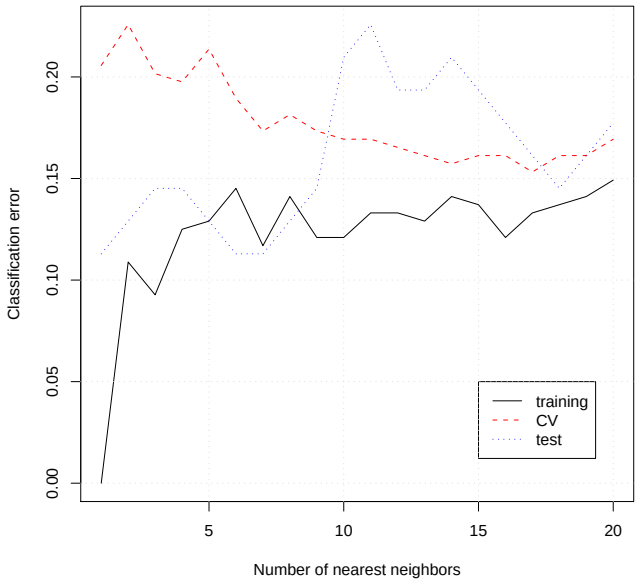
Solution: For $x = 1$, the 5 nearest neighbors are $\{2, 4, 6, 8, 10\}$ and the majority class is 1.

For $x = 10$, the 5 nearest neighbors are $\{4, 5, 8, 10, 15\}$ and the majority class is -1.

For $x = 100$, the 5 nearest neighbors are $\{80, 85, 90, 95, 100\}$ and the majority class is 1.

7. (6 points) This dataset contains data on patients belonging to one of three categories: Normal (N), Disk Hernia (H), or Spondylolisthesis (S). The latter two are both abnormal spine conditions. The dataset contains measurements of the following five quantitative variables derived from the shape and orientation of the pelvis and the lumbar spine: pelvic incidence (PI), lumbar lordosis angle (LL), sacral slope (SS), pelvic radius (PR) and grade of spondylolisthesis (GS).

   The researchers were interested in using the five numerical variables to predict the diagnosis (N, H, or S). There are 100 patients in class N, 60 in H, and 150 in S. The training data was obtained by selecting 80% of each class at random to be in the training set; the remaining observations were put in the test set. LDA, QDA, and kNN classifier were applied. The plot below shows the training, cross-validated, and test errors as functions of the number of nearest neighbors k used by the kNN classifier. The table shows training and test errors of all three methods, with k for kNN chosen by cross-validation.



	LDA	QDA	kNN
Training error	0.194	0.121	0.133
Test error	0.145	0.145	0.161

(a) Class priors for LDA and QDA were estimated from training data. What is the estimated class prior for class N?

Solution:

$$\hat{\pi}_N = \frac{100}{100 + 60 + 150} = 0.323$$

(b) For which values of k (the number of nearest neighbors) is the test error lower than the training error? Are there any values of k for which this cannot happen in principle? Explain your answer.

Solution: The test error is lower than the training error for $k = 6, 7, 8$. This cannot happen for $k = 1$ because for $k = 1$ the training error is always 0.

(c) What do the training and test errors for LDA, QDA, and kNN suggest about the feature distributions within classes? Comment specifically on normality and on class covariances. Explain your answer.

Solution: Since kNN performs fairly similar to both QDA and LDA, there is not much to be gained by removing the normality assumption, suggesting the feature distributions within classes are not too far from normal. LDA and QDA have the same test error (QDA has a lower training error than LDA, but that is always the case because QDA fits more parameters), so this suggests that the assumption of equal covariance matrices for all classes is also reasonable for this dataset.