Chapter 2.4 Exercise Solutions

Jason Moreau

January 18, 2020

Conceptual

Question 1a.

A flexible statistical learning method would be **worse** if the sample size n is extremely large, and the number of predictors p is small because a flexible method requires more variables (parameters) to reduce the errors.

Question 1b.

A flexible statistical learning method would be **better** if the number of predictors p is extremely large, and the number of observations n is small because it would provide a better fit.

Question 1c.

A flexible statistical learning method would be be **better** if the predictors and response is highly non-linear because the flexible method would provide a better fit of the observations.

Question 1d.

A flexible statistical learning method would be **worse** if the variance of the error terms, i.e. $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high because it would overfit the model and provide an inaccurate reading for the analyst.

Question 2a.

This scenario is a regression problem and we are most interested in inference. We are trying to determine the relationship between the dependent and independent variables/predictors.

n = 500 firms in the U.S.
p = profit, number of employees, industry, and the CEO salary

Question 2b.

This scenario is a classification problem and we are most interested in prediction. We are trying to determine whether or not the product will be a *success* or *failure.*

n = 20 similar products that were previously launched
p = success or failure, price charged, marketing budget, competition price, and ten other variables

Question 2c.

This scenario is a regression problem and we are most interested in prediction. We are trying to determine the relationship between the dependent (% change in the USD/Euro exchange rate) and independent variables.

n = Weekly data for all of 2012
p = % change each week in USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market

Question 4a.

Classification
- To find how many dogs vs. cats are in animal shelters in the United States
- To find which Japanese vehicle brand is purchased more often − Honda or Toyota
- To find which college major is most popular

Regression
- Find relationship between stock prices and bond rates
- Find relationship between life expectancy and income
- Find relationship between engine cylinders and miles per gallon

Cluster
- Find ethnicities within a city
- Find the gender demographic of a university
- Find the income demographic of a town

Question 5

The advantage of a very flexible (verses a less flexible) approach for regression or classification is that it provides a better level when conducting prediction modeling. The disadvantage is that level of interpretability suffers and it is not always accurate because it tends to overfit the model.

Question 6

A parametric approach is that allows the analyst to estimate using a set of parameters (ex. β_0, β_1, β_2,..β_p). A non-parametric approach attempts to estimate by getting a close a possible to fitting the data. An advantage of a parametric approach is that it makes estimating easier because the analyst doesn't have to fit the data the *f*, function of used to estimate the population, Y.

$$\mathrm{Y} = f(\mathrm{X}) + \epsilon$$

A disadvantage of a parametric approach is that it might not allow the analyst to pick the right model to obtain the true value of *f*.

Question 7a.

In the dataset provided we are to make a prediction for Y when $\mathrm{X}_1 = \mathrm{X}_2 = \mathrm{X}_3 = 0$ using K-nearest neighbors.

To find the Euclidean distance in two-dimensions when X = 0 we subtract each observation point by zero.

Observation 1:
(0 - 0) = 0, (3 - 0) = 3, (0 - 0) = 0

We then square each of the differences to find the absolute value:
$(0)^2 = 0$, $(3)^2 = 9$, $(0)^2 = 0$

We then find the sum of the squares: 0 + 9 + 0 = 9

Finally, we take the square root of the sum: $\sqrt{9} = \textcircled{3}$

Observation 2: 2
Observation 3: 3.16
Observation 4: 2.23
Observation 5: 1.41
Observation 6: 1.73

Question 7b.

The prediction when $K = 1$ is **GREEN** because the Observation 5 (1.41) is closest to 1.

Question 7c.

The prediction when $K = 3$ is **RED** because Observations 2, 5, 6 are closest to each other (take average of Euclidian distance of three observations that are closest to each other and less than 3). Based on probabilities, since Observation 2 and Observation 6 are both Red ($\frac{2}{3}$) and Observation 5 is Green ($\frac{1}{3}$), KNN will predict **RED**.

Question 7d.

If the the Bayes decision boundary in this problem is highly non-linear we expect the best value for K to be small. When K is small the decision boundary more non-linear. As K grows, the decision boundary becomes more linear.