Student Sourced Solution Manual for "An Introduction to Statistical Learning: with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

Ilya Kavalerov

August 28, 2014

The conceptual exercise solutions here were compiled into pdf with pandoc.

Chapter 2 Conceptual Exercise Solutions

2.1. Flexible vs inflexible methods performance

a. When sample size n is large, and the number of predictors p is small: fitting a flexible method often requires a great number of parameters, so this is difficult when the sample space is large.

b. when num of predictors p is large, and the number of observations is small, then it will be easy to fit an inflexible method

c. when the relationship between the predictors and response is very non-linear, then a flexible method will perform better

d. when the variance of the error terms is high, a more flexible method will perform better

2.2. Classification or regression?

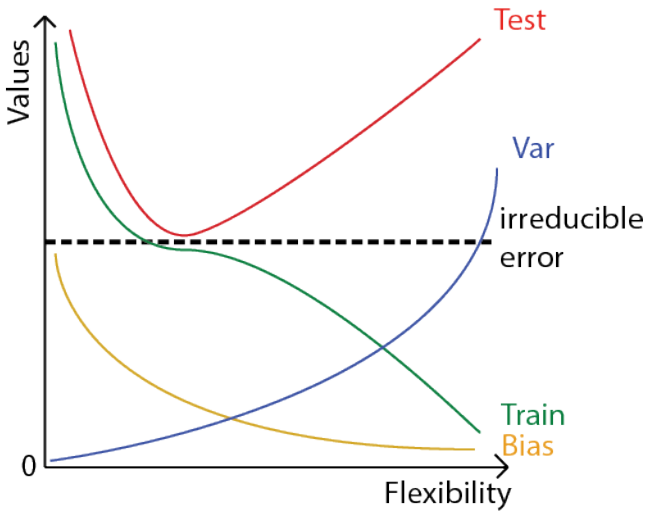
a. Inference, regression, p = 4, n = 500

b. Classification, prediction, n = 20, p = 14

c. Prediction, regression, p = 4, n = 52

2.3.

a.



2.4. c) Cluster analysis is useful when you want to classify a set of points into several groups, that have many different dimensions or characteristics

2.5. Flexible: Pros) Will always fit training data well, is preferred in non-linear cases Cons) prone to overfitting * Inflexible: Pros) the model may coincide with the true function that you are trying to approximate, preferred in linear cases Cons) makes assumptions about the underlying function, hard to correct these if the inflexible model was chosen wrongly

2.6. Parametric - pick a model, and optimize the parameters of that model, simpler. Non-parametric - is prone to overfitting the model, but fits the data very well.

Chapter 3 Conceptual Exercise Solutions

3.1. The null hypothesis was originally that there is no relation between tv, radio, and newspaper advertising budgets and product sales. It also tells us that in the absence of tv and radio expenditures, sales are non-zero. The tiny p-value for the intercept tells us that we can reject the null hypothesis, i.e. that ß0 == 0. The tiny p-values for tv and radio accordingly tells us that we can reject the hypotheses that ß1 == 0 and ß2 == 0, i.e. that there is no relationship between tv and sales, or radio and sales. Since the p-value for newspaper is substantially more than 5%, we can conclude that there is no relationship between newspaper advertisements and product sales, when tv and radio ads are kept constant.

3.2. KNN regression estimates the f(x0) of a test point by looking at the average y of the K points closest to that test point. In contrast, the KNN classifier estimates the conditional probability of a test point equaling each variation of f(x0) and then assigns f(x0) to the value that is most likely.

3.3. The estimated coefficient for IQ and the interaction between GPA and IQ are near zero.

a. Answer is iii

b. ß3 is positive and the gender dummy variable assigns 1 for females and 0 for males, then this means that the response, salary, is $35,000 dollars higher for females than males.

ii. At the same time, the interaction between gender and GPA is pretty negative, so overall, the ß5 interaction term should outweigh the ß3 term with a decent GPA (should be above 3.5).

iii. This is the correct answer.

iv. This would be the correct answer if it said "females earn more provided that the GPA is low enough." Odd, but think of a point where an identical male and female have the same 3.5 GPA, then think that if both of their GPA's were to increase, then the ß5 would start to outweigh the ß3 boost that females get, and their salary would start to decrease. For males, both ß3 and ß5 are always 0, since their gender is 0.

b. `50 + 20*4.0 + 0.07*110 + 35*1 + 0.01*4.0*110 + -10*4.0*1 = 137.1` in thousands of dollars.

c. We should be looking at the p-values in the model for the ß4 coeficient in order to determine whether or not the interaction effect is significant to the model. The coefficient only tells us how significant the interaction is relevant to the other parameters in the model, while the p-value would tell us how significant the interaction is with all other things kept constant.

3.4. In the case of an accurate fit, ß2 and ß3 would be near zero.

a. Since the cubic regression would be prone to end up overfitting the training data, and thereby reducing some of the random error in the true model, the RSS for the cubic fit would probably be lower.

b. For the case of the testing data, then the RSS should be lower for the linear regression than for the cubic regression, as a result of the overfitting described above.

c. Without knowing the true model, it is possible to justify that the RSS for either the cubic or the linear would be lower.

d. Same as above.

Extra: Why do we minimize the square of the error between our fit and the data?

- It is a result of the following hypothesis: The error terms of the true model have a Gaussian distribution.
  - strong assumption about both the shape of the distribution and that the variance is the same for all points
    - IID - independently identically distributed
  - errors are independent at different points
    - product of errors, when you take the log of the product, it becomes the sum of the errors, and they are all the same, so you multiply by the number of observations
- We will notice a squared x in the Gaussian distribution.
- calculate likelihood function, probabilities of a certain error given an error. Maximize that you got the observation.

Chapter 4 Conceptual Exercise Solutions

4.1.

We need to show: $y = \frac{x}{1+x}$ (4.2) $\Leftrightarrow \frac{y}{1-y} = x$ (4.3) i.e. that: $1 + x = \frac{1}{1-y}$

$$\begin{aligned} y &= \frac{x}{1+x} \\ y + yx &= x \\ -xy - y + x &= 0 \\ 1 - xy - y + x &= 1 \\ (1-y)(1+x) &= 1 \\ 1 + x &= \frac{1}{1-y} \end{aligned}$$

Chapter 10 Conceptual Exercise Solutions

10.6. Genes in tissue samples

a. The first principle component's corresponding loading vector is an eigenvector of the data's covariance matrix. It also has a corresponding eigenvalue which signifies the variance in the principle component. The proportion of variance explained is just the ratio of the 1st eigenvalue of the data's covariance matrix to the sum of all of the eigenvalues of the covariance matrix.

b. Some preparation before answering the question: From equation 10.2, we know:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip} = \sum_{k=1}^{p} \phi_{k1}x_{ik}$$

Which is the same as:

$$z_{ij} = \sum_{k=1}^{p} \phi_{kj}x_{ik}$$

In matrix notation this is:

$$\begin{bmatrix} | & & | \\ Z_1 & \cdots & Z_p \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ X_1 & \cdots & X_p \\ | & & | \end{bmatrix} \begin{bmatrix} \phi_{11} & \cdots & \phi_{1k} \\ \vdots & \ddots & \vdots \\ \phi_{p1} & \cdots & \phi_{pk} \end{bmatrix}$$

In the last matrix, k = p since the principal component loading matrix is square. Rearranging a little bit we get:

$$\begin{aligned} Z &= X\phi \\ Z\phi^T &= X\phi\phi^T \\ Z\phi^T &= X \ (\textit{since } \phi \textit{ is orthonormal}) \end{aligned}$$

$$\begin{bmatrix} | & & | \\ X_1 & \cdots & X_p \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ Z_1 & \cdots & Z_p \\ | & & | \end{bmatrix} \begin{bmatrix} \phi_{11} & \cdots & \phi_{p1} \\ \vdots & \ddots & \vdots \\ \phi_{1k} & \cdots & \phi_{pk} \end{bmatrix}$$

Stated as a sum, let's call this equation *q1*:

$$x_{ij} = \sum_{k=1}^{p} z_{ik}\phi_{jk}$$

The question says that the researcher wants to replace each (i,j)th element of X with: $x_{ij} - z_{i1}\phi_{j1}$

This means that, from equation *q1*, each element in X is no longer $\sum_{k=1}^{p} z_{ik}\phi_{jk}$ but is instead: $\sum_{k=2}^{p} z_{ik}\phi_{jk}$. The only difference is now, the projection of X onto the first principal component is left out.

The question mentions that the first principle component had a strong linear trend from earlier samples to later ones, and insinuates that this could be due to the researcher decreasing his use of machine A and increasing his use of machine B over time. By subtracting it, perhaps the researcher hopes to correct for using two different machines.

Ultimately though, he hopes to distinguish between the Control and Treatment groups for the tissue samples with a t-test for each of the 1,000 genes in the matrix - i.e. he wants to distinguish between the mean values of each gene between the C and T groups. This is an odd approach, since, as we will see in applied exercise 10.8b, plotting the top principal components against each other should reveal clustering between two different classes having different means.

This is because the first n components of PCA represent a hyperplane that gets as close as possible to the data. And to find a plane closest to the points, we're asking that the projection of the points onto the plane are spread out as much as possible, since we are looking for the top n directions that maximize the variance of our data. This view of clustering is shown in Figure 10.2, and is also shown with R code in exercise 10.8.