Stats 202: Data Mining and Analysis

Fall 2019

Problem 1

Chapter 5, Exercise 3 (Sec. 5.4, p. 198).

For part (a), instead of using simple English to explain, write pseudo R-code. That is, write some directions you might give a computer in an R session. We don't expect code that would actually run, but try to be realistic.

Problem 2

Chapter 5, Exercise 5 (Sec. 5.4, p. 198).

Problem 3

Chapter 5, Exercise 6 (Sec. 5.4, p. 199).

Problem 4

Chapter 5, Exercise 9 (Sec. 5.4, p. 201). Instead of **medv** use the variable **crim**.

Problem 5

The book only deals with Bootstrapping in the context of supervised learning. However, this method can be useful in unsupervised learning as well. In this exercise, we will apply bootstrapping to estimate standard errors in Principal Components Analysis (PCA).

The PCA algorithm uses the sample variance-covariance matrix. This is a statistic derived from a finite sample; therefore, it can deviate from the true covariance matrix of the variables. Because of this, every quantity estimated in PCA has some error.

1. Using the `USArrests` dataset, plot a histogram of the proportion of variance explained by the first 2 principal components in 1000 Bootstrap resamplings of the data.
2. Estimate a standard error and 95% confidence intervals for the proportion of variance explained by the first 2 principal components.
3. Suppose we compute the first principal component from each of 1000 Bootstrap resamplings of the data. Using the resulting 1000 vectors, we estimate the standard error of each entry or loading using Eq. 5.8 in the textbook. Explain why this would be problematic.
4. There is a way around the problem alluded to in part 3. Write a function in R which, given a data.frame:
    - Computes the vector of loadings for the first principal component and defines i to be the index of the element with highest absolute value.
    - For each of 1000 bootstrap resamplings of the data.frame, computes the vector of loadings for the first principal component and multiplies it by the sign of its ith element to generate *signed loadings*.
    - Plots a boxplot of the signed loadings in the boostrap samples.
5. Apply the function to `USArrests`.
6. The function described in part 4 yields estimates of the standard error for each loading of the first principal component. On what assumption does this method rely? Would this give good standard error estimates for principal components beyond the first few? Explain.