Elements of Statistical Learning
Homework Sheet 5

Optional: you can hand in solutions to questions 2, 3 and/or 4 for marking if you like. These questions are marked below in blue and with $*$. This is for formative assessment only and not official coursework and does not count towards your final course grade. Hand in solutions via Blackboard. Ensure your name is CLEARLY written on them. If you submit more than one page, ensure that the pages uploaded as a single document and in the right order, please!

1. Construct a synthetic data set (e.g. in R) that demonstrates the 'lack of continuity' of classification or regression trees. That is, if you move a *single* point, you obtain a completely different tree.

2. $*$ (This question is a slightly modified and corrected version of exercise 6.10 from the book *An introduction to the Bootstrap*, by Efron, B. and Tibshirani, R.J. (Chapman and Hall, 1994). The book has many other exercises you might consider doing). Consider the artificial data set consisting of the 8 numbers

$$1.2, 3.5, 4.7, 7.3, 8.6, 12.4, 13.8, 18.1$$

Let $\hat{\theta}$ be the 25% trimmed mean, computed by deleting the smallest and largest two numbers, and then taking the average of the remaining four numbers.

   (a) Calculate $\hat{\theta}$ on the data set.

   (b) For building confidence intervals, etc, one might need to calculate the standard error of $\hat{\theta}$ but, (i) it might be challenging to do this theoretically and (ii) we don't know the true distribution underlying the data and/or be not willing to assume one. The bootstrap standard error, $\hat{\text{se}}_B$ is defined by

$$\hat{\text{se}}_B = \left[\sum_{b=1}^{B} \{\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\}^2/(B-1) \right]^{1/2}, \tag{1}$$

   where B is the number of bootstrap resamples, $\hat{\theta}^*(b) = \hat{\theta}(x_b)$ is the trimmed mean of the bth bootstrap resample x_b, and $\hat{\theta}^*(\cdot) = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*(b)$. Calculate $\hat{\text{se}}_B$ for $B = 25, 100, 200, 500, 1000, 2000$.

   (c) Repeat the previous part using ten different random number seeds (`set.seed()` in R)) and, hence, assess the variability in the estimates. How large should we take B to provide satisfactory accuracy?

3. $*$ This is question 15.4 from the supplementary reading material "Elements of Statistical Learning". Suppose $x_i, i = 1, ..., N$ are independent and identically distributed (μ, σ^2). Let $\bar{x}_1^*$ and $\bar{x}_2^*$ be two bootstrap realizations of the sample mean. Show that the sampling correlation

$$\text{corr}(\bar{x}_1^*, \bar{x}_2^*) = \frac{n}{2n-1} \approx 50\%.$$

   Along the way, derive $\text{var}(\bar{x}_1^*)$ and the variance of the bagged mean $\bar{x}_{\text{bag}}$. Here, $\bar{x}$ is a linear statistic; bagging produces no reduction in variance for linear statistics.

4. $*$ Suppose you have access to a large dataset consisting of N observations of some numeric feature. For the purpose of this question suppose N is so large that this dataset cannot fit on a single laptop (of some moderate memory capabilities). We are interested in investigating various ways of implementing a bootstrap procedure for this dataset. Suppose we want to resample the dataset with replacement to obtain N observations and then take the mean of these. We will assume that we have access to multiple laptops/computers and that we can spread the work out across these computers and then aggregate the results.

   - As N is assumed to be very large someone naively suggests distributing the sampling procedure across x laptops/computers available by partitioning the data into x manageable parts and distributing these to each computer. At each computer we then draw N/x independent samples with replacement, sum the values up and then send these sums to a single computer where they are all summed together and then a final division by N yields an estimate of the mean. Does this method of distributing the sampling procedure yield the required independent bootstrap sample required? Explain your reasoning.
   - Assume the feature has n unique levels where $n \ll N$. Can you come up with a method to compress the dataset into a smaller form to allow us to perform the bootstrap on a single machine?