Bootstrap and Resampling Methods, fall 2012/3

Homework exercise 4

Due date: 29 January 2013 in my mailbox or by email to me

Submission format: Please include your code in your submission as an appendix. That is, write a proper HW submission giving your results, tables etc. and separately print out and include the code you used to generate the results.

1. **Bagging and sources of prediction error.**
   Using the code linked from the class homepage, load the movies training data, and do the following:

   (a) Draw 50 random samples of size 1000. For each one, build a regression tree using rpart(with parameters of your choice, but try to make the trees big). Apply all trees to one random sample of 100 observations from the validation set. Estimate the prediction error of your method by averaging the mean squared prediction error of the 50×100 predictions you observe. Estimate $Var(\hat{y})$ at each of these 100 validation observations using your 50 predictions in each one.

   (b) Now, for each of the 50 random samples of size 1000, run 20 bagging iterations and build a *bagged* prediction model. Repeat the calculations of the previous item on the same 100 validation observations. Estimate $Var(\hat{y})$ for the bagged models in the same approach. Also, estimate $Var(\hat{y}^*)$ at each of the 100 validation points (that is, empirically estimate the variance of the predictions from the 20 different trees, and average these estimates across the 50 samples).

   (c) Compare the prediction errors with and without bagging, and compare the variance of the different approaches, and the bootstrap-sampling-based variance. Discuss your conclusions.

   Note: It is important to compare all variances both at a specific validation observation (e.g., choose a few interesting ones and concentrate on them) and across all 100 observations chosen.

2. **MCMC on normal distribution.**
   For a standard normal distribution $N(0,1)$, implement a Metropolis algorithm using a $N(0,\sigma^2)$ "random walk" sampling function for different values of σ (in this mode, $x_{\text{new}} = x_{(t)} + N(0,\sigma^2)$). Start all chains from a random draw from $N(0,1)$.

   (a) Empirically estimate the autocorrelation curves for $\sigma \in \{1, 1.5, 2, 2.5, 3, 5, 10\}$. Plot them (as a function of time) and discuss which one is best and why.

   (b) With the same values of σ, calculate estimates of $E(X^4)$ for $X \sim N(0,1)$, using MCMC with $n = \{1000, 10000, 100000\}$ samples.

(c) Calculate the true expectation (integration in parts...) or find its true value, and compare to your results above.

(* Extra credit) Can you design a different MCMC algorithm, not using the normal random walk sampling, that will give guaranteed better estimates of $E(X^4)$?
Hint: Think how to accomplish the lowest possible level of dependence between consecutive samples.

3. **Gibbs sampling on a binormal example.**
Consider a simple bivariate normal distribution:

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

Implement a Gibbs sampling approach to generate data according to this model using only univariate random draws. That is, write down the conditional distribution X_2 given X_1 and vice versa, and use this to performa Gibbs sampling. Perform 100 repetitions of random-scan sampling (where you choose X_1 or X_2 randomly), and 100 of systematic-scan sampling (where you alternate between drawing X_1 given X_2 and vice versa). Run all 200 of the Gibss sampling experiments for 10000 iterations each, and use the resulting sequence to generate 200 estimates of

$$P(1 \leq X_2 \leq 2 | 1 \leq X_1 \leq 2)$$

Comparing to the true value of this probability, evaluate the performance of the two Gibbs samplers in terms of bias and variance.