Stat 462/862 Assignment 4

(Due in my mailbox at Jeffery Hall 406 on Dec 5h, 2019)

1. This problem involves the OJ data set which is part of the $ISLR$ package.

   (a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

   (b) Fit a tree to the training data, with $Purchase$ as the response and the other variables as predictors. Use the $summary()$ function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

   (c) Create a plot of the tree, and interpret the results.

   (d) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

   (e) Apply the $cv.tree()$ function to the training set in order to determine the optimal tree size.

   (f) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

   (g) Which tree size corresponds to the lowest cross-validated classification error rate?

   (h) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

   (i) Compare the training error rates between the pruned and unpruned trees. Which is higher?

   (j) Compare the test error rates between the pruned and unpruned trees. Which is higher?

2. Consider the problem of generating sample from a Beta distribution $Be(\alpha, \beta)$.

   (a) One result is, if two Gamma random variables are $X_1 \sim Ga(\alpha, 1)$ and $X_2 \sim Ga(\beta, 1)$, then

$$X = \frac{X_1}{X_1 + X_2} \sim Be(\alpha, \beta).$$

Use this result to construct an algorithm to generate a Beta random sample. Provide a density histogram to emulate the performance.

(b) Compare the algorithm in (a) with the rejection method based on (i) the uniform distribution; (ii) the truncated normal distribution.

3. Consider estimating the integral

$$\theta = \int_0^\infty \exp(-(\sqrt{x} + 0.5x))sin^2(x)dx$$

where the pdf of x is $f(x) = 0.5\exp(-0.5x)$.

(a) Conduct the Monte Carlo (MC) integration for estimating θ.

(b) Conduct MC integration using importance sampling with the following proposal functions

$$\begin{aligned} g_1(x) &= \frac{1}{2}\exp(-|x|), \text{(Laplace Distribution)} \\ g_2(x) &= \frac{1}{2\pi}\frac{1}{1+x^2/4}, \\ g_3(x) &= \frac{1}{\sqrt{2\pi}}\exp(-x^2/2). \end{aligned}$$

For sample size $M = 100, 500, 1000, 2000$, compare the mean and standard deviations of the estimates.

(c) (For graduate students 862 only) Implement MC integration using self-normalized importance sampling with $g(x)$ from a mixture normal density. Explain the procedure and integrate your results clearly.

4. (a) Provide a Metropolis-Hastings algorithm to generate samples from a binomial distribution $Bino(n, p)$ with

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, \ldots, n.$$

Use uniform distribution in $\{0, \ldots, n\}$ as proposal distribution and use independent chains. Compare estimated means and variances with the known theoretical means and variances of the binomial distribution.

(b) Provide a Metropolis-Hastings algorithm to generate samples from a standard normal distribution. The proposal distribution is the normal distribution with the mean being the current value in the chain and the variance being 0.25,0.01,100, respectively. Compare the estimated means and variance with the known theoretical means and variance of a standard normal distribution.