B365 Final Exam: 120 points

1. A group of 1000 people were asked four yes/no questions with the number of people answering a particular way tabulated in the R array count. For example count[“yes”,”yes”,”no”,”no”] gives the number of people that answered yes to the first two questions and no to the last two questions.

   (a) (5 pts) Write R code to compute an estimate of the probability of answering “no” to all four questions?

   (b) (7 pts) Write R code to compute the estimated marginal distribution on the answers to the first two questions. That is, the probability of (yes,yes), (yes,no), (no,yes), (no,no) for the first two questions.

   (c) (7 pts) Write R code to compute the estimated conditional distribution on the first question, given that the answers to the last two questions were both “yes.”

   (d) (8 pts) Suppose you have two such 1000-person random samples that produce estimates of .5 and .6 for the proportion of people who answer no to all questions. Is it reasonable to believe these two samples came from the same population. Explain your answer in detail — most of the credit for this problem is in the explanation.

2. This problem deals with the data set “x_and_y.csv” from the Canvas site.

   (a) (10 pts) Suppose that you would like to fit a third degree polynomial to the data. That is, you would like to approximate the y values by $\hat{y} = ax^3 + bx^2 + cx + d$ For some unknown constants a, b, c, d. Estimate the four unknown constants from this data set.

   (b) (5 pts) On the same plot show the (x, y) pairs as points in black and the $(x, \hat{y})$ pairs in red.

   (c) (5 pts) Compute the sum of squared errors (SSE) between y and $\hat{y}$.

3. Consider the system of linear equations:

$$\begin{array}{ccccccc} 4a_1 & + & 2a_2 & + & 2a_3 & = & 5 \\ 5a_1 & - & 3a_2 & - & 4a_3 & = & 6 \\ 7a_1 & - & 4a_2 & - & 2a_3 & = & 2 \\ 2a_1 & + & 2a_2 & - & 2a_3 & = & 10 \end{array}$$

   (a) (5 pts) Is it possible to solve the system of equations for a_1, a_2, a_3? Explain why or why not.

   (b) (5 pts) Write the system as a matrix equation: $Xa = y$, clearly identifying exactly what X, a, y are.

   (c) (8 pts) Suppose you want to find the a_1, a_2, a_3 that comes as close to solving the system as possible. Using R, compute a_1, a_2, a_3.

   (d) (5 pts) Describe precisely in what sense your a_1, a_2, a_3 are the best choice.

4. Consider the tree structured classifier for a two-class depicted below, using the usual convention for the numbering of nodes (the children of node k are $2k$ and $2k + 1$).

node	+/-	terminal
1	50/50	0
2	40/10	0
3	10/40	0
4	35/5	0
5	5/5	1
6	8/2	0
7	2/38	1
8	30/5	0
9	5/0	1
12	4/1	1
13	4/1	1
16	29/1	1
17	1/4	1

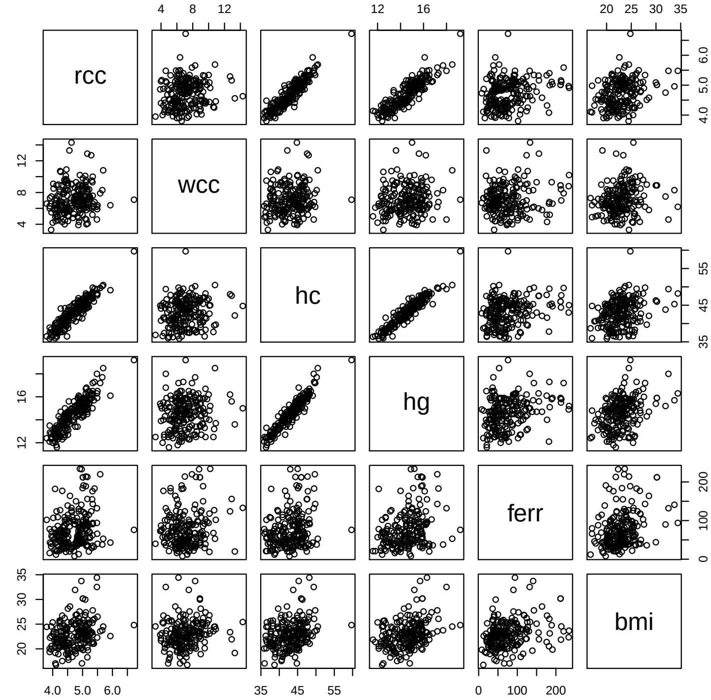
(a) (5 pts) Find the pruning of the tree that gives the best (minimum) value of $R(T) + \alpha|T|$ where $\alpha = 0.0$. Here, as usual, $R(T)$ is the error rate of the tree, T, while $|T|$ is the number of splits in T. Draw the tree labeling each node with its number and p/q where p is the number of + examples and q is the number of - examples reaching the node.

(b) (8 pts) Do the same for $\alpha = .02$.

(c) (5 pts) Do the same for $\alpha = 1$.

(d) (7 pts) Suppose we have an additional collection of labeled examples, not used in the construction of the tree above. Explain in detail how would you use this collection to decide between the three trees constructed above?

5. Consider the pairs plot shown below.



(a) (5 pts) Suppose we want to perform regression using the rcc variable as our response, with all of the other variables as predictor variables. If we implement the forward selection process, which variable would be selected in the first iteration of the algorithm? Explain your answer.

(b) (5 pts) If we call the response variable y and the selected variable x, approximate α in the regression equation $\hat{y} = \alpha x$.

6. Suppose that the prior probability of finding life on a randomly chosen planet in the universe is $P(L) = .01$. We know that the probability of having an oxygen-rich environment for a planet containing life is $P(O|L) = .9$, while the corresponding probability when life is not present is is $P(O|\bar{L}) = .01$. Similarly, the probability of having an Earth-like range of temperatures when there is life on the planet is $P(T|L) = .95$ while $P(T|\bar{L}) = .3$. After observing the two binary variables: oxygen and temperature status, we wish to classify the planet as containing life or not.

(a) (5 pts) Does the problem give enough information to compute the Bayes classifier?. If not, say exactly what is needed that is not given.

(b) (10 pts) For the 4 possible scenarios: $(O, T), (\bar{O}, T), (O, \bar{T}), (\bar{O}, \bar{T})$ say how the Naive Bayes classifer would classify the planet, showing your calculations completely.