Stats 315B: Homework 1

Joe Higgins

Due 5/6/2018

Question 1

1. (15) Data Mining Marketing. The data set age_stats315B.csv represents an extract from a commercial marketing database. The goal is to fit a regression tree to predict the age of a person from 13 demographic attributes and interpret the results. Note that some of the variables are categorical: be sure to mark them as such using the R function as.factor, before running rpart. Use the RPART implementation of the decision tree algorithm to fulfill this task. Write a short report about the relation between the age and the other demographic predictors as obtained from the RPART output and answer the following questions:

(a) Were surrogate splits used in the construction of the optimal tree you obtained? What does a surrogate split mean? Give an example of a surrogate split from your optimal decision tree. Which variable is the split on? Which variable(s) is the surrogate split on?

(b) Using your optimal decision tree, predict your age.

```
n <- 2
```

Question 2

2. (15) Multi-Class Classification: Marketing Data. The data set housetype_stats315B.csv comes from the same marketing database that was used for problem 1. Refer to the documen- tation housetype_stats315B.txt for attributes names and order. From the original pool of 9409 questionnaires, those with non-missing answers to the question "What is your type of home?" were selected. There are 9013 such questionnaires. The goal in this problem is to construct a classification tree to predict the type of home from the other 13 demographics attributes. Give an estimate of the misclassification error of an 1 optimal tree. Plot the optimal tree if possible (otherwise plot a smaller tree) and interpret the results.

```
n <- 2
```

3. (5) What are the two main reasons why a model that accurately describes the data used to build it, may not do a good job describing future data?

- Overfitting to training data
- Future data comes from different distribution than training data

4. (5) Why can't the prediction function be chosen from the class of all possible functions?

Because then we would have a problem: there are many (infinite) solutions that minimize empirical risk and we can overfit. If our function class was all functions, we would get many bad and overfitting functions if we don't have enough data or we don't regularize our score with certain penalties.

5. (5) What is the definition of the target function for a given problem? Is it always an accurate function for prediction? Why/why not?

The target function is the function that minimizes prediction risk (expected value of loss). No it is not always an accurate function for prediction because...

6. (5) Is the empirical risk evaluated on the training data always the best surrogate for the actual (population) prediction risk? Why/why not? In what settings would it be expected to be good?

Often empirical risk evaluated on training data works well. However, this is not always the case.

7. (10) Suppose the loss for an incorrect classification prediction is the same regardless of either the predicted calue ck of the true value cl of the outcome y. Show that in this case misclassification risk reduces the classification error rate. What is the Bayes rule for this case in terms of the probabilities of y realizing each of its values PR(y=ck)Kk=1? Derive this rule from the gneral (unequal loss) Bayes rule, for this particular loss structure Lkl = 1(k != l).

8. (5) Does a low error rate using a classification rule derived by substituting probability KK estimates {Pr(y = ck)}k=1 in place of the true probabilities{Pr(y = ck)}k=1 in the Bayes rule imply accurate estimates of those probabilities? Why?

9. (5) Explain the bias-variance trade-off.

The bias-variance trade-off states that in general if we increase model complexity, we tend to reduce squared bias but increase variance.

An example of a high-bias low-variance estimating function is a constant (for a target function that is not a constant): there is a consistent and predictable bias, and there is no variance.

However we can reduce bias by introducing more model complexity. Specifically, we can change the estimating function to rely on the data. As we increase model complexity by having it rely on the data, we inherently introduce variance since the it is assumed there is a random component to the data and thus our loss is random.

10. Why not choose surrogate splits to best predict the outcome variable y rather than the primary split.

11. Show the values of cm that minimize the squared-error risk score criterion:

$$
\begin{aligned}
F(\mathbf{x}) &= \sum_{m=1}^{M} c_m I(\mathbf{x} \in R_m) \\
S(\mathbf{x}) &= \sum_{i=1}^{N} (y_i - F(\mathbf{x}_i))^2 \\
S(\mathbf{x}) &= \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} c_m I(\mathbf{x}_i \in R_m))^2 \\
S(\mathbf{x})_m &= \sum_{i=1i}^{N} (y_i - c_m I(\mathbf{x}_i \in R_m))^2 \\
\frac{dS_m}{d\mathbf{x}} &= \sum_{i=1i}^{N} [2(y_i - c_m I(\mathbf{x}_i \in R_m))(-c_m I(\mathbf{x}_i \in R_m))] \\
\text{set } \frac{dS_m}{d\mathbf{x}} &= 0... \\
0 &= \sum_{i=1i}^{N} [(y_i - c_m I(\mathbf{x} \in R_m))(c_m I(\mathbf{x} \in R_m))] \\
0 &= \sum_{i=1i}^{N} y_i c_m I(\mathbf{x} \in R_m) - c_m^2 I(\mathbf{x} \in R_m) \\
\sum_{i=1i}^{N} c_m^2 I(\mathbf{x} \in R_m) &= \sum_{i=1i}^{N} y_i c_m I(\mathbf{x} \in R_m) \\
c_m^2 \sum_{i=1i}^{N} I(\mathbf{x} \in R_m) &= c_m \sum_{i=1i}^{N} y_i I(\mathbf{x} \in R_m) \\
c_m &= \frac{\sum_{i=1i}^{N} y_i I(\mathbf{x} \in R_m)}{\sum_{i=1i}^{N} I(\mathbf{x} \in R_m)}
\end{aligned}
$$

Since this c_m minimizes each $S(\mathbf{x})_m$ and $\sum_m S(\mathbf{x})_m = S(\mathbf{x})$ then this choice of c_m minimizes $S(\mathbf{x})$.