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