Lab 6 Decision Trees

Problem Statement

We are going to study the `titanic` dataset and use the decision tree to predict whether a person would survive or not.

Dataset

Install the packages `readr` and `dplyr` to read in the dataset and process the dataset. Use the following code to obtain the dataset. Note `dplyr` imports `magrittr` which uses the `%>%` syntax used below.

```r
library(readr)
library(dplyr)

titanic3 <- "https://goo.gl/At238b" %>%
     read_csv %>% # read in the data
     select(survived, embarked, sex, sibsp, parch, fare) %>%
     mutate(embarked = factor(embarked), sex = factor(sex))
```

Each row in the data is a passenger. Columns are features:

- `survived`: 0 if died, 1 if survived
- `embarked`: Port of Embarkation (Cherbourg, Queenstown,Southampton)
- `sex`: Gender
- `sibsp`: Number of Siblings/Spouses Aboard
- `parch`: Number of Parents/Children Aboard
- `fare`: Fare Payed

Remark: `dplyr` is a very useful package and is widely used in practice.

Here is a tutorial on `dplyr`: http://genomicsclass.github.io/book/pages/dplyr_tutorial.html

Here is another one: https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

Questions

1) `survived` is a numeric value. We need to first transform it to a categorical value. Use `titanic3\$survived = as.factor(titanic3\$survived)` to do so.

2) Fit a classification tree using all the observations. Find out which variables actually contribute to building this tree. Plot the tree.

3) Now we are going to estimate the test error:

- a. Split the observations into a training set and a test set.
- b. Build the tree using the training set, and plot the tree.

- c. Evaluate its performance on the test data.

4) Next, let's find out whether pruning the tree might lead to improved results.

a. Use `cv.tree()` to determine the optimal level of tree complexity.

b. According to the result, do you think pruning is necessary? Why or why not?

c. If you think it is necessary, or would like to give it a try, use `prune.misclass()` to prune the tree and evaluate the performance of the pruned tree.