

Exercises 02582  
Module 6  
Spring 2018

March 13, 2019

## Topics: Classification and Regression Trees (CART)

Exercises:

- 1 The first few exercises concern the fictitious movie dataset below. A real database with information on movies and TV-series is available in different formats from <http://www.imdb.com/interfaces>. The goal here is to predict the user rating of an upcoming movie as soon as the information on cast and budget is known, and to be able to explain in words why the movie will go straight to the Oscar's – or straight to oblivion.

Observation	Actor	Budget (\$ million)	IMDB User Rating
1	Nicholas Cage	100	2.8
2	Scarlett Johansson	50	8.3
3	Scarlett Johansson	150	4.0
4	Nicholas Cage	20	2.9
5	Al Pacino	75	7.8
6	Al Pacino	150	8.1
7	Al Pacino	115	3.0
8	Nicholas Cage	115	3.0

- (a) Is the described problem a classification or regression problem? Motivate your answer.
- (b) Which input variables are categorical and which are continuous?
- (c) For a categorical variable with  $k$  unique categories, what is the number of possible splits into two groups? Note that empty groups are not allowed, and that groupings

are commutative in the sense that e.g. the split  $\{1,2,3\},\{4,5\}$  is equal to the split  $\{4,5\},\{1,3,2\}$ .

- (d) What is the total number of splits to investigate at the root node for the movie dataset?
  - (e) Build a tree that predicts IMDb ratings.
- 2 Matlab and R have functions for building, pruning, evaluating and viewing classification and regression trees (Python is missing pruning at this point). We are going to use them to diagnose heart problems based on a set of 13 clinical variables from 303 patients and healthy controls. The data is in the file `ClevelandHeartData.csv`. The first 13 columns are different features and the 14th column is an indicator for heart problem/healthy. You can read more about the data in `ClevelandHeartDataDescription.txt`.
- (a) Read the help files for the tree methods in your preferred language to familiarize yourself with the possibilities.
  - (b) Build a large tree with the minimum number of observations (*minLeaf*) in at leaf set to 1 and view the tree.
  - (c) Choose optimal tree size by tuning the parameter *MinLeaf* value using cross validation.
  - (d) View the optimal tree and try to interpret it such that it makes sense for a doctor.

Exercises 02582  
Module 6  
Spring 2018

March 13, 2019

## Topics: Classification and Regression Trees (CART)

Resources for this exercise:

### Listing 1: Resources in Matlab

```
fitctree % for fitting trees  
cvLoss % for cross validation  
predict % for predicting  
view % for viewing trees
```

### Listing 2: Resources in R

```
library(rpart) # fitting trees  
read.csv('Name_of_file') # read .csv file
```

### Listing 3: Resources in Python

```
from sklearn.tree import DecisionTreeClassifier # loading tree  
import pandas as pd  
pd.read_csv # read csv file
```

End of exercise