BST 219
Core Principles of Data Science

Lecture 27: Machine Learning continued
December 10, 2024

Recipe of the Day!

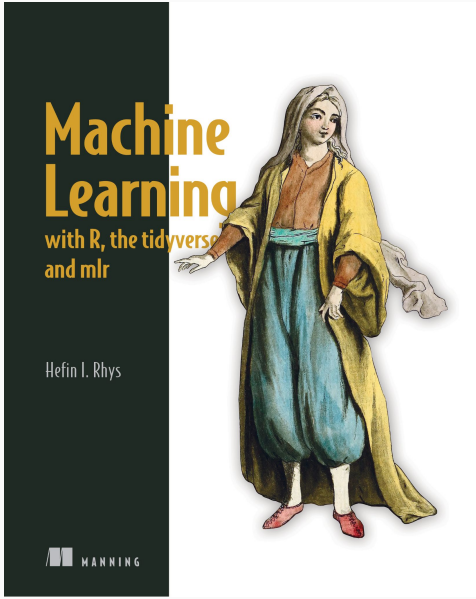
Korean Fried Cauliflower

Chef Emre

Agenda

- Announcements
  - Lab this week!
  - Homework #4 due December 16th by 11:59pm
  - Final projects have been assigned (see Canvas announcement for repo access) and TFs have been assigned
- Continue Machine Learning module
  - Decision trees
  - Random Forests

Machine Learning with R



Coding Question of the Day!

Using the gapminder dataset, fit a classification tree model that predicts fertility (low vs high) using life expectancy, infant mortality, population, and gdp as predictors and data from the year 1989 only. Print the tree. Which variables were used to construct the tree?

The code that categorizes fertility into low (fertility <= 4) and high (fertility > 4) groups, with 0 indicating low fertility and 1 high fertility, has been provided. The training and test sets using 70% of the data for the training set and 30% for the test set have also been coded for you.

Make sure to run this code first (it's also available on the course repository in the coding question of the day folder)

```
library(dplyr)
library(ggplot2)
library(caret)
library(dslabs)
library(tree)
library(rpart)

data(gapminder)
set.seed(9)

gapminder_1989 <- gapminder %>%
  filter(year == 1989) %>%
  mutate(fertility_cat = ifelse(fertility <= 4, 0, 1))

y <- gapminder_1989$fertility_cat

train_index <- createDataPartition(y, times = 1, p = 0.7, list = FALSE) # Partition data

train_set <- gapminder_1989[train_index, ] # Create training set
test_set <- gapminder_1989[-train_index, ] # Create test set
```