IEOR 242 Lab 4

CART in R

Mengxin Wang

IEOR, UC Berkeley

11/09/2019

Berkeley
UNIVERSITY OF CALIFORNIA

Today's Agenda

(Pair programming, please)

- General coding notes
- Classification and Regression Trees (review)
- Cross-Validation (review)
- Parole Violator Prediction using CART
- Bayes Theorem Review

Let's practice finding functions

Find a function that can take turn $\vec{x} = \{1, 2, 3\}$ and $\vec{y} = \{a, b\}$ into Table 1.

x_1	x_2
1	a
2	a
3	a
1	b
2	b
3	b

Table 1: Your Challenge

Let's practice finding functions

Find a function that takes $\vec{x} = \{1, 2, 3, 4, 5, 6\}$ and $\vec{y} = \{6, 1, 4, 8, 2, 1\}$ and returns $\{6, 2, 4, 8, 5, 6\}$

CART

- Build a tree by splitting on independent variables Can predict
  - a continuous outcome: regression tree
  - a discrete outcome: classification tree (multiclass classification is natural)
- Trees lead to a partition of the feature space: k distinct regions $R_1, .., R_k$
  - For regression trees, make a prediction based on the mean response value (w.r.t. to the training data) in the corresponding bucket
  - For classification trees, make a prediction based on the most commonly occurring class in the corresponding bucket
- Hyperparameters: cp, minbucket,...

Cross Validation

- Training and test set
- Leave-one-out Cross Validation
- k-fold Cross Validation

Questions:

- When is k-fold cross validation equivalent to LOOCV?
- Which one is the least sensitive to the random number generator?

Download

- Lab4Part1.R
- Lab4Part2.R
- NYCparole.csv

New Packages

- rpart
- rpart.plot
- caret (for CV)

Remember to install all the packages before using them!
install.packages(c(”rpart”, ”rpart.plot”, ”caret”))

Bayes Theorem

Given two random events A and B. We know the $P(A)$, the prior probability of A and the conditional probability of B given A, which is $P(B|A)$. How to get the posterior probability $P(A|B)$?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{1}$$

Example

$A = \{\text{Rain in Berkeley}\}$.
$B = \{\text{We here rain sound indoor}\}$.
We know $P(A) = 0.1$. This is a prior belief of whether it rains in Berkeley. And we also know the conditional probability $P(B|A) = 1$ and $P(B|A^c) = 0.2$ (might be some water pipes broken!).
Suppose we really hear rain sound indoor, we will update our belief of whether it is raining in Berkeley based on this new information. In other words, we calculate the posterior probability $P(A|B)$,

$$P(A|B) = \frac{0.1 * 1}{0.1 * 1 + 0.9 * 0.2} = 0.35$$

Bayes Theorem (cont.)

Given discrete random variable X and continuous random variable Y,

$$P(X = x|Y = y) = \frac{f_{Y|X=x}(y)P(X = x)}{f_Y(y)} \tag{2}$$

Python Resources

- CART: sklean.tree
- CV: https://scikit-learn.org/stable/modules/cross_validation.html