05.03.2018
MSL

Module 8: Tree-based methods

Regression trees

univariate continuous random variable

(x_i, Y_i) $i = 1, \dots, n$

p-dim predictor

$$Y = f(x) + \varepsilon$$

1) Divide predictor space into J non-overlapping regions, $R_1, \dots, R_J$
2) Prediction in R_j is $\hat{y}_j$ = mean of training obs that fall into R_j

How to decide on $R_1, \dots, R_J$?

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Recursive binary splitting:

at the current node split into $R_1(j,s) = \{x \mid x_j < s\}$

$R_2(j,s) = \{x \mid x_j \geq s\}$

and choose j and s to minimize

$$\sum_{i: x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2} (y_i - \hat{y}_{R_2})^2$$

Ozone $\leftarrow$ match R's and tree nodes.

Classification trees   K classes $\{1,..,K\}$

1) Prediction in R_j: majority vote or chosen cut-off on

$$\hat{p}_{jk} = \frac{1}{N_j} \sum_{i: x_i \in R_j} I(y_i = k)$$

region   class

2) Splitting criterion: minimize misclassification rate: $1 - \max \hat{p}_{jk}$

When to split?   → too crude

Q: are the child nodes on average "purer" than their parents? For region j

If one $p = 1$ → very pure: 0
all $p = \frac{1}{K}$ → not pure: large

$-\sum_k p_k \log p_k$ cross entropy

$\sum p_{jk}(1 - p_{jk})$ Gini

expected error rate if label is chosen at random from the class distribution at the node

statisticians prefer Gini