Decision Tree

Presenter: Jaehyeong Ahn

Department of Applied Statistics, Konkuk University

jayahn0104@gmail.com

Contents

Introduction

- Decision tree is one of the most widely used supervised learning method in machine learning
- Its big appeal is that the decision process is very much akin to how we humans make decisions
- Therefore it is easy to understand and accept the results coming from the tree-style decision process
- It used to be an baseline model when constructing some other algorithms (e.g. Random Forest, Boosting)

Decision Tree in a nutshell

Decision Tree Construction Process

1. Recursively partition the feature space $\mathfrak{X}$ into the subregions $t_1, \cdots, t_m$ based on the specific **splitting rule**
2. Stop splitting the feature space (or stop growing tree) when the **stopping rule** holds
3. Prune the grown tree based on the **pruning rule**
4. Make a (local) prediction on each subregion in pruned tree based on the specific **estimation method**

Four Ingredients in Decision Tree Construction

- **Splitting rule**
  - How does the split works?
  - What is the criterion for splitting the feature space?
- **Stopping rule**
  - What kind of threshold is used for stopping?
- **Pruning rule**
  - Why prune the tree?
  - How does it works?
- **Estimation method**
  - What kind of estimation method is used?
  - How does it works?

Notation

- $\mathfrak{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$: the dataset,
- Where $x^{(i)} = (x_1^{(i)}, \cdots, x_d^{(i)}) \in \mathfrak{X}$ and $y^{(i)} \in \mathfrak{Y}$, where $\mathfrak{Y} = \{1, \cdots, K\}$
- Let t be a node, which is also identified with a subregion of $\mathfrak{X}$
- $\mathfrak{D}(t) = \{(x^{(i)}, y^{(i)}) \in \mathfrak{D} : (x^{(i)}, y^{(i)}) \in t\}$: the set of data points in the node t
- Let T be a tree
- $N = |\mathfrak{D}|$: the total number of data points
- $N(t) = |\mathfrak{D}(t)|$: the number of data points in the node t
- $N_j(t) = |\{(x^{(i)}, y^{(i)} \in \mathfrak{D}(t) : y^{(i)} = j\}|$: the number of data points in the node t with class label $j \in \mathfrak{Y}$
- $p(j,t) = \frac{N_j}{N} \cdot \frac{N_j(t)}{N_j} = \frac{N_j(t)}{N}$
- $p(t) = \sum_j p(j,t) = \frac{N(t)}{N}$
- $p(j|t) = \frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N(t)}$

Splitting Rule

- For each node, determine a splitting variable x_j and splitting criterion c
- For continuous splitting variable, splitting criterion c is a number.
  - For example, if an observation $x^{(i)}$ is the case that its splitting variable $x_j^{(i)} < c$
  - Then tree assigns it to the left child node. Otherwise tree assigns it to the right child
- For categorical variable, the splitting criterion divides the range of the splitting variable in two parts
  - For example, let splitting variable $x_j \in \{1, 2, 3, 4\}$
  - And let the splitting criterion is $\{1, 2, 4\}$
  - If $x_j^{(i)} \in \{1, 2, 4\}$, tree assigns it to the left child. Otherwise, tree assigns it to the right child

Splitting Rule

- A split is determined based on the impurity
- Impurity (or purity) is the measure of homogeneity for a given node
- For each node, we select a splitting variable and a splitting criterion which minimizes the sum of impurities of the two child nodes
- Impurity is calculated by an $impurity\ function\ \phi$ which satisfies the following conditions

Splitting Rule

- **Definition 1.** An **impurity function** ϕ is a function $\phi(p_1, \cdots, p_K)$ defined for $p_1, \cdots, p_K$ with $p_j \geq 0$ for all j and $p_1 + \cdots + p_K = 1$ such that

  (i) $\phi(p_1, \cdots, p_K) \geq 0$

  (ii) $\phi(1/K, \cdots, 1/K)$ is the maximum value of ϕ

  (iii) $\phi(p_1, \cdots, p_K)$ is symmetric with regard to $p_1, \cdots, p_K$

  (iv) $\phi(1, 0, \cdots, 0) = \phi(0, 1, \cdots, 0) = \phi(0, \cdots, 0, 1) = 0$

- **Definition 2.** For node t, its impurity (measure) $i(t)$ is defined as

$$i(t) = \phi(p(1|t), \cdots, p(K|t))$$

Splitting Rule

Examples of impurity functions

- *Entropy impurity*

$$\phi(p_1, \cdots, p_K) = -\sum_j p_j \log p_j,$$

(Where we use the convention $0 \log 0 = 0$)

- *Gini impurity*

$$\phi(p_1, \cdots, p_K) = \frac{1}{2} \sum_j p_j (1 - p_j)$$

Splitting Rule

- **Definition 3.** The decrease in impurity
  - Let t be a node and let s be split of t into two child nodes $t_L \ and \ t_R$

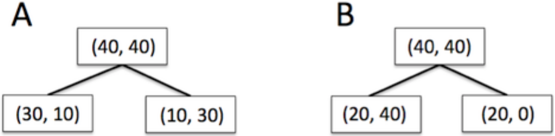
$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

  - Where $p_L = \frac{p(t_L)}{p(t)}$ and $p_R = \frac{p(t_R)}{p(t)}, \quad p_L + p_R =$1
  - Then $\Delta i(t) \geq$0 (see Proposition 4.4. in Breiman et al. (1984))
- Hence the splitting rule at t is s^* such that we take the split s^* among all possible candidate splits that decreases the cost most

$$s^* = \underset{s}{\operatorname{argmax}} \, \Delta i(s,t)$$

Splitting Rule

Why does not use misclassification error as an impurity function?



- Recall that $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ and $s^* = \underset{s}{\text{argmax}}\, \Delta i(s,t)$
- Misclassification error: $i_M(t) = 1 - \max_j p_j$, (where $p_j = p(j|t)$)
    - A: $\Delta i_M(s_A,t) = (1 - \frac{1}{2}) - (\frac{40}{80}) * (1 - \frac{3}{4}) - (\frac{40}{80}) * (1 - \frac{3}{4}) = \frac{1}{4}$
    - B: $\Delta i_M(s_B,t) = (1 - \frac{1}{2}) - (\frac{60}{80}) * (1 - \frac{4}{6}) - (\frac{20}{80}) * (1 - 1) = \frac{1}{4}$
- Entropy impurity: $i_E(t) = -\sum_j p_j log(p_j)$
    - A: $\Delta i_E(s_A,t) = 0.130812$
    - B: $\Delta i_E(s_B,t) = 0.2157616$

Stopping Rule

- Stopping rules terminate further splitting
- For example
  - All observations in a node are contained in one group
  - The number of observations in a node is small
  - The decrease of impurity is small
  - The depth of a node is larger than a given number

Pruning Rule

- A tree with too many nodes will have large prediction error rate for new observations
- It is appropriate to prune away some branch of tree for good prediction error rate
- To determine the size of tree, we estimate prediction error using validation set or cross validation

Pruning Rule

Pruning Process

- For a given tree T and positive number α, cost-complexity pruning is defined by

$$\text{cost-complexity}(\alpha) = \text{error rate of } T + \alpha|T|$$

(Where $|T|$ is the number of nodes)

- In general, the larger tree(the larger $|T|$), the smaller error rate (see Proposition 1. in Hyeongin Choi (2017)). But cost-complexity does not decrease as $|T|$ increases.

- For the grown tree T_{max}, $T(\alpha)$ is a subtree which minimizes cost-complexity(α)

- In general, the larger α, the smaller $|T(\alpha)|$

Pruning Process

- One important property of $T(\alpha)$ is that if $\alpha_1 \leq \alpha_2$, then $T(\alpha_1) \geq T(\alpha_2)$ which means $T(\alpha_2)$ is a subtree of $T(\alpha_1)$
- Let $T_1 = T(\alpha_1),\ T_2 = T(\alpha_2), \cdots$ then we can get the following sequence of pruned subtrees

$$T_1 \geq T_2 \geq \cdots \geq \{t_1\},$$

$$\text{where } 0 = \alpha_1 < \alpha_2 < \cdots$$

- For a given α, we estimate the generalization error of $T(\alpha)$ by validation set or cross-validation
- Choose α^* (and corresponding $T(\alpha^*)$) which minimizes the (estimated) generalization error. (see Hyeongin Choi (2017) for details)

Estimation Method

- Once a tree is fixed, a prediction at each subregion (terminal node) can be determined from that tree
- Since the estimation is operated at each subregion, it is called a local estimation
- For this local estimation, any estimation method can be obtained. What kind of method to use is dependent on the model assumption.
- For example, for classification, $majority\ vote$ or $logistic\ regression$ is obtained by modeler's probability assumption

Some Algorithms for Decision Tree

CHAID (CHi-squared Automatic Interaction Detector)

- by J. A. Hartigan 1975
- Employes χ^2 statistic as impurity
- No pruning process, it stops growing at a certain size

CART (Classification And Regression Tree)

- by Breiman and et al. 1984
- Only binary split is operated
- Cost-complexity pruning is an important unique feature

C5.0 (successor of ID3 and C4.5)

- by J. Ross Quinlan 1993
- Multisplit is available
- For categorical input variable, a node splits into the number of categories.

Advantages and Disadvantages

Advantages

- Easy to interpret and explain
- Trees can be displayed graphically
- Trees can easily handle both continuous and categorical variables without the need to create dummy variables

Disadvantages

- Poor prediction accuracy compared to other models
- When depth is large, not only accuracy but interpretation are bad

References

1 Breiman, L. Friedman, J.H, Olshen, R.A. and Stone, C.J., *Classification And Regression Trees*, Chapman & Hall/CRC (1984)

2 Hyeongin Choi (2017), Lecture 9: Classification and Regreesion Tree (CART), http://www.math.snu.ac.kr/∼*hichoi/machinelearning/lecturenotes/CART.pdf*

3 Yongdai Kim (2017), Chapter 6. Decision Tree, https://stat.snu.ac.kr/ydkim/courses/2017-1/addm/Chap6-DecisionTree.pdf