tree

Andrea Spano’

2023-12-18

Introduction

- Decision Tree is one of the commonly used exploratory data analysis and objective segmentation techniques.
- Great advantage with Decision Tree is that its output is relatively easy to understand or interpret.
- A decision tree is a recursive hierarchical partitioning of the input data: at each node (step) a specific value of one of the independent variables is used for the partition.
- Tree-based methods can be used for both regression and classification problems.

Introduction

- The building of a tree is usually produced in two phases: *growth* and *pruning*.
- To grow a classification tree, a binary splitting is used.

Splitting

- To split the nodes, the minimum *within-node variability*, is searched.
- Variability is usually measured with three alternative indices:
    - Gini index.
    - Entropy
    - Classification Error

Assume a class made of: $4A$, $3B$ and $3C$ for a total of 10 observations, the probability (frequency) of each class:

- $P(A) = 0.4$, $P(B) = 0.3$ and $P(C) = 0.3$

Gini index

$$G = 1 - \sum_i {p_i}^2 = 0.66$$

- $G = 0$ for a *pure* class
- $max(G) = 1$
- The value of G is always between 0 and 1 regardless the size of N

Entropy

$$E = \sum_i -p_i \times log_2(p_i) = 1.571$$

- $E = 0$ for a *pure* class
- $max(E) = -n \times p \times log_2(p)$
- The value of E is larger than 1 if the number of classes is larger than 2
- The value of $max(E)$ increases as N increases

Classification Error

$$CE = 1 - max(p_i)$$

- $CE = 0$ for a *pure* class
- $max(CE) = 1$
- The value of CE is always between 0 and 1 regardless the size of N

Splitting

- The R rpart algorithm offers both entropy and Gini index methods as splitting choices
- The algorithm stops splitting when cp: complexity parameters reaches a given threshold
- There is a fair amount of fact and opinion about which method is better
- The answer as to which method is the best is: it depends. Try both

Splitting

- The algorithm works by making the best possible choice at each particular stage, without any consideration of whether those choices remain optimal in future stages.
- That is, the algorithm makes a locally optimal decision at each stage
- It is thus quite possible that such a choice at one stage turns out to be sub-optimal in the overall scheme
- In other words, the algorithm does not find a globally optimal tree.

bias-variance tradeoff

- Bias-variance tradeoff in machine learning is a tradeoff between:
  - the degree to which a model fits the training data
  - its predictive accuracy
- This refers to the general rule that beyond a point, it is counterproductive to improve the fit of a model to the training data as this increases the likelihood of overfitting
- It is easy to see that deep trees are more likely to overfit the data than shallow ones.

bias-variance tradeoff

- One obvious way to control such overfitting is to construct shallower trees by stopping the algorithm at an appropriate point based on whether a split significantly improves the fit.
- Another is to grow a tree unrestricted and then prune it back using an appropriate criterion.
- The rpart algorithm takes the latter approach.

bias-variance tradeoff

- The algorithm minimises the cost, $C_\alpha(T)$, a quantity that is a linear combination of:
    - the error $R(T)$
    - the number of leaf nodes in the tree, $|\tilde{T}|$:

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

- The error being:
    - The fraction of misclassified instances for a discrete variable
    - Variance in the case of a continuous variable,

bias-variance tradeoff

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

- When $\alpha = 0$, this simply returns the original fully grown tree.
- As α increases, we incur a penalty that is proportional to the number of leaf nodes
- In practice we vary α and pick the value that gives the subtree that results in the smallest cross-validated prediction error. ed to do is pick the value of the coefficient that gives the lowest cross-validated error
- We usually set a lower threshold for α. $\alpha = 0.01$ by default in rpart

Pruning

- Pruning the tree is about selecting the number of terminal nodes that minimize the cost $C_\alpha(T)$
- In practice this is achieved by imposing a desired cp threshold