Decision Trees

Nguyen Trung Hieu

Hanoi University of Science and Technology

May 29, 2019

Outline

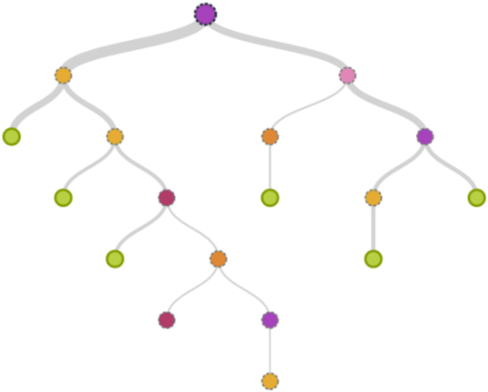
Introduction

Getting started

A mathematical theorem

Decision tree model

"A decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path." [1]



[1] Data science from Scratch

Decision-tree learning algorithms

Some notable algorithms

1. **Iterative Dichotomiser (ID3)**:
   for data with categorical features
2. **C4.5**:
   can handle both categorical and numerical features
3. **Classification And Regression Tree (CART)**:
   improved version of ID3

Building blocks of a DTL algorithm

- **Loss function** entropy, gini index
- **Stopping criteria**
- **Pruning**

Loss function

For a the sample set S which contains n classes: $C_1, C_2, ..., C_n$.
Let $p(C_i)$ be the portion of class C_i in S.

Entropy

$$H(S) = -\sum_{i=1}^{n} p(C_i)\log_2 p(C_i)$$

Gini index

$$G(S) = \sum_{i=1}^{n} p(C_i)(1 - p(C_i))$$

Loss function

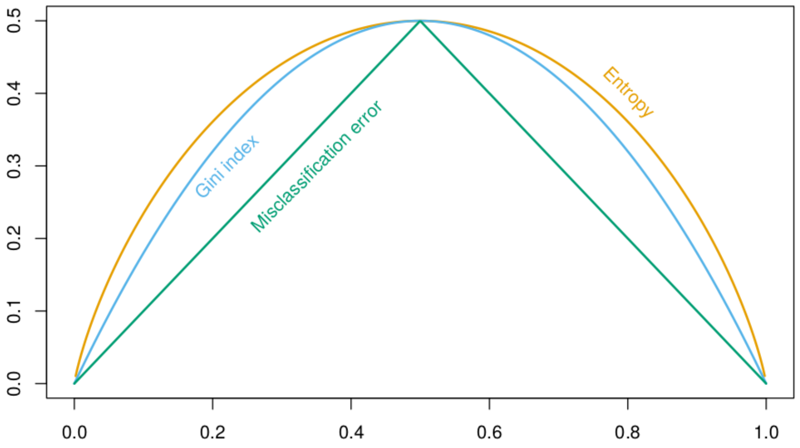


Figure: Node impurity measures [1]

[1] Elements of Statistical Learning, p.309

Loss function

"The truth is, most of the time it does not make a big difference: they lead to similar trees. Gini impurity is slightly faster to compute, so it is a good default. However, when they differ, Gini impurity tends to isolate the most frequent class in its own branch of the tree, while entropy tends to produce slightly more balanced trees." [1]

[1] Hands-on machine learning with Scikit-Learn and TensorFlow, p.184

Stopping criteria

- Entropy achieves zero.
- Number of samples belong to a note gets below a threshold.
- Reach tree depth limit.
- Reach number of nodes limit.
- Information gain is less than a threshold.

Pruning

- Reduced error pruning
- Cost complexity pruning

Cost complexity pruning

- Generate a series of trees $T_0, ..., T_m$.
  tree T_i is generated from tree T_{i-1} by replace a subtree by a leaf node.
- The subtree to be removed is chosen by:

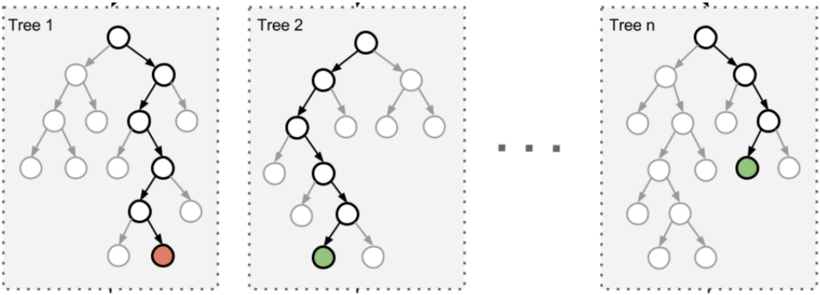
$$error\_rate\_per\_pruned\_leaf = \frac{err(prune(T,t), S - err(T,S)}{|leaves(T)| - |leaves(prune(T,t))|} \tag{1}$$

  The function $prune(T,t)$ define the tree gotten by remove sub tree t from T, $err(T,S)$ is the error of tree T with respect to the set S
- The best tree is chosen by a measure such as cross-validation

Random forest

Motivation

In order to reduce the effect of overfitting of a model with the training set, the output is averaged over the results of multiple models.



Implementations

Implementations

- **Decision tree with sklearn:** Decision tree sklearn
- **Random forest with sklearn:** Random forest sklearn

Q & A

References

References

Grus Joel
Data science from scratch: first principles with python
2019 O'Reilly Media

Trevor Hastie, Robert Tibshirani, Jerome Friedman
The Elements of Statistical Learning
2013 Springer

Vu Huu Tiep
Machine learning co ban
`https://machinelearningcoban.com/2018/01/14/id3/`

Géron, Aurélien
Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems
2017 O'Reilly Media