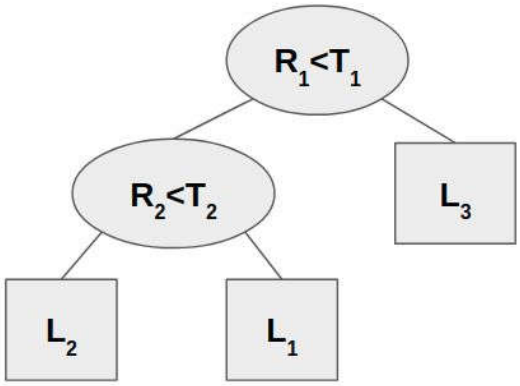
Decision Trees and Random Forests - Summary

Lucas D. Lo Vercio - Statistical Learning Study Group

Decision tree

A decision tree is a hierarchical data structure implementing the divide-and-conquer strategy. It is an efficient nonparametric method, which can be used for both classification and regression.



It is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree is composed of internal decision nodes and terminal leaves (leaf). (Alpaydin, 2010).

The boundaries of the regions are defined by the discriminants that are coded in the internal nodes.

Impurity measures

A split is pure if after the split, for all branches, all the instances choosing a branch belong to the same class.

Entropy is one option: $I_m = -\sum_{i=1}^{K} p_m{}^i log_2 p_m{}^i$,

where K is the number of classes and $p_m{}^i$ is the probability of class *i* to node m.

Also, for 2-class problems:
Gini index (Breiman et al. 1984): $\phi(p, 1-p) = 2p(1-p)$
Misclassification error: $\phi(p, 1-p) = 1 - max(p, 1-p)$

Improving generalization

Growing the tree until it is maximally pure, in the presence of noisy samples and unclear regions, may result in a deep tree. As a result, the decision tree will overfit to the training data at the cost of an appropriate generalization.

To avoid overfitting, a tolerance to the impurity can be set. Additionally, limiting the number of minimum samples per leaf, via pre-pruning or post-pruning, can also help.

Random Forest

Random forests (Breiman, 2001) use bagging, which builds a large collection of de-correlated noisy trees and then averages the output. Before each split, random input variables or features (*q*) are selected as candidates for splitting.

Recommended parameters

Number of trees = 100-500
Random input variables = $\sqrt{q}$ - $q/2$
Minimum node size = 1

Out-of-bag error

For each sample *z* in the training set, average only those trees where *z* was not included. This out-of-bag error (OOB) estimation is close to the N-fold cross validation.

Variable importance

At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable. (Hastie et al, 2009)

References

- Leo Breiman. Random forests. Machine Learning, 45(1):5{32, 2001. ISSN 1573-0565.
- Ethem Alpaydin. Introduction to Machine Learning. 2nd. The MIT Press, 2010. ISBN 978-0-262-01243-0
- Trevor Hastie, Robert Tibshirani, y Jerome Friedman. The elements of statistical learning: Data mining, inference, and prediction. Springer, 2009.