



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Featured content](#)  
[Current events](#)  
[Random article](#)  
[Donate to Wikipedia](#)  
[Wikipedia store](#)

Interaction  
[Help](#)  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact page](#)

Tools  
[What links here](#)  
[Related changes](#)  
[Upload file](#)  
[Special pages](#)  
[Permanent link](#)  
[Page information](#)  
[Wikidata item](#)  
[Cite this page](#)

Print/export  
[Create a book](#)  
[Download as PDF](#)  
[Printable version](#)

Languages  
[Español](#)  
[Français](#)  
[한국어](#)  
[עברית](#)  
[Latviešu](#)  
[中文](#)

 [Edit links](#)

[Create account](#) [Log in](#)

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#)

# Decision tree learning

From Wikipedia, the free encyclopedia

*This article is about decision trees in machine learning. For the use of the term in decision analysis, see [Decision tree](#).*

**Decision tree learning** uses a [decision tree](#) as a [predictive model](#) which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in [statistics](#), [data mining](#) and [machine learning](#). Tree models where the target variable can take a finite set of values are called **classification trees**. In these tree structures, [leaves](#) represent class labels and branches represent [conjunctions](#) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](#)) are called **regression trees**.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and [decision making](#). In [data mining](#), a decision tree describes data but not decisions; rather the resulting classification tree can be an input for [decision making](#). This page deals with decision trees in [data mining](#).

## Contents

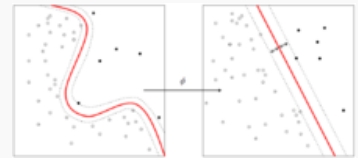
- 1 General
- 2 Types
- 3 Metrics
  - 3.1 Gini impurity
  - 3.2 Information gain
  - 3.3 Variance reduction
- 4 Decision tree advantages
- 5 Limitations
- 6 Extensions
  - 6.1 Decision graphs
  - 6.2 Alternative search methods
- 7 See also
- 8 Implementations
- 9 References
- 10 External links

## General

Decision tree learning is a method commonly used in data mining.<sup>[1]</sup> The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each [interior node](#) corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning<sup>[*citation needed*]</sup>. For this section, assume that all of the features have finite discrete domains, and there is a single target feature called the classification. Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

## Machine learning and data mining



### Problems

[Classification](#) · [Clustering](#) · [Regression](#) · [Anomaly detection](#) · [Association rules](#) · [Reinforcement learning](#) · [Structured prediction](#) · [Feature learning](#) · [Online learning](#) · [Semi-supervised learning](#) · [Unsupervised learning](#) · [Learning to rank](#) · [Grammar induction](#)

### Supervised learning

([classification](#) · [regression](#))

**Decision trees** · [Ensembles \(Bagging, Boosting, Random forest\)](#) · [k-NN](#) · [Linear regression](#) · [Naive Bayes](#) · [Neural networks](#) · [Logistic regression](#) · [Perceptron](#) · [Support vector machine \(SVM\)](#) · [Relevance vector machine \(RVM\)](#)

### Clustering

[BIRCH](#) · [Hierarchical](#) · [k-means](#) · [Expectation-maximization \(EM\)](#) · [DBSCAN](#) · [OPTICS](#) · [Mean-shift](#)

### Dimensionality reduction

[Factor analysis](#) · [CCA](#) · [ICA](#) · [LDA](#) · [NMF](#) · [PCA](#) · [t-SNE](#)

### Structured prediction

[Graphical models \(Bayes net, CRF, HMM\)](#)

### Anomaly detection

[k-NN](#) · [Local outlier factor](#)

### Neural nets

[Autoencoder](#) · [Deep learning](#) · [Multilayer perceptron](#) · [RNN](#) · [Restricted Boltzmann machine](#) · [SOM](#) · [Convolutional neural network](#)

### Theory

[Bias-variance dilemma](#) · [Computational learning theory](#) · [Empirical risk minimization](#) · [PAC learning](#) · [Statistical learning](#) · [VC theory](#)



[Machine learning portal](#)



[Computer science portal](#)



[Statistics portal](#)

v · t · e

A tree can be "learned" by splitting the source [set](#) into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called [recursive partitioning](#). The [recursion](#) is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of *top-down induction of decision trees* (TDIDT) <sup>[2]</sup> is an example of a [greedy algorithm](#), and it is by far the most common strategy for learning decision trees from data.

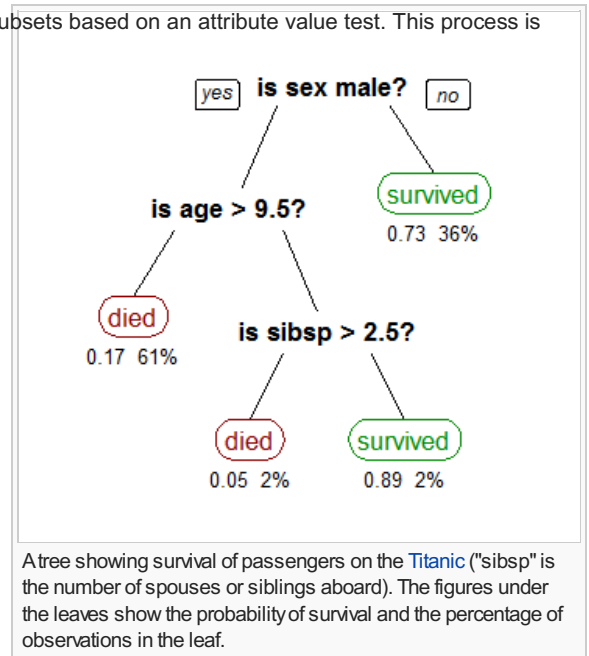
In [data mining](#), decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data.

Data comes in records of the form:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, \dots, x_k, Y)$$

The dependent variable,  $Y$ , is the target variable that we are trying to understand, classify or generalize.

The vector  $\mathbf{x}$  is composed of the input variables,  $x_1, x_2, x_3$  etc., that are used for that task.



## Types <sup>[edit]</sup>

Decision trees used in [data mining](#) are of two main types:

- **Classification tree** analysis is when the predicted outcome is the class to which the data belongs.
- **Regression tree** analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

The term **Classification And Regression Tree (CART)** analysis is an [umbrella term](#) used to refer to both of the above procedures, first introduced by [Breiman](#) et al.<sup>[3]</sup> Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.<sup>[3]</sup>

Some techniques, often called *ensemble* methods, construct more than one decision tree:

- **Bagging** decision trees, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction.<sup>[4]</sup>
- A **Random Forest** classifier uses a number of decision trees, in order to improve the classification rate.
- **Boosted Trees** can be used for regression-type and classification-type problems.<sup>[5][6]</sup>
- **Rotation forest** - in which every decision tree is trained by first applying [principal component analysis](#) (PCA) on a random subset of the input features.<sup>[7]</sup>

**Decision tree learning** is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

There are many specific decision-tree algorithms. Notable ones include:

- **ID3** (Iterative Dichotomiser 3)
- **C4.5** (successor of ID3)
- **CART** (Classification And Regression Tree)
- **CHAID** (CHI-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.<sup>[8]</sup>
- **MARS**: extends decision trees to handle numerical data better.
- **Conditional Inference Trees**. Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid overfitting. This approach results in unbiased predictor selection and does not require pruning.<sup>[9][10]</sup>

ID3 and CART were invented independently at around the same time (between 1970 and 1980)<sup>[citation needed]</sup>, yet follow a similar approach for learning decision tree from training tuples.

## Metrics <sup>[edit]</sup>

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best

splits the set of items.<sup>[11]</sup> Different algorithms use different metrics for measuring "best". These generally measure the homogeneity of the target variable within the subsets. Some examples are given below. These metrics are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split.

### Gini impurity [\[edit\]](#)

*Not to be confused with [Gini coefficient](#).*

Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

To compute Gini impurity for a set of items, suppose  $i \in \{1, 2, \dots, m\}$ , and let  $f_i$  be the fraction of items labeled with value  $i$  in the set.

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i) = \sum_{i=1}^m (f_i - f_i^2) = \sum_{i=1}^m f_i - \sum_{i=1}^m f_i^2 = 1 - \sum_{i=1}^m f_i^2 = \sum_{i \neq k} f_i f_k$$

### Information gain [\[edit\]](#)

*Main article: [Information gain in decision trees](#)*

Used by the ID3, C4.5 and C5.0 tree-generation algorithms. [Information gain](#) is based on the concept of [entropy](#) from [information theory](#).

$$I_E(f) = - \sum_{i=1}^m f_i \log_2 f_i$$

### Variance reduction [\[edit\]](#)

Introduced in CART,<sup>[3]</sup> variance reduction is often employed in cases where the target variable is continuous (regression tree), meaning that use of many other metrics would first require discretization before being applied.

The variance reduction of a node  $N$  is defined as the total reduction of the variance of the target variable  $x$  due to the split at this node:

$$I_V(N) = \frac{1}{|S|^2} \sum_{i \in S} \sum_{j \in S} \frac{1}{2} (x_i - x_j)^2 - \left( \frac{1}{|S_t|^2} \sum_{i \in S_t} \sum_{j \in S_t} \frac{1}{2} (x_i - x_j)^2 + \frac{1}{|S_f|^2} \sum_{i \in S_f} \sum_{j \in S_f} \frac{1}{2} (x_i - x_j)^2 \right)$$

where  $S$ ,  $S_t$  and  $S_f$  are the set of presplit sample indices, set of sample indices for which the split test is true, and set of sample indices for which the split test is false, respectively. Each of the above summands are indeed [variance](#) estimates, though, written in a form without directly referring to the mean.

## Decision tree advantages [\[edit\]](#)




This section **does not cite any references or sources**. Please help improve this section by [adding citations to reliable sources](#). Unsourced material may be challenged and [removed](#). *(July 2015)*

Amongst other data mining methods, decision trees have various advantages:

- **Simple to understand and interpret.** People are able to understand decision tree models after a brief explanation.
- **Requires little data preparation.** Other techniques often require data normalisation, [dummy variables](#) need to be created and blank values to be removed.
- **Able to handle both numerical and [categorical](#) data.** Other techniques are usually specialised in analysing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.)
- **Uses a [white box](#) model.** If a given situation is observable in a model the explanation for the condition is easily explained by boolean logic. (An example of a black box model is an [artificial neural network](#) since the explanation for the results is difficult to understand.)
- **Possible to validate a model using statistical tests.** That makes it possible to account for the reliability of the model.
- **Robust.** Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- **Performs well with large datasets.** Large amounts of data can be analysed using standard computing

resources in reasonable time.

## Limitations [\[edit\]](#)

- The problem of learning an optimal decision tree is known to be **NP-complete** under several aspects of optimality and even for simple concepts.<sup>[12][13]</sup> Consequently, practical decision-tree learning algorithms are based on heuristics such as the **greedy algorithm** where locally-optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally-optimal decision tree. To reduce the greedy effect of local-optimality some methods such as the dual information distance (DID) tree were proposed.<sup>[14]</sup> <sup>[1]</sup> 
- Decision-tree learners can create over-complex trees that do not generalise well from the training data. (This is known as **overfitting**.<sup>[15]</sup>) Mechanisms such as **pruning** are necessary to avoid this problem (with the exception of some algorithms such as the Conditional Inference approach, that does not require pruning <sup>[9][10]</sup>).
- There are concepts that are hard to learn because decision trees do not express them easily, such as **XOR**, **parity** or **multiplexer** problems. In such cases, the decision tree becomes prohibitively large. Approaches to solve the problem involve either changing the representation of the problem domain (known as propositionalisation)<sup>[16]</sup> or using learning algorithms based on more expressive representations (such as **statistical relational learning** or **inductive logic programming**).
- For data including categorical variables with different numbers of levels, **information gain in decision trees** is biased in favor of those attributes with more levels.<sup>[17]</sup> However, the issue of biased predictor selection is avoided by the Conditional Inference approach.<sup>[9]</sup>

## Extensions [\[edit\]](#)

### Decision graphs [\[edit\]](#)

In a decision tree, all paths from the root node to the leaf node proceed by way of conjunction, or *AND*. In a decision graph, it is possible to use disjunctions (ORs) to join two more paths together using **Minimum message length** (MML).<sup>[18]</sup> Decision graphs have been further extended to allow for previously unstated new attributes to be learnt dynamically and used at different places within the graph.<sup>[19]</sup> The more general coding scheme results in better predictive accuracy and log-loss probabilistic scoring.<sup>[citation needed]</sup> In general, decision graphs infer models with fewer leaves than decision trees.

### Alternative search methods [\[edit\]](#)

Evolutionary algorithms have been used to avoid local optimal decisions and search the decision tree space with little *a priori* bias.<sup>[20][21]</sup>



It is also possible for a tree to be sampled using MCMC.<sup>[22]</sup>

The tree can be searched for in a bottom-up fashion.<sup>[23]</sup>

## See also [\[edit\]](#)

- **Decision tree pruning**
- **Binary decision diagram**
- **CHAID**
- **CART**
- **ID3 algorithm**
- **C4.5 algorithm**
- **Decision stump**
- **Incremental decision tree**
- **Alternating decision tree**
- **Structured data analysis (statistics)**
- **Logistic model tree**

## Implementations [\[edit\]](#)

Many data mining software packages provide implementations of one or more decision tree algorithms. Several examples include Salford Systems CART (which licensed the proprietary code of the original CART authors<sup>[3]</sup>), **IBM SPSS Modeler**, **RapidMiner**, **SAS Enterprise Miner**, **Matlab**, **R** (an open source software environment for statistical computing which includes several CART implementations such as **rpart**, **party** and **randomForest** packages), **Weka** (a free and open-source data mining suite, contains many decision tree algorithms), **Orange** (a free data mining software suite, which includes the tree module **orngTree** ) , **KNIME**, **Microsoft SQL Server** <sup>[2]</sup> , and **scikit-learn** (a free and open-source machine learning library for the **Python** programming language).

## References [[edit](#)]

- ↑ Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.
- ↑ Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers
- ↑  <sup>*a*  *b*  *c*  *d*</sup> Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.
- ↑ Breiman, L. (1996). Bagging Predictors. "Machine Learning, 24": pp. 123-140.
- ↑ Friedman, J. H. (1999). *Stochastic gradient boosting*. Stanford University.
- ↑ Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The elements of statistical learning : Data mining, inference, and prediction*. New York: Springer Verlag.
- ↑ Rodriguez, J.J. and Kuncheva, L.I. and Alonso, C.J. (2006), Rotation forest: A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1619-1630.
- ↑ Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data". *Applied Statistics* **29** (2): 119–127. doi:10.2307/2986296. JSTOR 2986296.
- ↑  <sup>*a*  *b*  *c*</sup> Hothorn, T.; Hornik, K.; Zeileis, A. (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* **15** (3): 651–674. doi:10.1198/106186006X133933. JSTOR 27594202.
- ↑  <sup>*a*  *b*</sup> Strobl, C.; Malley, J.; Tutz, G. (2009). "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests". *Psychological Methods* **14** (4): 323–348. doi:10.1037/a0016973.
- ↑ Rokach, L.; Maimon, O. (2005). "Top-down induction of decision trees classifiers-a survey". *IEEE Transactions on Systems, Man, and Cybernetics, Part C* **35** (4): 476–487. doi:10.1109/TSMCC.2004.843247.
- ↑ Hyafil, Laurent; Rivest, RL (1976). "Constructing Optimal Binary Decision Trees is NP-complete". *Information Processing Letters* **5** (1): 15–17. doi:10.1016/0020-0190(76)90095-8.
- ↑ Murthy S. (1998). Automatic construction of decision trees from data: A multidisciplinary survey. *Data Mining and Knowledge Discovery*
- ↑ Ben-Gal I. Dana A., Shkolnik N. and Singer (20). "Efficient Construction of Decision Trees by the Dual Information Distance Method"  (PDF). Quality Technology & Quantitative Management (QTQM), 11( 1), 133-147. Check date values in: |date= (help)
- ↑ "Principles of Data Mining". 2007. doi:10.1007/978-1-84628-766-4. ISBN 978-1-84628-765-7.
- ↑ Horváth, Tamás; Yamamoto, Akihiro, eds. (2003). "Inductive Logic Programming". Lecture Notes in Computer Science **2835**. doi:10.1007/b13700. ISBN 978-3-540-20144-1.
- ↑ Deng,H.; Runger, G.; Tuv, E. (2011). *Bias of importance measures for multi-valued attributes and solutions*. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). pp. 293–300.
- ↑ http://citeseer.ist.psu.edu/oliver93decision.html
- ↑ Tan & Dowe (2003)
- ↑ Papagelis A., Kalles D.(2001). Breeding Decision Trees Using Evolutionary Techniques, Proceedings of the Eighteenth International Conference on Machine Learning, p.393-400, June 28-July 01, 2001
- ↑ Barros, Rodrigo C., Basgalupp, M. P., Carvalho, A. C. P. L. F., Freitas, Alex A. (2011). *A Survey of Evolutionary Algorithms for Decision-Tree Induction*. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42, n. 3, p. 291-312, May 2012.
- ↑ Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "Bayesian CART model search." Journal of the American Statistical Association 93.443 (1998): 935-948.
- ↑ Barros R. C., Cerri R., Jaskowiak P. A., Carvalho, A. C. P. L. F., *A bottom-up oblique decision tree induction algorithm*. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011).

## External links [[edit](#)]

- Building Decision Trees in Python From O'Reilly.
- An Addendum to "Building Decision Trees in Python" From O'Reilly.
- Decision Trees Tutorial using Microsoft Excel.
- Decision Trees page at aitopics.org, a page with commented links.
- Decision tree implementation in Ruby (AI4R)
- Evolutionary Learning of Decision Trees in C++
- Java implementation of Decision Trees based on Information Gain
- A very explicit explanation of information gain as splitting criterion

Categories:  Decision trees | Classification algorithms

This page was last modified on 28 August 2015, at 19:00.

Text is available under the  Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the  Terms of Use and  Privacy Policy. Wikipedia® is a registered trademark of the  Wikimedia Foundation, Inc., a non-profit organization.

