Дерево решений (Decision Trees)

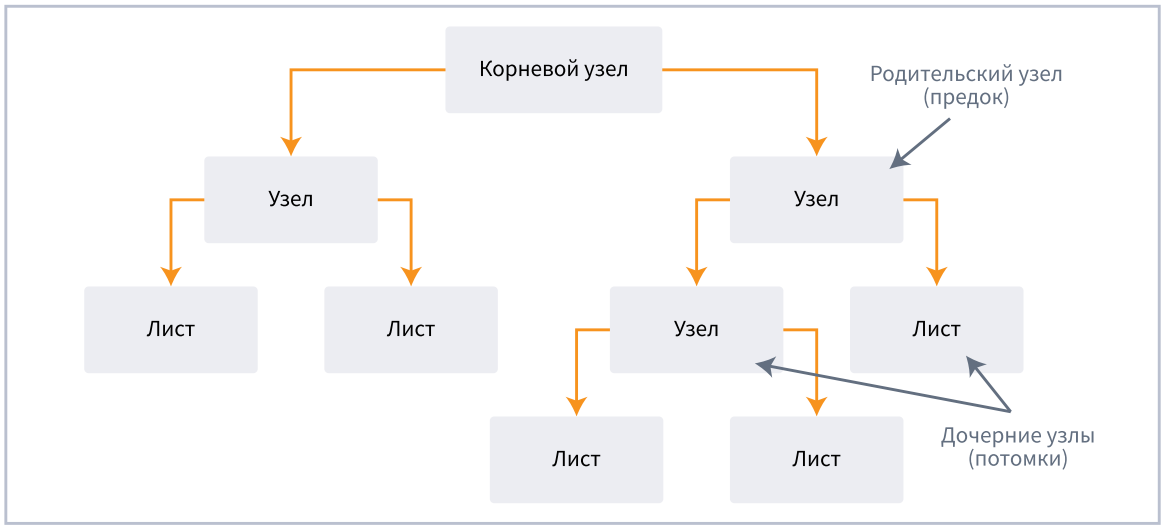
> Синонимы: Дерево классификаций, Classification Tree
> Разделы: Алгоритмы

Дерево решений — классификатор, построенный на основе решающих правил вида «если, то», упорядоченных в древовидную иерархическую структуру.

В основе работы дерева решений лежит процесс рекурсивного разбиения исходного множества объектов на подмножества, ассоциированные с предварительно заданными классами. Разбиение производится с помощью решающих правил, в которых осуществляется проверка значений атрибутов по заданному условию.

Строятся на основе обучения с учителем. В качестве обучающего набора данных используется множество наблюдений, для которых предварительно задана метка класса.

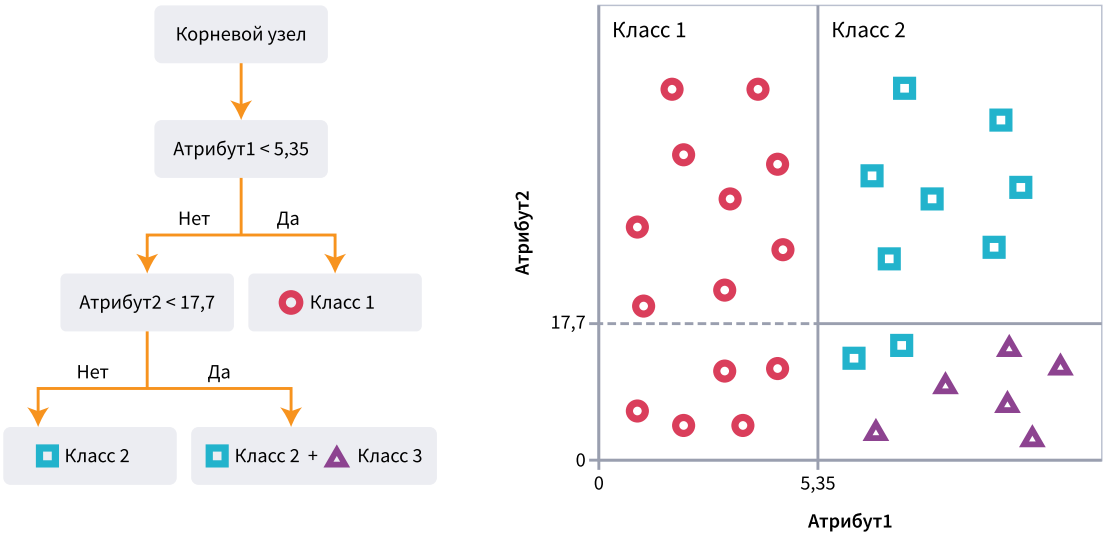
Структурно дерево решений состоит из объектов двух типов — узлов (node) и листьев (leaf). В узлах расположены решающие правила и подмножества наблюдений, которые им удовлетворяют. В листьях содержатся классифицированные деревом наблюдения: каждый лист ассоциируется с одним из классов, и объекту, который распределяется в лист, присваивается соответствующая метка класса.



Визуально узлы и листья в дереве хорошо различимы: в узлах указываются правила, разбивающие содержащиеся в нем наблюдения, и производится дальнейшее ветвление. В листьях правил нет, они помечаются меткой класса, объекты которого попали в данный лист. Ветвление в листьях не производится, и они заканчивают собой ветвь дерева (поэтому их иногда называют терминальными узлами).

Если класс, присвоенный деревом, совпадает с целевым классом, то объект является распознанным, в противном случае — нераспознанным. Самый верхний узел дерева называется корневым (root node). В нем содержится весь обучающий или рабочий набор данных.

Дерево решений является линейным классификатором, т.е. производит разбиение объектов в многомерном пространстве плоскостями (в двумерном случае — линиями).



Дерево, представленное на рисунке, решает задачу классификации объектов по двум атрибутам на три класса.

На рисунке кружки представляют объекты класса 1, квадраты — класса 2, а треугольники — класса 3. Пространство признаков разделено линиями на три подмножества, ассоциированных с классами. Эти же подмножества будут соответствовать и трем возможным исходам классификации. В классе «треугольников» имеются нераспознанные примеры («квадраты»), т.е. примеры, попавшие в подмножества, ассоциированные с другим классом.

Теоретически, алгоритм может генерировать новые разбиения до тех пор, пока все примеры не будут распознаны правильно, т.е. пока подмножества, ассоциированные с листьями, не станут однородными по классовому составу. Однако это приводит к усложнению дерева: большое число ветвлений, узлов и листьев усложняет его структуру и ухудшает его интерпретируемость. Поэтому на практике размер дерева ограничивают даже за счет некоторой потери точности. Данный процесс называется упрощением деревьев решений и может быть реализован с помощью методов ранней остановки и отсечения ветвей.

Деревья решений являются жадными алгоритмами. Могут быть дихотомичными (бинарными), имеющими только два потомка в узле, и полихотомичными — имеющими более 2-х потомков в узле. Дихотомичные деревья являются более простыми в построении и интерпретации.

В настоящее время деревья решений стали одним из наиболее популярных методов классификации в интеллектуальном анализе данных и бизнес-аналитике. Поэтому они входят в состав практически любого аналитического ПО.

Разработано большое количество различных алгоритмов построения деревьев решений. Наиболее известным является семейство алгоритмов, основанное на критерии прироста информации (information gain) — ID3, C4.5, C5.0, — предложенное Россом Куинленом в начале 1980-х.

Также широкую известность приобрел алгоритм CART (Classification and Regression Tree — дерево классификации и регрессии), который, как следует из названия, позволяет решать не только задачи классификации, но и регрессии. Алгоритм предложен Лео Брейманом в 1982 г.

Широкая популярность деревьев решений обусловлена следующими их преимуществами:

- правила в них формируются практически на естественном языке, что делает объясняющую способность деревьев решений очень высокой;
- могут работать как с числовыми, так и с категориальными данными;
- требуют относительно небольшой предобработки данных, в частности, не требуют нормализации, создания фиктивных переменных, могут работать с пропусками;
- могут работать с большими объемами данных.

Вместе с тем, деревьям решений присущ ряд ограничений:

- неустойчивость — даже небольшие изменения в данных могут привести к значительным изменениям результатов классификации;
- поскольку алгоритмы построения деревьев решений являются жадными, они не гарантируют построения оптимального дерева;
- склонность к переобучению.

В настоящее время деревья решений продолжают развиваться: создаются новые алгоритмы (SHAID, MARS, Random Forest) и их модификации, изучаются проблемы построения ансамблей моделей на основе деревьев решений.

Деревья решений становятся важным инструментом управления бизнес-процессами и поддержки принятия решений. Общие принципы работы и области применения описаны в статье «Деревья решений: общие принципы».