

# Дерево решений

Материал из Википедии — свободной энциклопедии

**Дерево принятия решений** (также называют деревом классификации или регрессионным деревом) — средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны признаки, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — признаки, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Каждый лист представляет собой значение целевой переменной, изменённой в ходе движения от корня по рёбрам дерева до листа. Каждый внутренний узел сопоставляется с одной из входных переменных.

Дерево может быть также «изучено» разделением исходных наборов переменных на подмножества, основанные на проверке значений признаков. Это действие повторяется на каждом из полученных подмножеств. Рекурсия завершается тогда, когда подмножество в узле имеет те же значения целевой переменной, таким образом, оно не добавляет ценности для предсказаний. Процесс, идущий «сверху вниз», индукция деревьев решений (TDIDT)<sup>[1]</sup>, является примером поглощающего «жадного» алгоритма, и на сегодняшний день является наиболее распространённой стратегией деревьев решений для данных, но это не единственная возможная стратегия.

В интеллектуальном анализе данных, деревья решений могут быть использованы в качестве математических и вычислительных методов, чтобы помочь описать, классифицировать и обобщить набор данных, которые могут быть записаны следующим образом:

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

Зависимая переменная Y является целевой переменной, которую необходимо проанализировать, классифицировать и обобщить. Вектор **x** состоит из входных переменных **x**<sub>1</sub>, **x**<sub>2</sub>, **x**<sub>3</sub> и т. д., которые используются для выполнения этой задачи.



## Содержание

- Основные определения
- Типология деревьев
- Алгоритмы построения дерева
- Достоинства метода
- Недостатки метода

## Регулирование глубины дерева

[Методы регулирования](#)

[Пример задачи](#)

[См. также](#)

[Примечания](#)

[Ссылки](#)

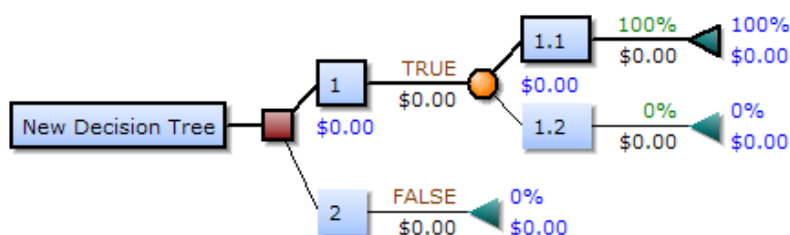
[Литература](#)

## Основные определения

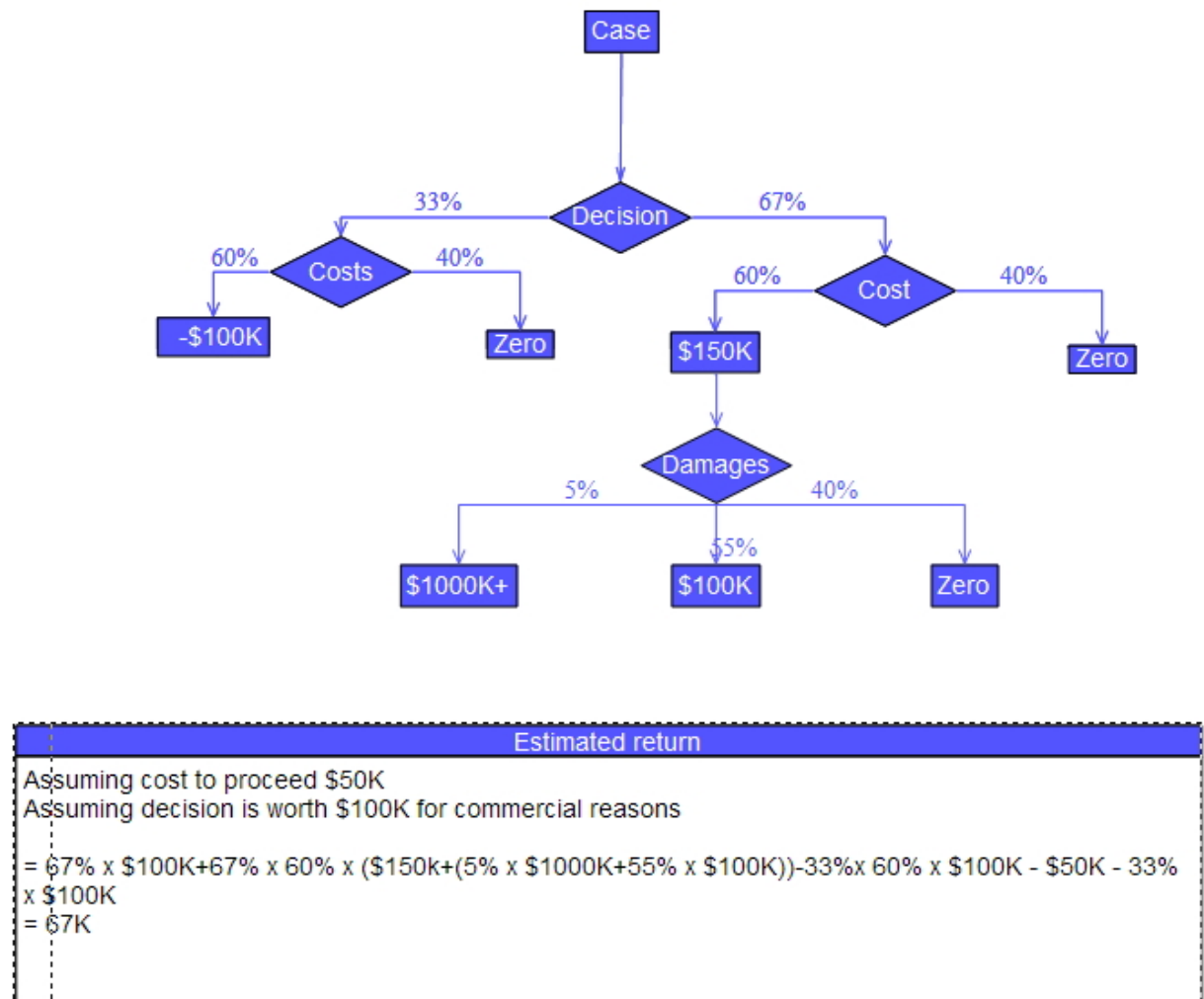
При анализе решений посредством «дерева решений» используют визуальный и аналитический инструмент поддержки принятия решений для расчёта ожидаемых значений (или ожидаемой пользы) конкурирующих альтернатив.

Дерево решений состоит из трёх типов узлов:

- Узлы решения — обычно представлены квадратами
- Вероятностные узлы — представляются в виде круга
- Замыкающие узлы — представляются в виде треугольника



На рисунке, представленном выше, дерево решений следует читать слева направо. Дерево решений не может содержать в себе циклические элементы, то есть каждый новый лист впоследствии может лишь расщепляться, отсутствуют сходящиеся пути. Таким образом, при конструировании дерева вручную, мы можем столкнуться с проблемой его размерности, поэтому, как правило, дерево решений мы можем получить с помощью специализированного программного обеспечения. Обычно дерево решений представляют в виде схематического чертежа, благодаря которому его проще воспринимать и анализировать.



## Типология деревьев

Деревья решений, используемые при добыче данных, бывают двух основных типов:

- Дерево для классификации, когда предсказываемый результат является классом, к которому принадлежат данные;
- Дерево для регрессии, когда предсказываемый результат можно рассматривать как вещественное число (например, цена на дом, или продолжительность пребывания пациента в больнице).

Упомянутые выше термины впервые были введены Брейманом и др.<sup>[2]</sup> Перечисленные типы имеют некоторые сходства (рекурсивный алгоритмы построения), а также некоторые различия, такие, как критерии выбора разбиения в каждом узле.<sup>[2]</sup>

Некоторые методы позволяют построить более одного дерева решений (ансамбли деревьев решений):

1. Бэггинг над деревьями решений, наиболее ранний подход. Строит несколько деревьев решений, неоднократно интерполируя данные с заменой (бутстреп), и в качестве консенсусного ответа выдаёт результат голосования деревьев (их средний прогноз).<sup>[3]</sup>
2. Классификатор «Случайный лес» основан на бэггинге, однако в дополнение к нему случайным образом выбирает подмножество признаков в каждом узле, с целью сделать деревья более независимыми;
3. Бустинг над деревьями может быть использован для задач как регрессии, так и классификации.<sup>[4]</sup> Одна из реализаций бустинга над деревьями, алгоритм XGBoost, неоднократно использовался победителями соревнований по анализу данных.
4. «Вращение леса» — деревья, в которых каждое дерево решений анализируют первым применением метода главных компонент (PCA) на случайные подмножества входных функций.<sup>[5]</sup>

# Алгоритмы построения дерева

---

Есть различные способы выбирать очередной признак:

- Алгоритм ID3, где выбор признака происходит на основании прироста информации (англ. *Gain*), либо на основании критерия Джини.
- Алгоритм C4.5 (улучшенная версия ID3), где выбор признака происходит на основании нормализованного прироста информации (англ. *Gain Ratio*).
- Алгоритм CART и его модификации — IndCART, DB-CART.
- Автоматический детектор взаимодействия Хи-квадрат (CHAID). Выполняет многоуровневое разделение при расчёте классификации деревьев;<sup>[6]</sup>
- MARS: расширяет деревья решений для улучшения обработки цифровых данных.

На практике, в результате работы этих алгоритмов часто получаются слишком подробные деревья, которые при их дальнейшем применении дают много ошибок. Это связано с явлением переобучения. Для сокращения деревьев используют отсечение ветвей (англ. *pruning*).

## Достоинства метода

---

В отличие от остальных методов добычи данных, метод дерева принятия решений имеет несколько достоинств:

- Прост в понимании и интерпретации.
- Не требует специальной подготовки данных, как например: нормализации данных, добавления фиктивных переменных, а также удаления пропущенных данных.
- Способен работать как с категориальными, так и с интервальными переменными. (Прочие методы работают лишь с теми данными, где присутствует лишь один тип переменных. Например, метод отношений может быть применён только на номинальных переменных, а метод нейронных сетей только на переменных, измеренных по интервальной шкале.)
- Использует модель «белого ящика», то есть если определённая ситуация наблюдается в модели, то её можно объяснить при помощи булевой логики. Примером «чёрного ящика» может быть искусственная нейронная сеть, так как полученные результаты сложно объяснить.
- Позволяет оценить модель при помощи статистических тестов. Это даёт возможность оценить надёжность модели.
- Метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включённые в модель.
- Позволяет работать с большим объёмом информации без специальных подготовительных процедур. Данный метод не требует специального оборудования для работы с большими базами данных.

## Недостатки метода

---

- Проблема получения оптимального дерева решений является NP-полной задачей, с точки зрения некоторых аспектов оптимальности даже для простых задач<sup>[7][8]</sup>. Таким образом, практическое применение алгоритма деревьев решений основано на эвристических алгоритмах, таких как алгоритм «жадности», где единственно оптимальное решение выбирается локально в каждом узле. Такие алгоритмы не могут обеспечить оптимальность всего дерева в целом.
- В процессе построения дерева решений могут создаваться слишком сложные конструкции, которые недостаточно полно представляют данные. Данную проблему называют переобучением<sup>[9]</sup>. Для того, чтобы её избежать, необходимо использовать метод «регулирования глубины дерева».
- Существуют понятия, которые сложно понять из модели, так как модель описывает их сложным путём. Данное явление может быть вызвано проблемами XOR, чётности или мультиплексарности. В этом случае мы имеем дело с непомерно большими деревьями. Существует несколько подходов решения данной проблемы, например, попытка изменить репрезентацию концепта в модели (составление новых суждений)<sup>[10]</sup>, или использование алгоритмов, которые более полно

описывают и репрезентируют концепт (например, метод статистических отношений, индуктивная логика программирования).

- Для данных, которые включают категориальные переменные с большим набором уровней (закрытий), большой информационный вес присваивается тем признакам, которые имеют большее количество уровней<sup>[11]</sup>.

## Регулирование глубины дерева

Регулирование глубины дерева — это техника, которая позволяет уменьшать размер дерева решений, удаляя участки дерева, которые имеют маленький вес.

Один из вопросов, который возникает в алгоритме дерева решений — это оптимальный размер конечного дерева. Так, небольшое дерево может не охватить ту или иную важную информацию о выборочном пространстве. Тем не менее, трудно сказать, когда алгоритм должен остановиться, потому что невозможно спрогнозировать, добавление какого узла позволит значительно уменьшить ошибку. Эта проблема известна как «эффект горизонта». Тем не менее, общая стратегия ограничения дерева сохраняется, то есть удаление узлов реализуется в случае, если они не дают дополнительной информации<sup>[12]</sup>.

Регулирование глубины дерева должно уменьшить размер обучающей модели дерева без уменьшения точности её прогноза или с помощью перекрестной проверки. Есть много методов регулирования глубины дерева, которые отличаются измерением оптимизации производительности.

### Методы регулирования

Сокращение дерева может осуществляться сверху вниз или снизу вверх. Сверху вниз — обрезка начинается с корня, снизу вверх — сокращается число листьев дерева. Один из простейших методов регулирования — уменьшение ошибки ограничения дерева. Начиная с листьев, каждый узел заменяется на самый популярный класс. Если изменение не влияет на точность предсказания, то оно сохраняется.

## Пример задачи

Предположим, что нас интересует, выиграет ли наша любимая футбольная команда следующий матч. Мы знаем, что это зависит от ряда параметров; перечислять их все — задача безнадёжная, поэтому ограничимся основными:

- выше ли находится соперник по турнирной таблице;
- дома ли играется матч;
- пропускает ли матч кто-либо из лидеров команды;
- идёт ли дождь.

У нас есть некоторая статистика на этот счёт:

Соперник	Играем	Лидеры	Дождь	Победа
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Ниже	В гостях	На месте	Нет	Да

Хочется понять, выиграет ли наша команда в очередной игре.



## См. также

---

- Random forest — классификатор, основанный на применении комитетов из решающих деревьев
- Переобучение
- Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач
- Таблица принятия решений

## Примечания

---

1. *Quinlan, J. R.* Induction of Decision Trees (<https://www.hunch.net/~coms-4771/quinlan.pdf>)  (англ.) // Machine Learning. — Kluwer Academic Publishers, 1986. — No. 1. — P. 81—106.
2. *Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J.* Classification and regression trees (англ.). — Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. — ISBN 978-0-412-04841-8.
3. *Breiman, L.* Bagging Predictors (англ.) // Machine Learning. — 1996. — No. 24. — P. 123—140.
4. *Friedman, J. H.* Stochastic gradient boosting (англ.). — Stanford University, 1999.
5. *Hastie, T., Tibshirani, R., Friedman, J. H.* The elements of statistical learning : Data mining, inference, and prediction (англ.). — New York: Springer Verlag, 2001.
6. *Kass, G. V.* An exploratory technique for investigating large quantities of categorical data (<https://www.jstor.org/stable/2986296>) (англ.) // Journal of the Royal Statistical Society. Series C (Applied Statistics). — Vol. 29, no. 2. — P. 119—127. — doi:10.2307/2986296 (<https://dx.doi.org/10.2307%2F2986296>).
7. *Hyafil, Laurent; Rivest, RL.* Constructing Optimal Binary Decision Trees is NP-complete (англ.) // Information Processing Letters. — 1976. — Vol. 5, no. 1. — P. 15—17. — doi:10.1016/0020-0190(76)90095-8 (<https://dx.doi.org/10.1016%2F0020-0190%2876%2990095-8>).
8. *Murthy S.* Automatic construction of decision trees from data: A multidisciplinary survey (<https://www2.cs.sfu.ca/CourseCentral/741/jpei/readings/murt98.pdf>)  (англ.) // Data Mining and Knowledge Discovery. — 1998. — No. 2. — P. 345—389.
9. *Max Bramer.* Principles of Data Mining (англ.). — Springer, 2007. — ISBN 978-1-84628-765-7.
10. Inductive Logic Programming (англ.) / Horváth, Tamás; Yamamoto, Akihiro, eds.. — Springer, 2003. — ISBN 978-3-540-20144-1.
11. *Deng, H., Runger, G., Tuv, E.* Bias of Importance Measures for Multi-valued Attributes and Solutions // Artificial Neural Networks and Machine Learning – ICANN 2011. ICANN 2011. Lecture Notes in Computer Science, vol 6792 (англ.) / Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds). — Berlin, Heidelberg: Springer, 2011. — ISBN 978-3-642-21737-1.
12. Fast, Bottom-Up Decision Tree Pruning Algorithm

## Ссылки

---

- Конспект лекции по деревьям принятия решений (<http://logic.pdmi.ras.ru/~sergey/teaching/ml/notes-01-dectrees.pdf>) 
- Decision trees applet provides several sample data sets of examples to learn and classify (<https://web.archive.org/web/20070502112247/http://www.cs.ubc.ca/nest/lci/CIspace/Version4/dTree/>)

## Литература

---

- *Левитин А. В.* Глава 10. Ограничения мощности алгоритмов: Деревья принятия решения // Алгоритмы. Введение в разработку и анализ — М.: Вильямс, 2006. — С. 409—417. — 576 с. — ISBN 978-5-8459-0987-9
- *Паклин Н.Б., Орешков В.И.* Глава 9. // Бизнес-аналитика: от данных к знаниям(+CD): Учебное пособие. 2-е изд.. — СПб.: Питер, 2013. — С. 428—472. — ISBN 978-5-459-00717-6.

---

**Эта страница в последний раз была отредактирована 2 апреля 2022 в 04:48.**

Текст доступен по лицензии Creative Commons Attribution-ShareAlike; в отдельных случаях могут действовать дополнительные условия.

Wikipedia® — зарегистрированный товарный знак некоммерческой организации Wikimedia Foundation, Inc.