

Машинное обучение

Лекция 03. Решающие деревья

Драль Алексей

<https://www.linkedin.com/in/alexey-dral>

27.02.2018, Москва, ФИВТ МФТИ

План

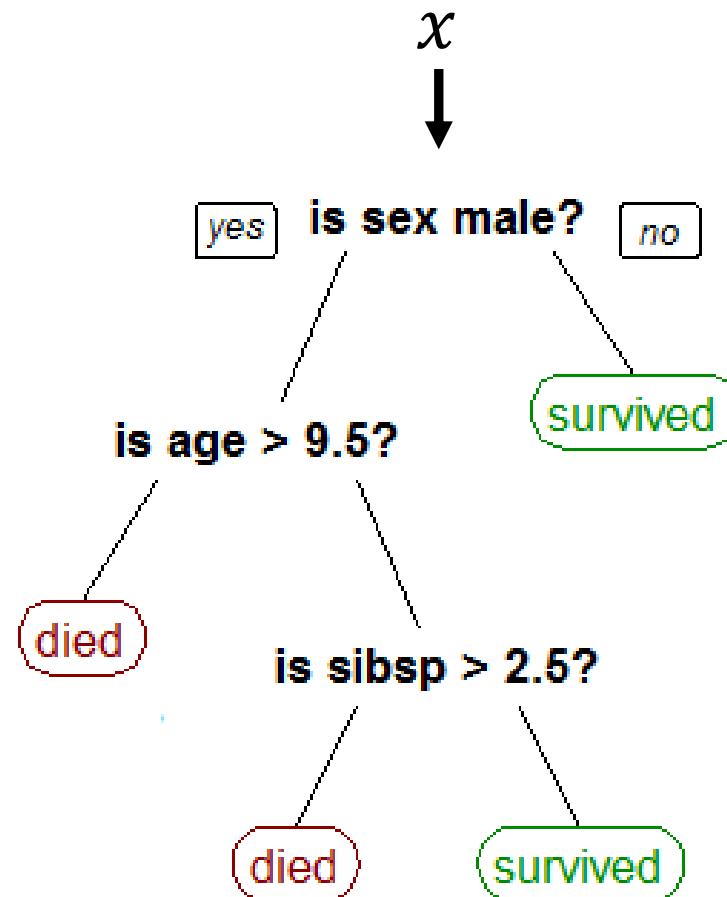
Решающее дерево (Decision Tree)

- Задача классификации и регрессии
- Критерии информативности
- Категориальные признаки
- Пропущенные значения
- Стрижка деревьев (pruning)
- Технические заметки (ID3 / C4.5 / CART)

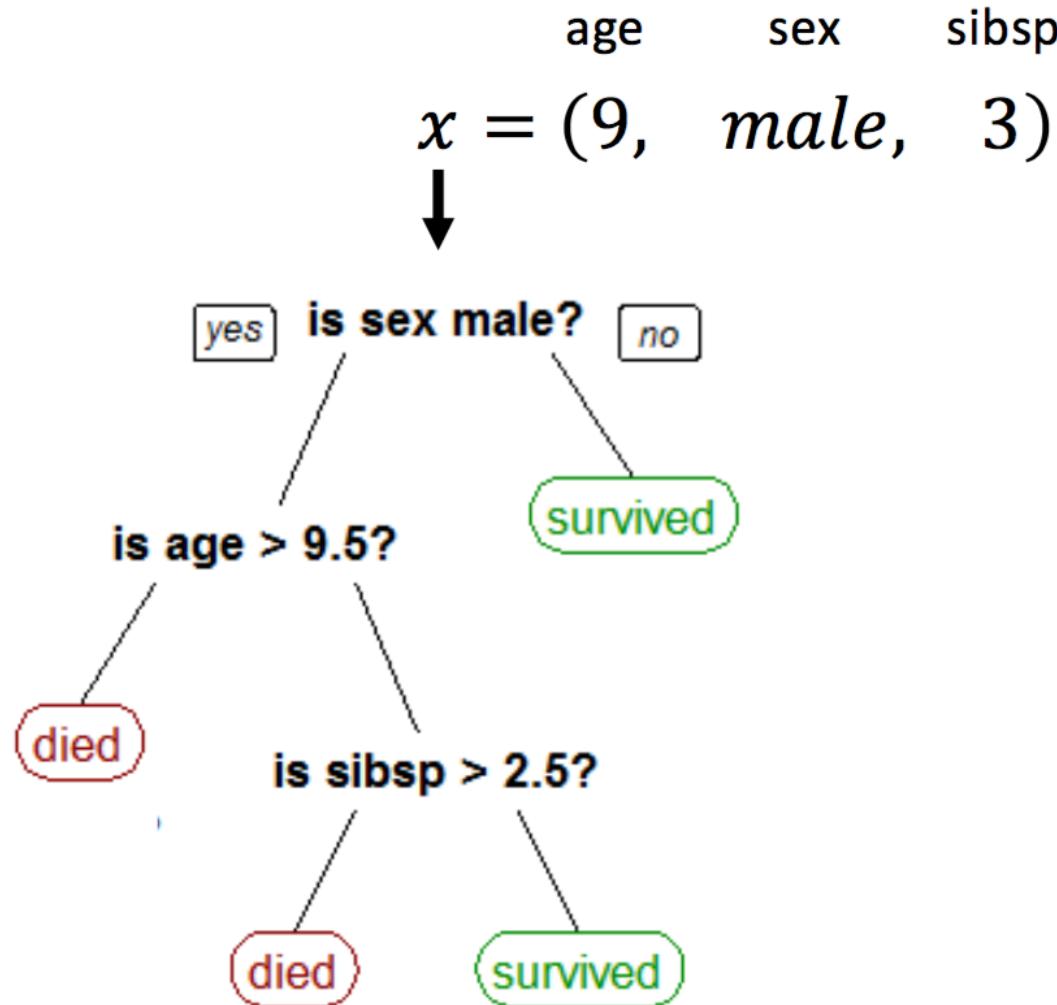
Решающее дерево (Decision Tree)



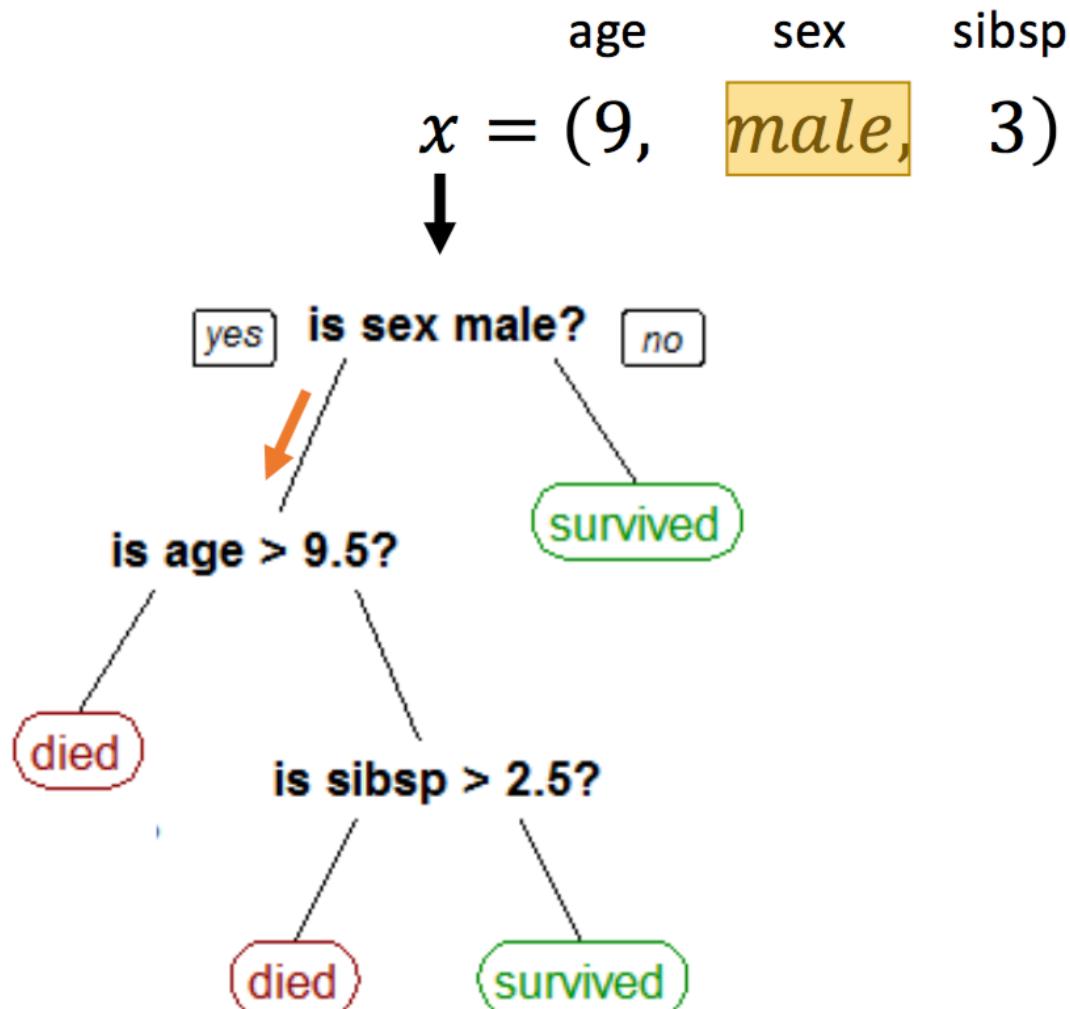
<https://www.kaggle.com/c/titanic>



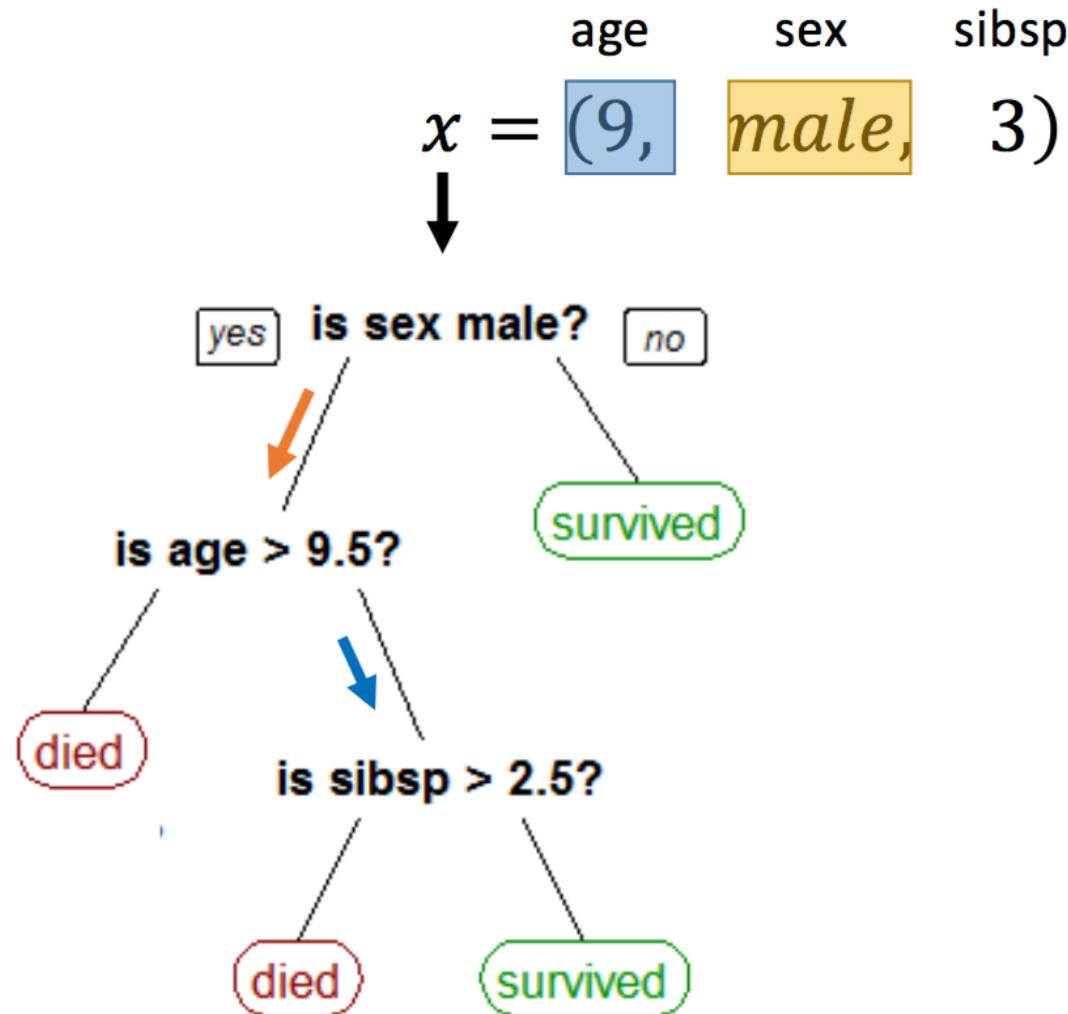
Решающее дерево (Decision Tree)



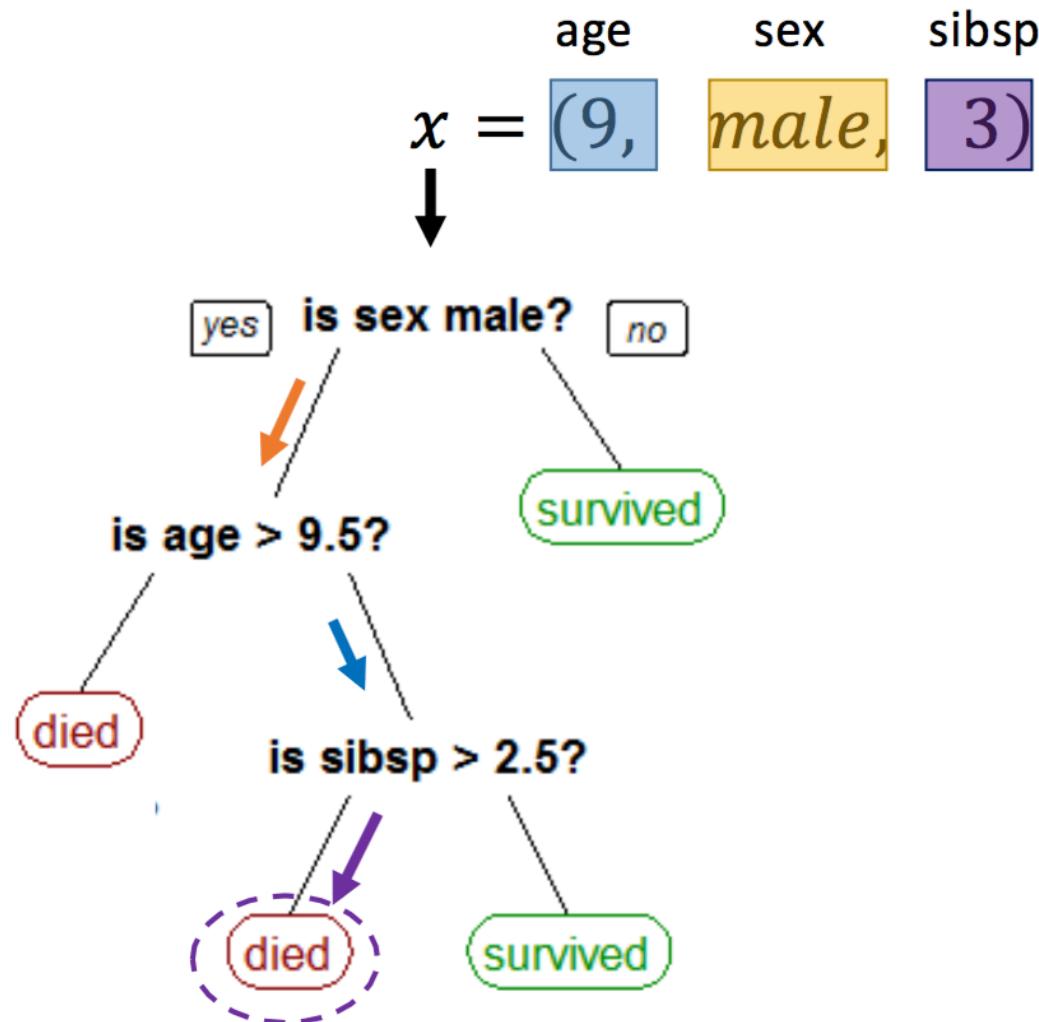
Решающее дерево (Decision Tree)



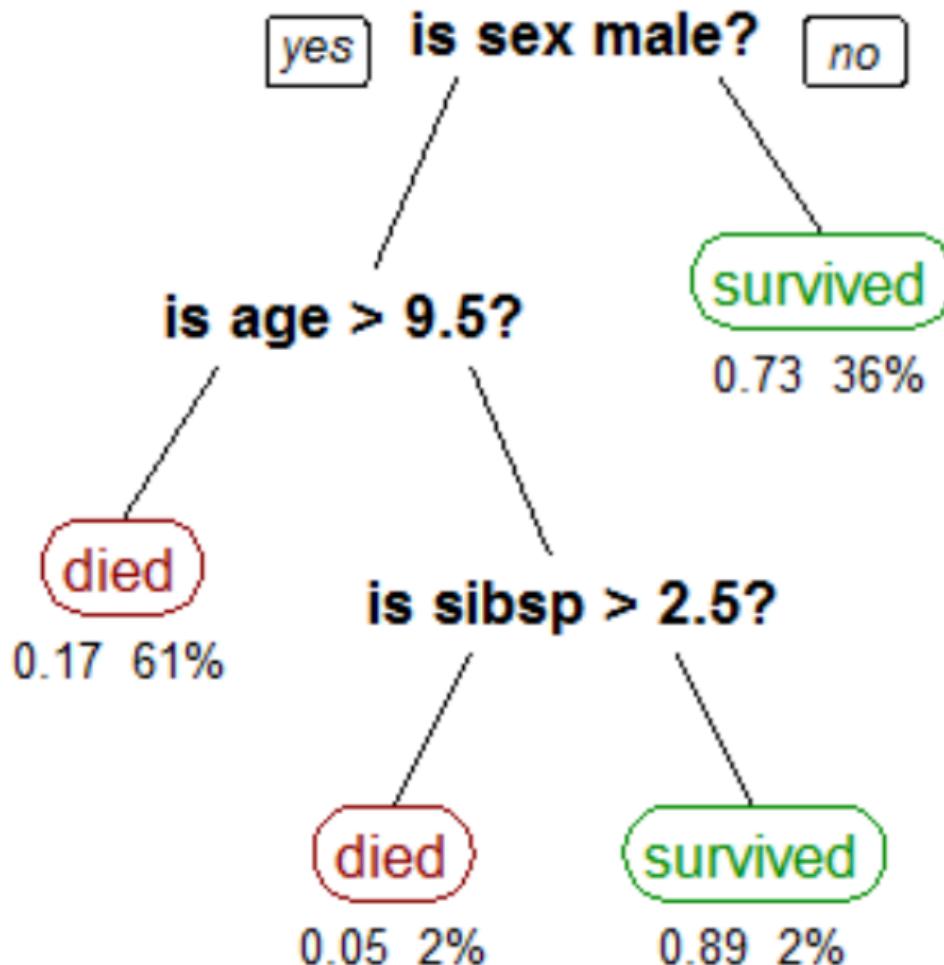
Решающее дерево (Decision Tree)



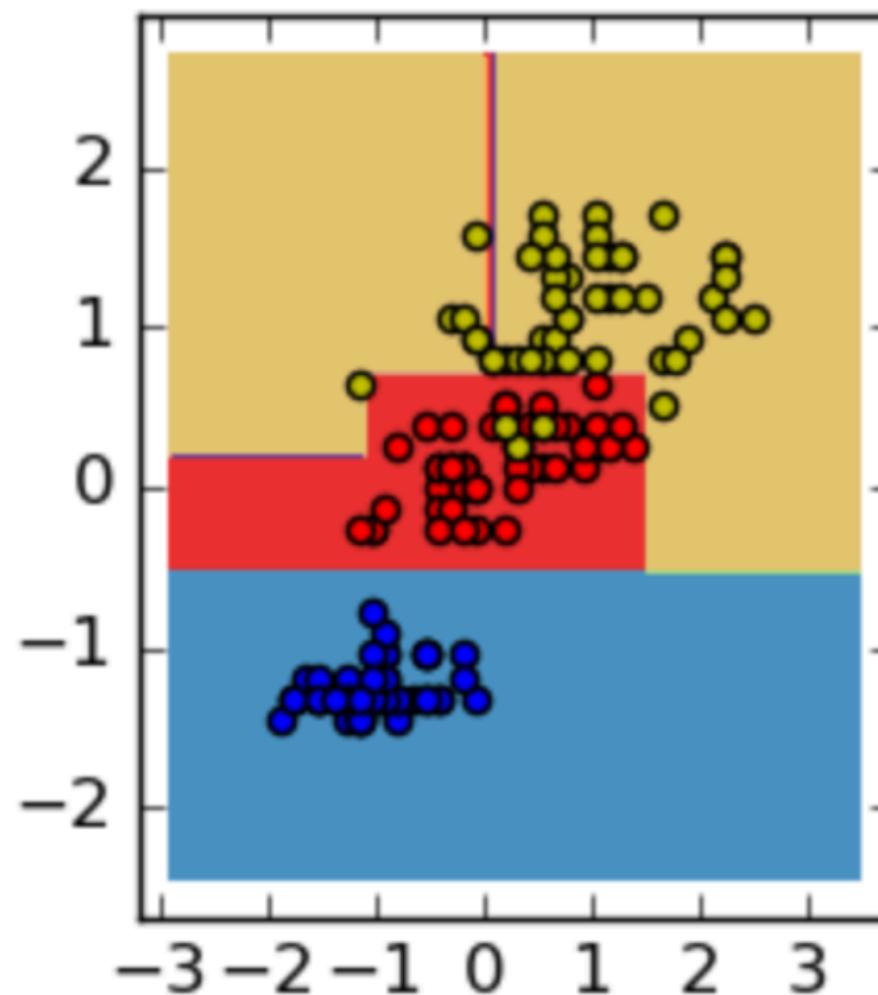
Решающее дерево (Decision Tree)



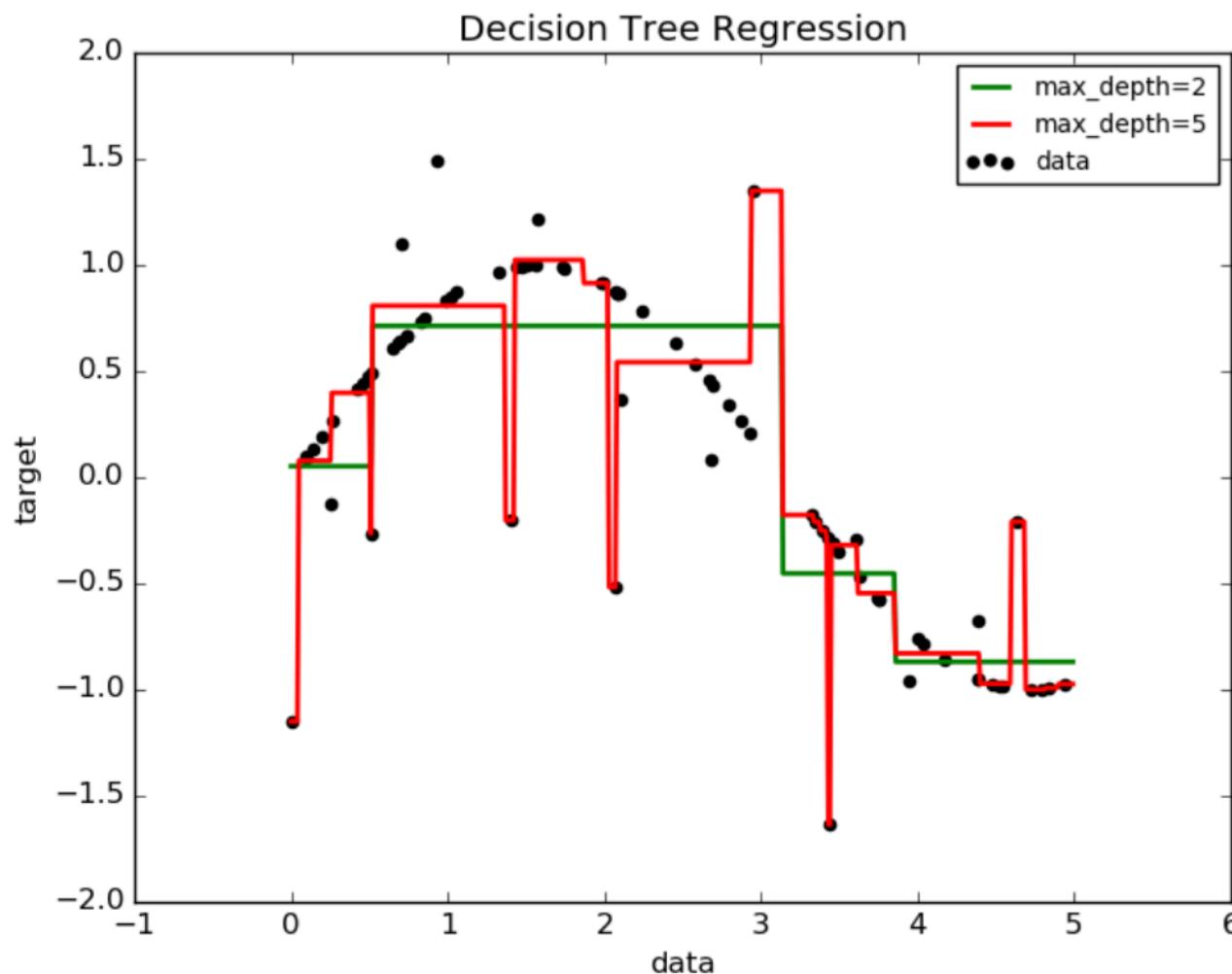
Решающее дерево: классификация



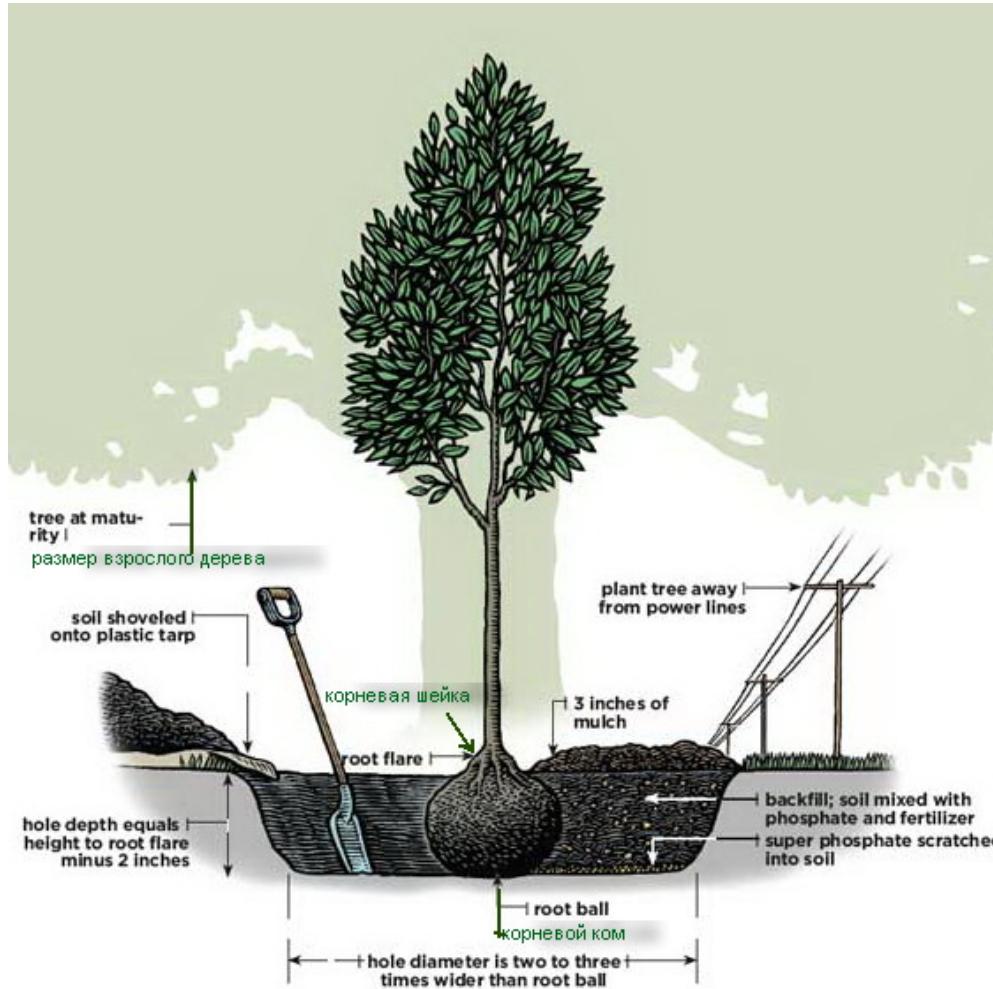
Решающее дерево: классификация



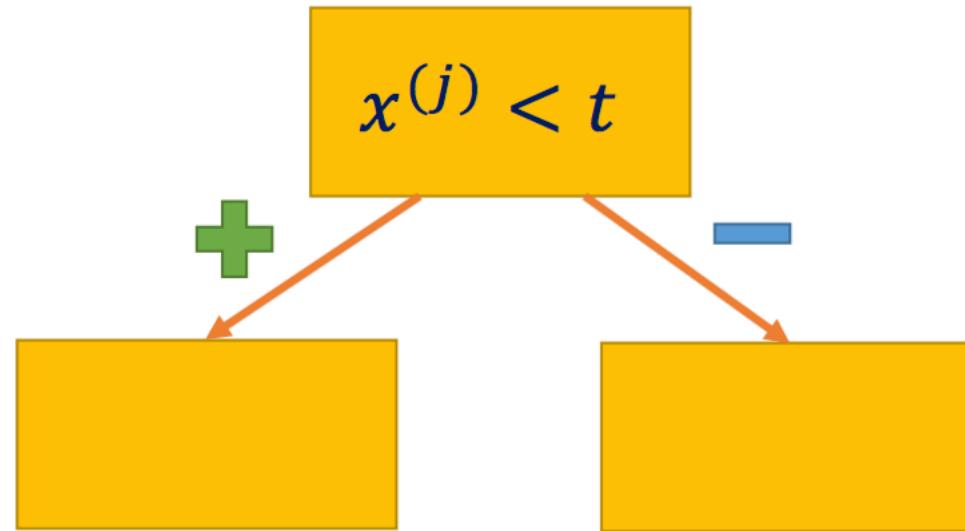
Решающее дерево: регрессия



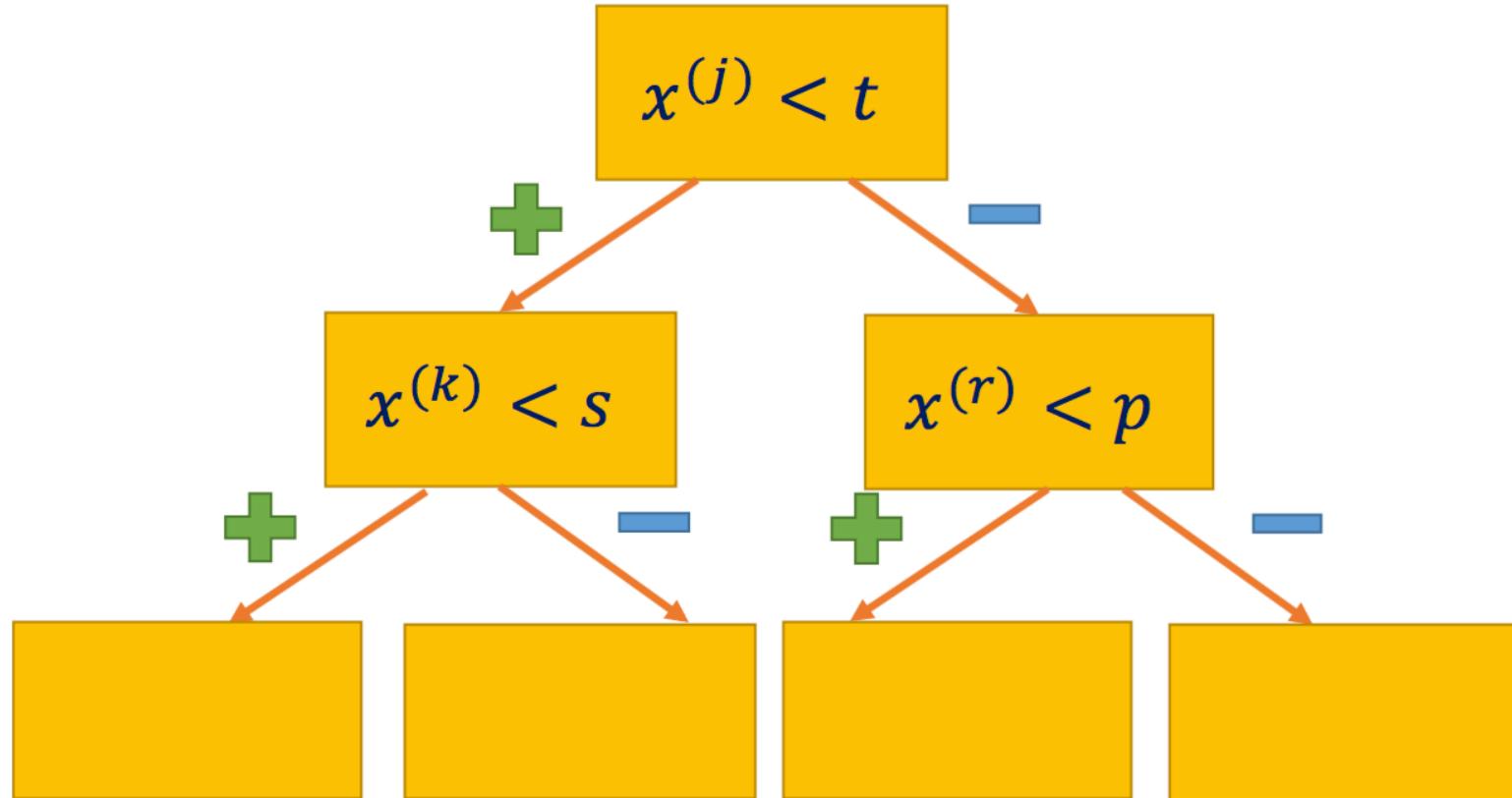
Инструкция по выращиванию деревьев



Инструкция по выращиванию деревьев

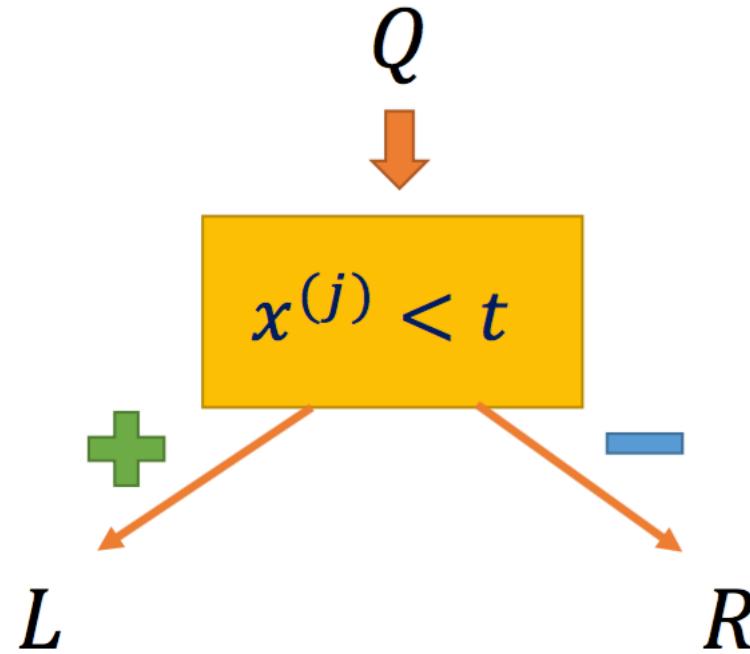


Инструкция по выращиванию деревьев



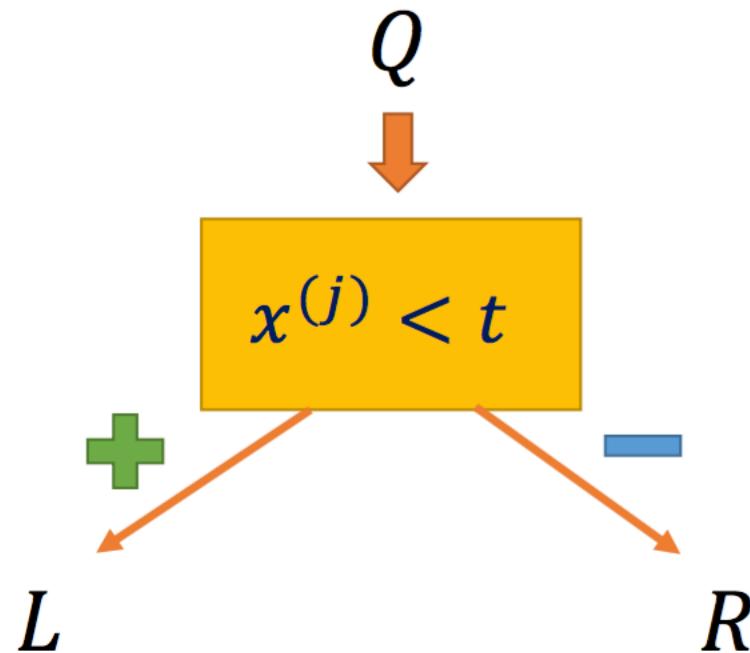
Q&A: выбор порога останова

Инструкция по выращиванию деревьев



Q&A: выбор критерия

Инструкция по выращиванию деревьев



$$G(j, t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \rightarrow \min_{j,t}$$

$$I(Q, j, t) = H(Q) - \frac{|L|}{|Q|} H(L) - \frac{|R|}{|Q|} H(R)$$

Критерии: регрессия

$H(R)$ – мера «неоднородности» выборки R

$$H(R) = \frac{1}{|R|} \sum_{x_i \in R} (y_i - \bar{y})^2$$

Критерии: бинарная классификация

$H(R)$ – мера «неопределенности» / «неоднородности» выборки R

p_0, p_1 – доли объектов классов 0 и 1 в выборке R

1. Misclassification
$$H(R) = 1 - \max\{p_0, p_1\}$$

2. Entropy
$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1$$

3. Gini
$$H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$$

Критерии: многоклассовая классификация

$H(R)$ – мера «неопределенности» / «неоднородности» выборки R

p_1, \dots, p_K – доли объектов классов 1, ..., K в выборке R

1. Misclassification $H(R) = 1 - p_{max}$

2. Entropy
$$H(R) = - \sum_{k=1}^K p_k \ln p_k$$

3. Gini
$$H(R) = \sum_{k=1}^K p_k(1 - p_k)$$

Критерии информативности

Пример:

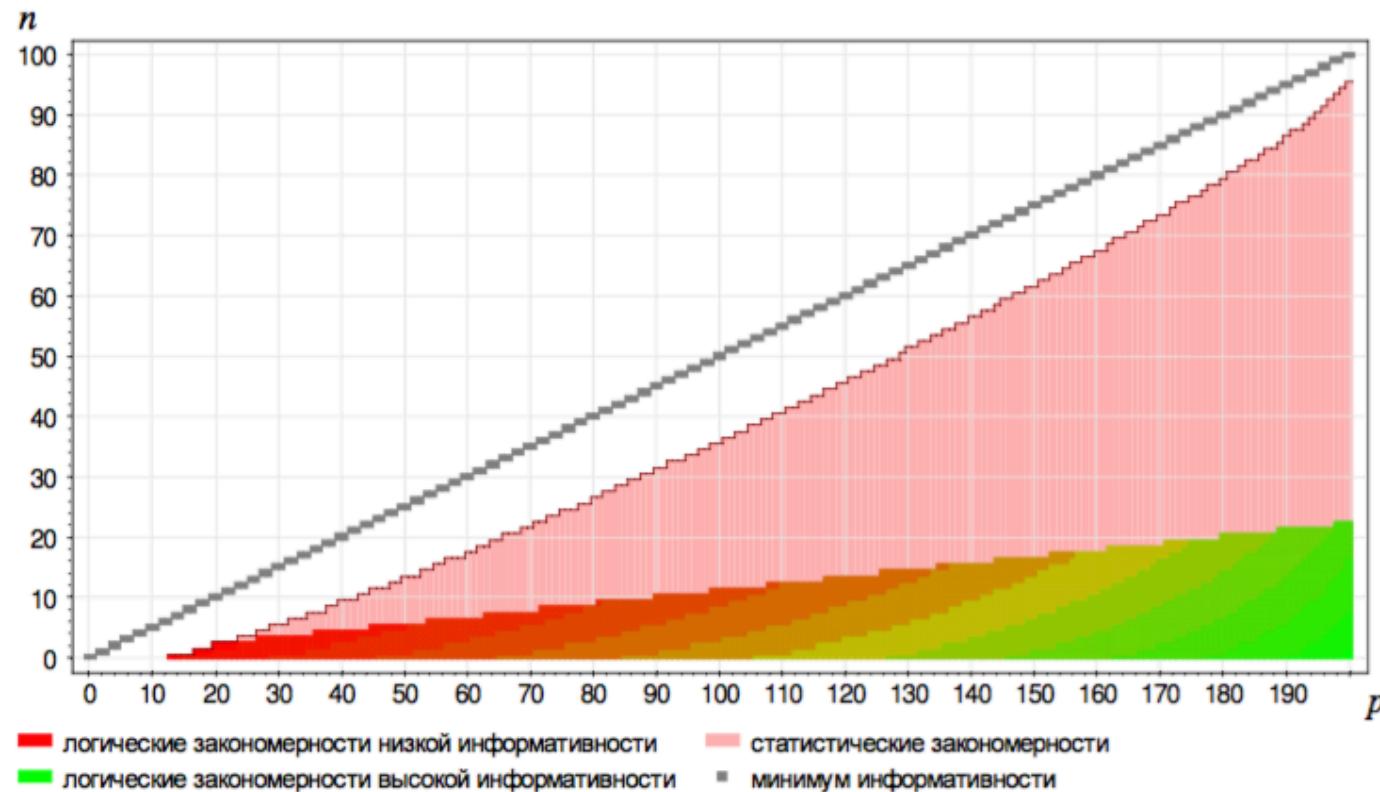
при $P = 200$, $N = 100$ и различных p и n .

Простые эвристики не всегда адекватны:

p	n	$p-n$	$p-5n$	$\frac{p}{P} - \frac{n}{N}$	$\frac{p}{n+1}$	IStat· ℓ	IGain· ℓ	$\sqrt{p} - \sqrt{n}$
50	0	50	50	0.25	50	22.65	23.70	7.07
100	50	50	-150	0	1.96	2.33	1.98	2.93
50	9	41	5	0.16	5	7.87	7.94	4.07
5	0	5	5	0.03	5	2.04	3.04	2.24
100	0	100	100	0.5	100	52.18	53.32	10.0
140	20	120	40	0.5	6.67	37.09	37.03	7.36

Критерии информативности

Логические закономерности: $\frac{n}{p+n} \leq 0.1$, $\frac{p}{P+N} \geq 0.05$.
Статистические закономерности: $IStat(p, n) \geq 3$.



Q&A: проблемы обученного дерева



Категориальные признаки

Допустим категориальный признак x^j имеет q различных значений, тогда нам нужно перебрать $\sim 2^q$ разбиений.

Дороге долго

Категориальные признаки

$R(u)$ – множество объектов, попавших в текущую вершину и имеющих $x^j = u$. $u_{(i)}$ упорядочены по числу объектов, относящихся к классу +1.

$$\frac{1}{|R(u_{(1)})|} \sum_{x_i \in R(u_{(1)})} [y_i = +1] < \dots < \frac{1}{|R(u_{(q)})|} \sum_{x_i \in R(u_{(q)})} [y_i = +1]$$

$u(k)$ кодируем с помощью k и работаем как с вещественным признаком. Результат для критерия Джини и энтропийного критерия такой же, как и для полного перебора.

(1984) Breiman et al.; (1996) Ripley

Категориальные признаки

$$\frac{1}{|R(u_{(1)})|} \sum_{x_i \in R(u_{(1)})} y_i < \dots < \frac{1}{|R(u_{(q)})|} \sum_{x_i \in R(u_{(q)})} y_i$$

Задача регрессии (минимизация MSE)

(1988) Fisher

Пропущенные значения (missing attributes)

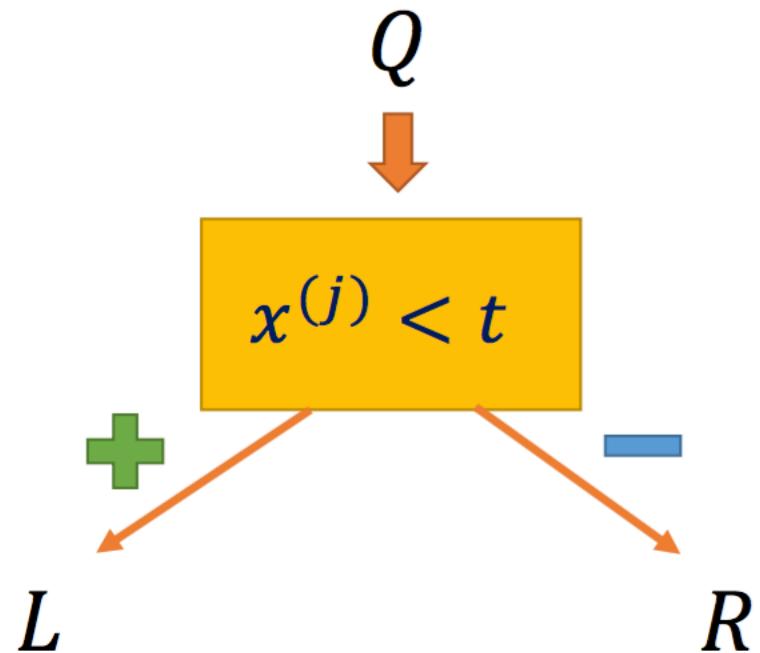
Варианты:

- Фильтровать объекты с пропусками
- Заполнить пропуски “impute” (пример: mean, median)

Специально для деревьев:

- Добавить категорию “missing”
- Суррогатные предикаторы

Пропущенные значения (missing attributes)



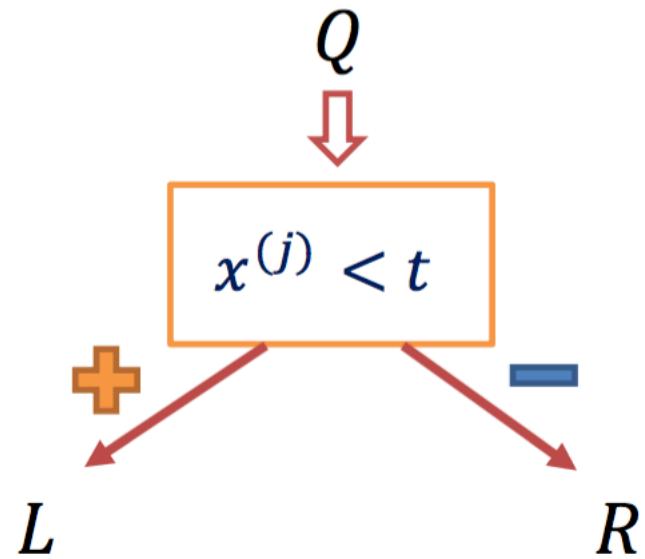
$$G(Q, j, t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R)$$

Пусть $x^{(j)}$ не определен для $V \subset Q$,
подправим $G(Q, j, t)$:

$$G(Q, j, t) = \frac{|Q \setminus V|}{|Q|} G(Q \setminus V, j, t)$$

Если разбиение по $x(j)$ окажется лучшим, то добавим объекты V в левое и в правое поддеревья

Пропущенные значения (missing attributes)



Также можно учитывать объекты из V с весом $\frac{|L|}{|Q|}$ в левом поддереве и $\frac{|R|}{|Q|}$ в правом

При применении также – например, усредняем с этими весами прогноз вероятности класса от левого и правого поддерева

Суррогатные предикаты

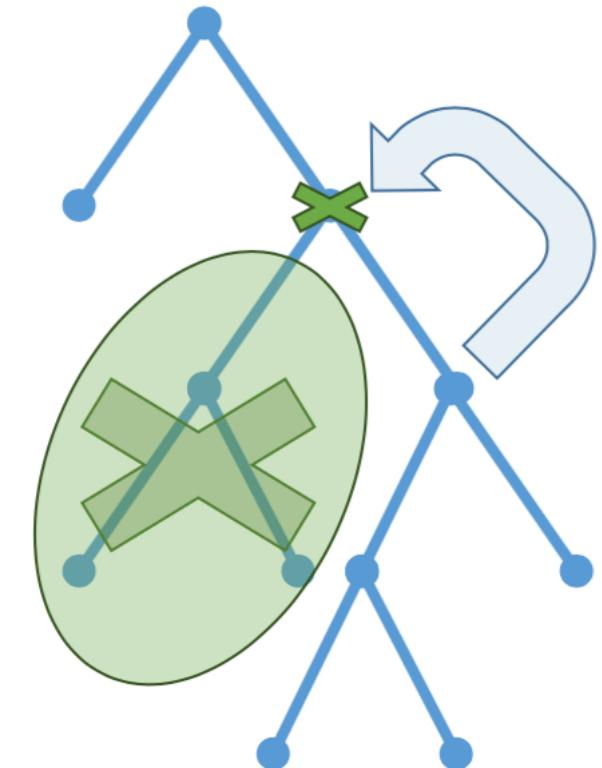
Разбиваем множество V по другому признаку (не пропущенному), разбиение по которому для остальных вершин максимально похоже на наилучшее.

Инструкция по стрижке деревьев (pruning)



Инструкция по стрижке деревьев (pruning)

- Pre-pruning
 - Максимальная глубина дерева
 - Минимальное число элементов в узле дерева
 - Минимальное число элементов в сплите дерева
 - Минимальный “Information gain”
 - ...
- Post-pruning
 - Упрощаем дерево после того, как оно было построено



Cost-complexity pruning

$$C_\alpha(T) = C(T) + \alpha|T| \rightarrow \min, \quad \alpha \geq 0$$

Оказывается, что существует последовательность:

$$T_K \subset T_{K-1} \subset \dots \subset T_0$$

где T_i минимизирует критерий $C_\alpha(T)$ в интервале $\alpha \in [\alpha_i, \alpha_{i+1})$

$$0 = \alpha_0 < \alpha_1 < \dots < \alpha_K < \infty$$

$|T|$ - число листьев дерева Т

T_K – тривиальное дерево, состоящее из корня

ID3: Iterative Dichotomiser 3 (1986)

- Умеет работать только с бинарными признаками
- Использует энтропийный критерий информативности
- Строит бинарное дерево до тех пор, пока уменьшается энтропия при разделении

Вопрос на понимание: в каком случае будет происходить деление узла, в котором находятся исключительно объекты одного класса?

ID3 нового поколения: C4.5 или J48 (1993)

- Умеет работать с вещественными признаками
- Умеет работать с пропущенными значениями (стратегия: пропускать в рамках вычислений энтропии, взвешивать в узлах поддерева)
- Post-pruning

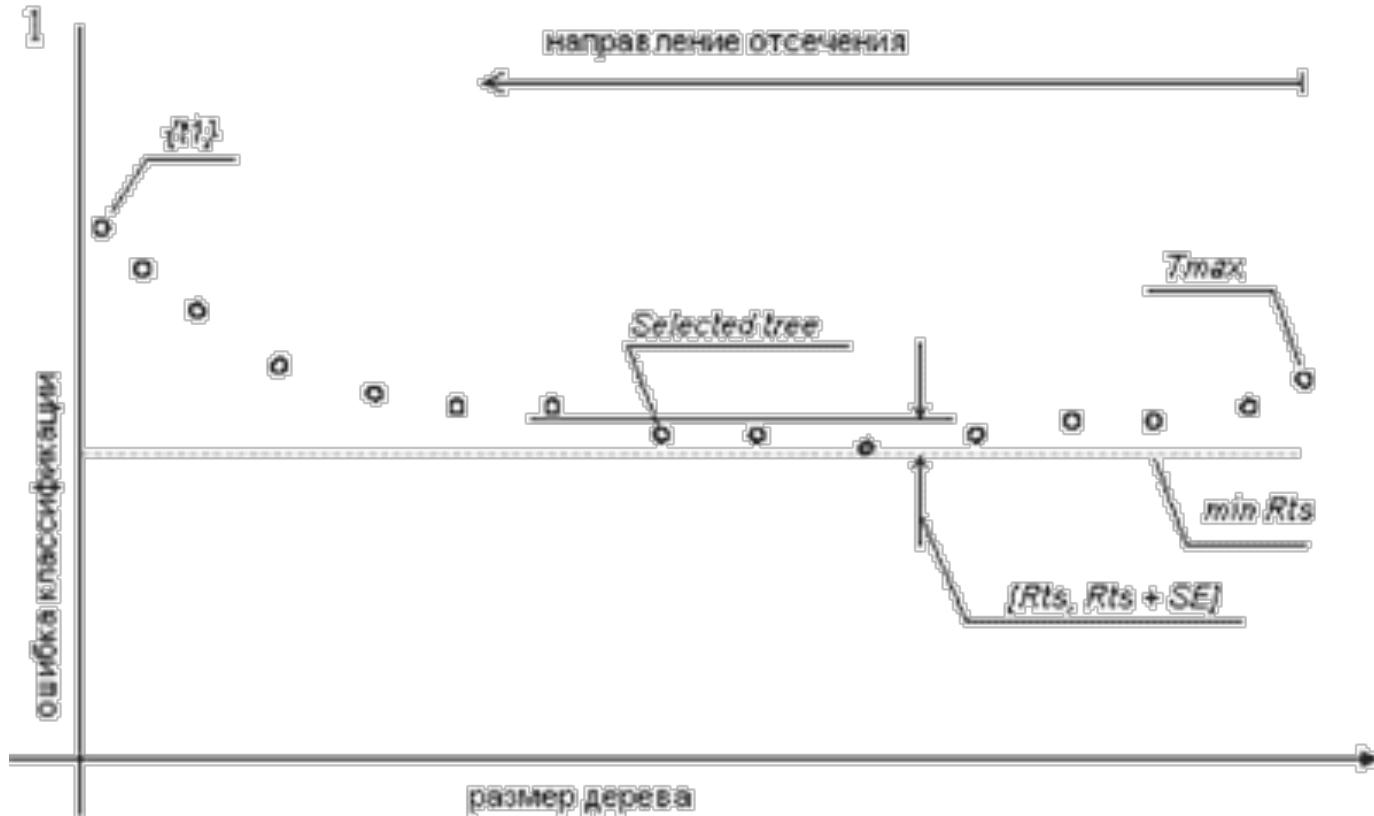
C5.0 – коммерческая оптимизация C4.5

CART: Classification and Regression Trees (1984)

- Использует критерий Gini
- Post-pruning с помощью cost-complexity pruning
- Для обработки пропусков используются суррогатные предикаты
- В дополнение к задачам классификации могут решать задачи регрессии (MSE)

Историческая справка – основано на работе Morgan и Sonquist 1963 года.

CART: 1-SE rule



Вычислительная сложность

Имплементация с учетом сортировки объектов по каждому признаку:

$$O(h m n \log(n))$$

Оптимизация `sklearn.tree`:

$$O(h (m + n) \log(n))$$

h – число узлов дерева

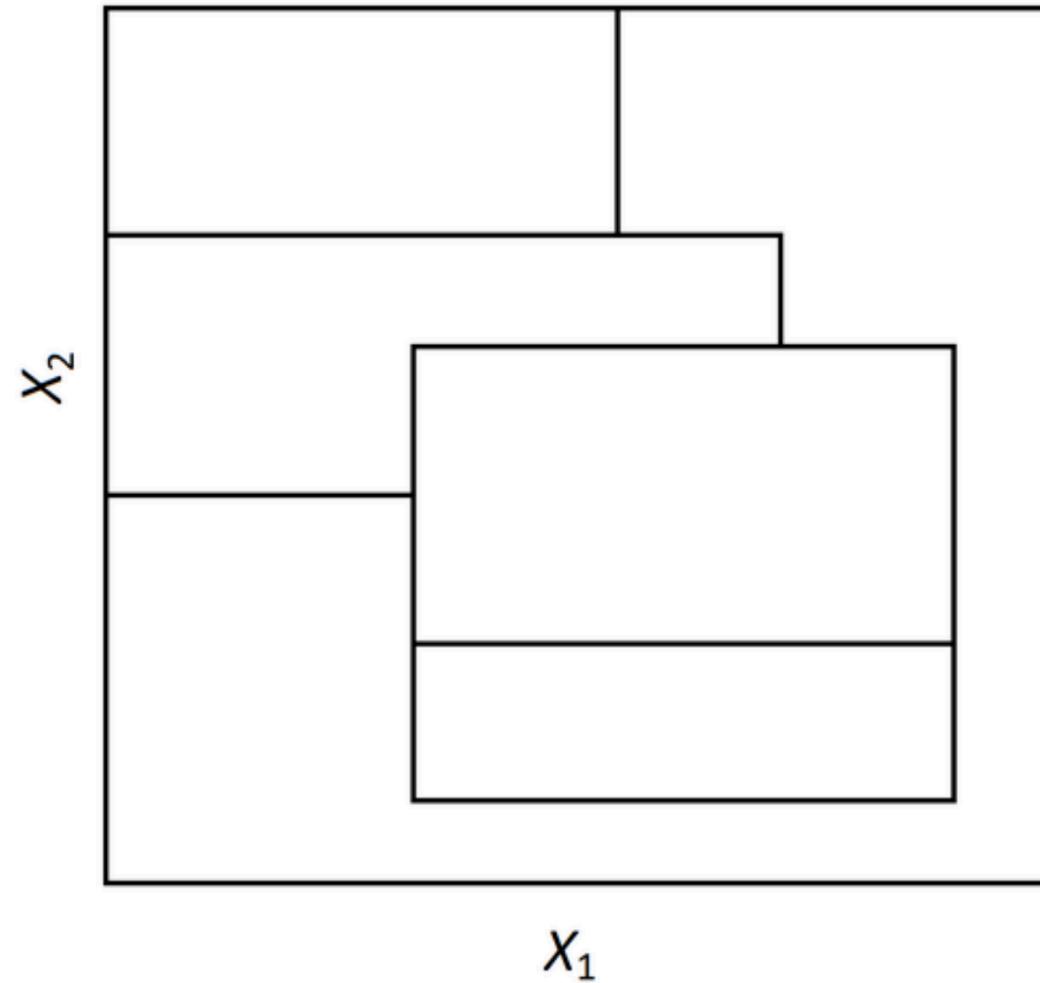
m – число признаков

n – число объектов в обучающей выборке

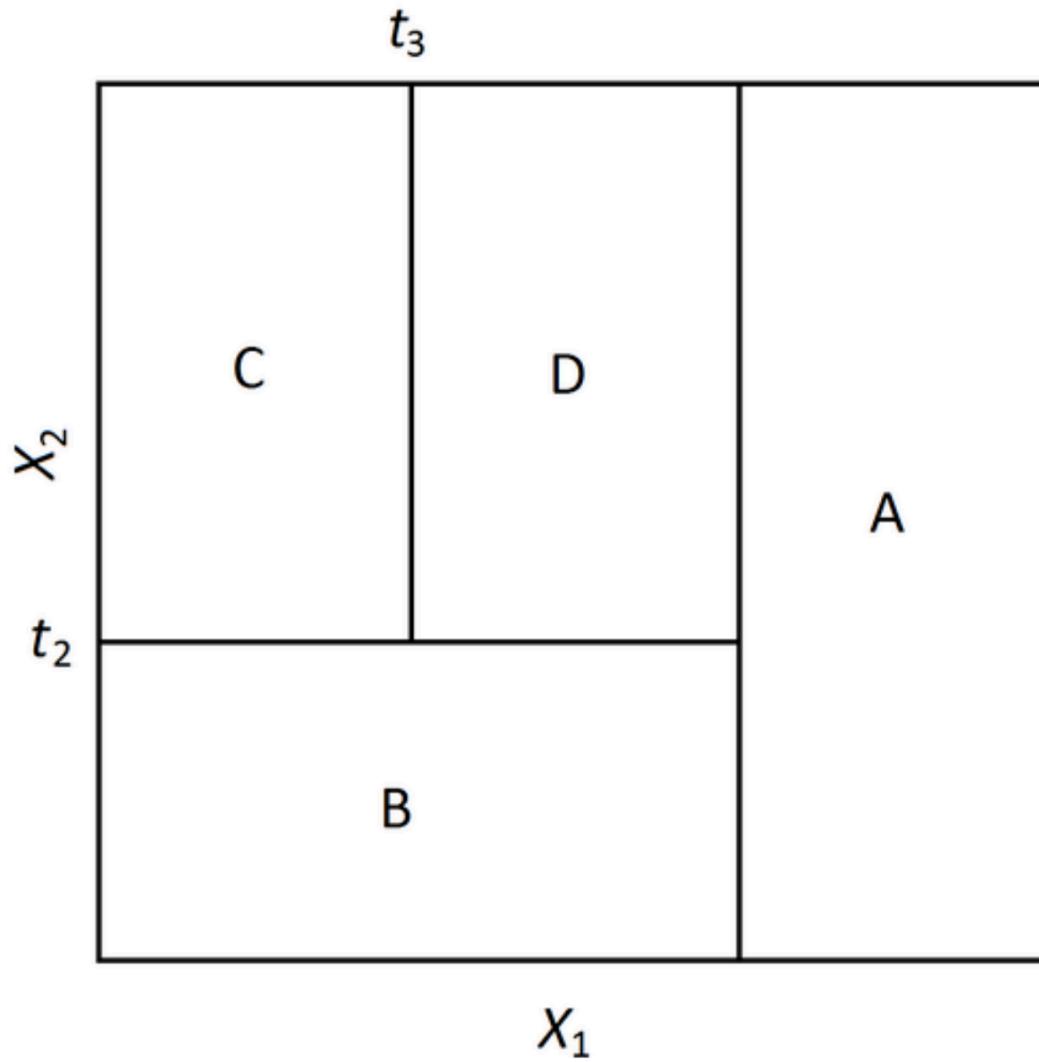
Q&A: как посчитать информативность признаков



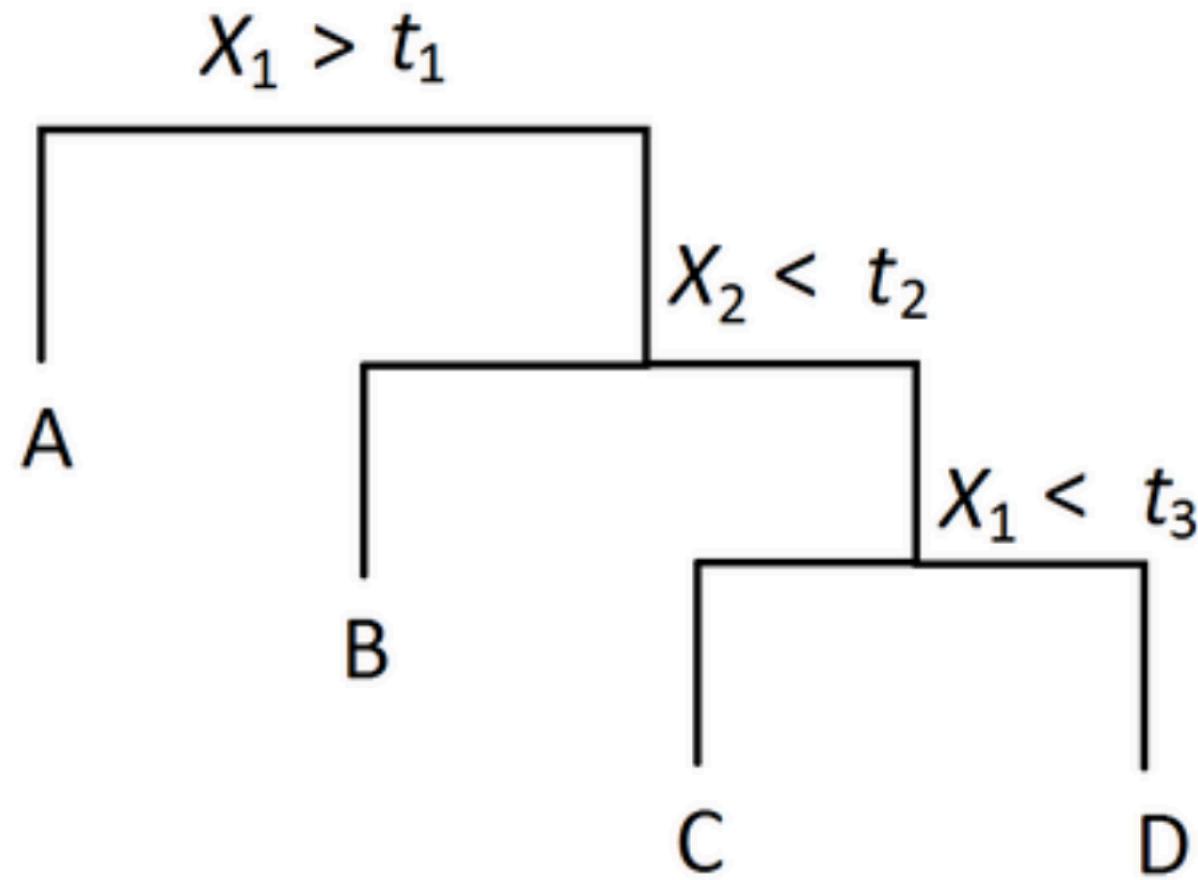
Q&A: «проверка связи»



Q&A: «проверка связи»



Q&A: «проверка связи»



Резюме

Преимущества решающих деревьев:

- интерпретируемость
- работа с разными типами данных
- работа с пропусками
- решение задач классификации и регрессии

Недостатки решающих деревьев:

- переобучение
- неустойчивость к шуму, выборке

Обратная связь

Отзывы о прошедших лекциях и семинарах просьба оставлять здесь:

https://ml-mipt.github.io/2018part1_Schedule/

Полезные материалы

- материалы курса ФИВТ МФТИ (включая архивы прошлых лет) -
<https://github.com/ml-mipt/ml-mipt-part1>
- James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning with Applications in R (Chapter 8. Tree-Based Methods)
- Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. (Chapter 9. Additive Models, Trees, and Related Methods)
- материалы курса ФКН ВШЭ -
http://wiki.cs.hse.ru/Машинное_обучение_1
- Щепин Е.В. Теория информации и распознавание образов

Задача со звездочкой #1: best split

Цель: доказать утверждение про возможность оптимизации перебора экспоненциального числа разбиений категориальных признаков до линейного перебора

Бонус: 1 балл (первым Зм приславшим)

Оформление: электронный документ (не скан / фото)

Тема письма: “ML-2018, tree, best split, ФИО (группа)”

Высыпать по адресу: aadral@gmail.com

Задача со звездочкой #2: complexity

Цель: провести исследование о зависимости времени обучения дерева от числа параметров (число узлов дерева / глубина, число признаков в обучающей выборке, число объектов в обучающей выборке)

Бонус: 1 балл (первым Зм приславшим)

Оформление: электронный документ (не скан / фото) + ссылка на репозиторий для воспроизведения результатов

Тема письма: “ML-2018, tree, complexity, ФИО (группа)”

Высыпать по адресу: aadral@gmail.com