Машинное обучение

Лекция 3

Решающие деревья в задачах классификации и регрессии

Виктор Кантор

На этой лекции

I. Решающие деревья
II. Ансамбли решающих деревьев

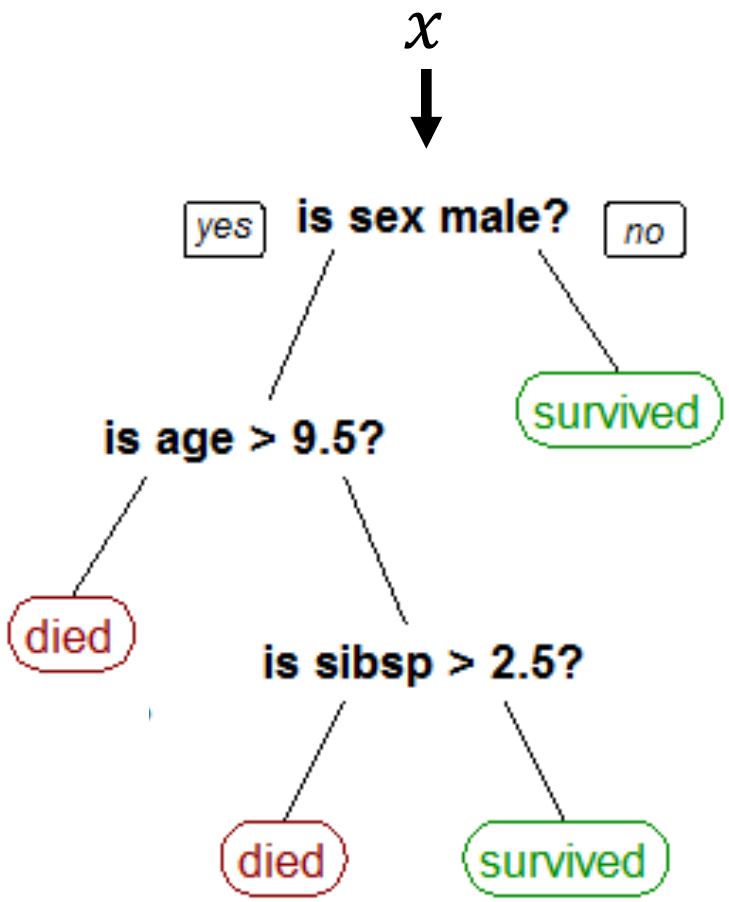
I. Решающие деревья

План

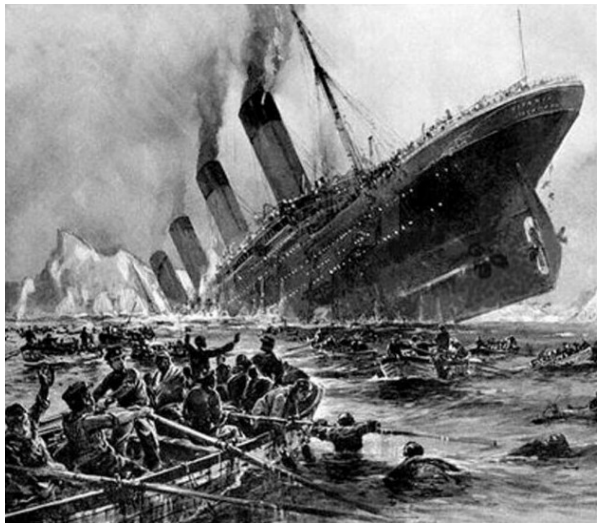
1. Что такое решающие деревья
2. Решающие деревья в классификации и регрессии
3. Как строить решающие деревья
4. Дополнительные темы

1. Что такое решающие деревья

Решающее дерево



Датасет

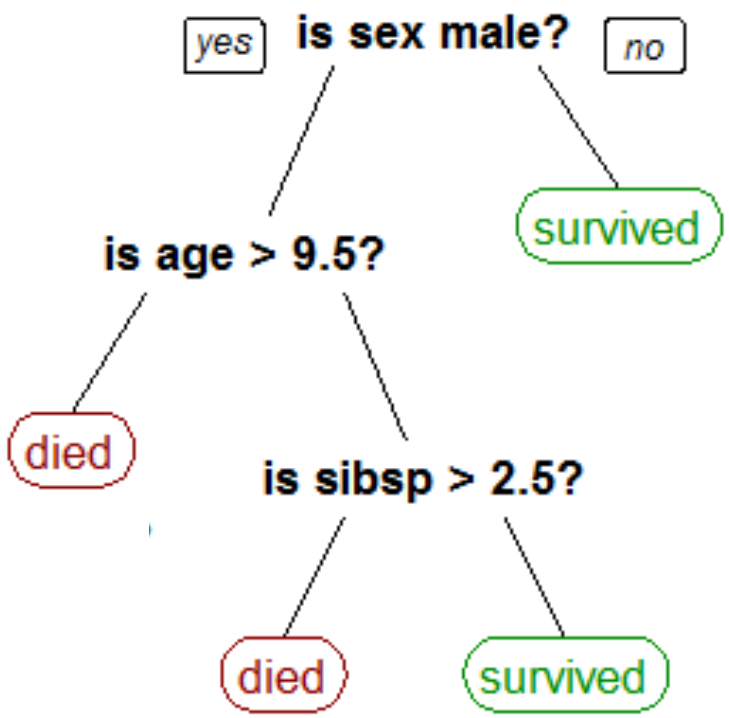


«Titanic Dataset» - список пассажиров Титаника, для которых даны возраст, пол, количество членов семьи на борту и другие признаки.

Целевые значения: выжил пассажир или нет (задача классификации)

Решающее дерево

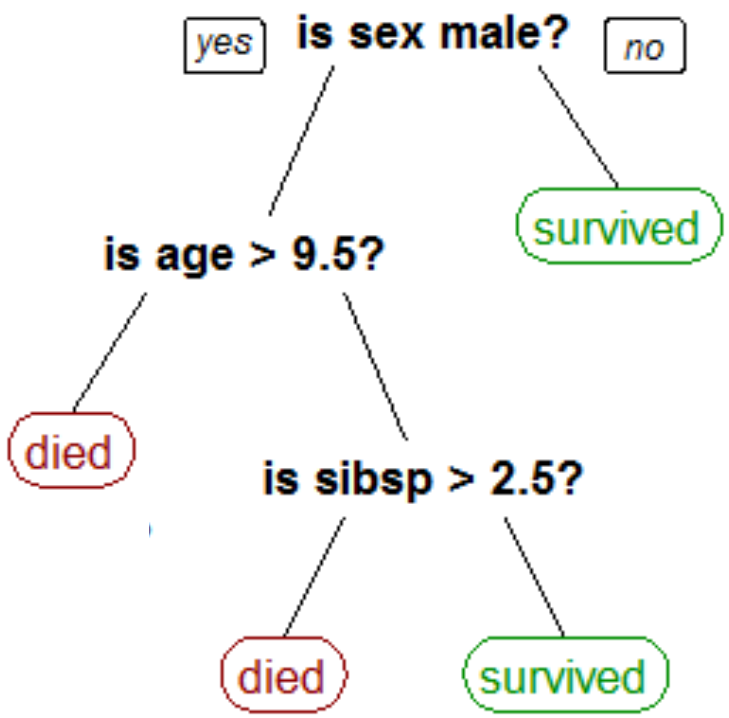
$$x = (9, \quad male, \quad 3)$$



Решающее дерево

age sex sibsp

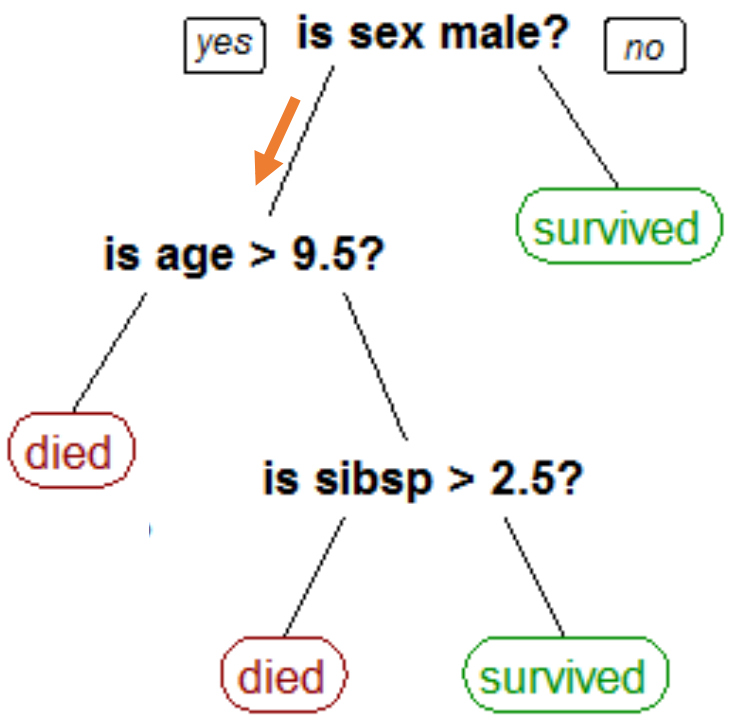
$$x = (9, \quad male, \quad 3)$$



Решающее дерево

age sex sibsp

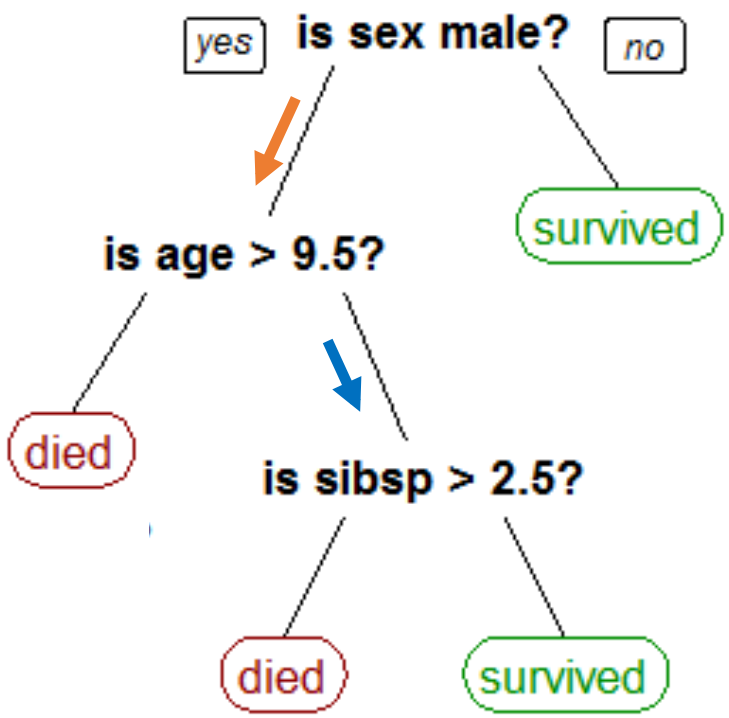
$$x = (9, \quad male, \quad 3)$$



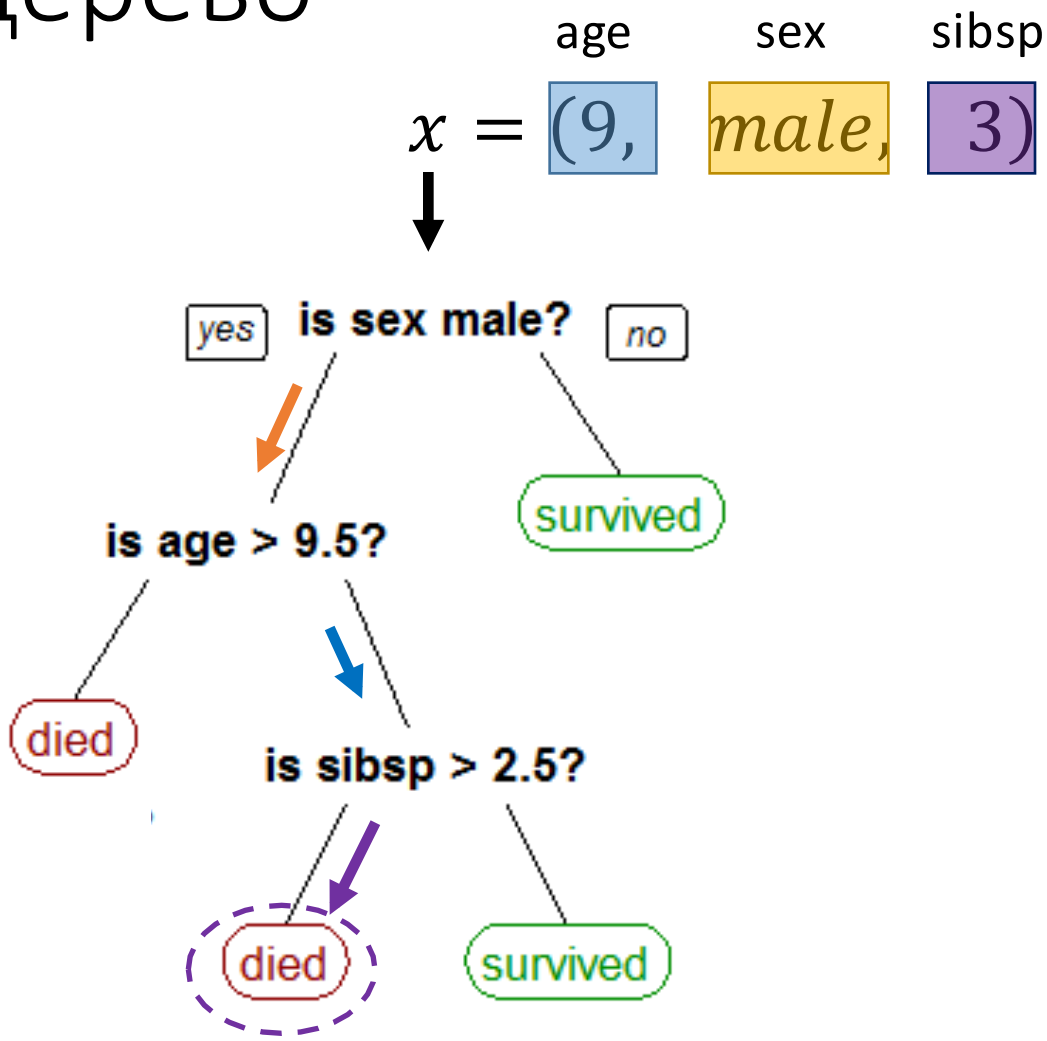
Решающее дерево

age sex sibsp

$x = (9, \ male, \ 3)$



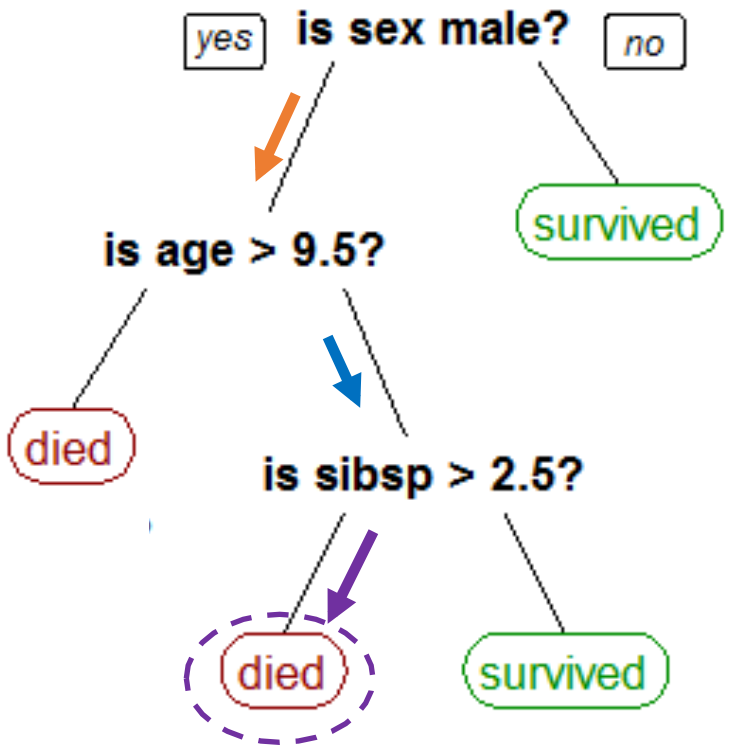
Решающее дерево



Решающее дерево

age sex sibsp

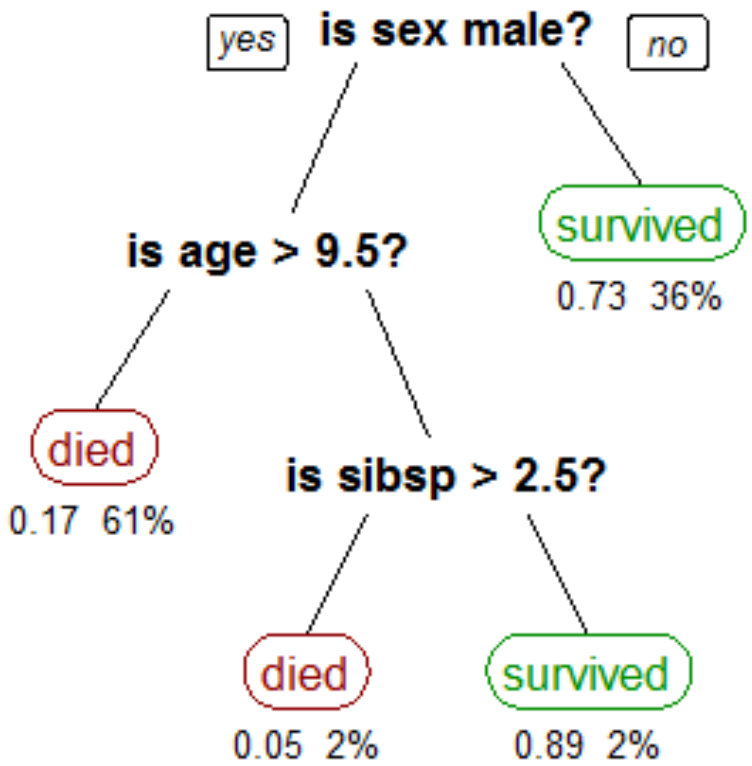
$x = (9, \ male, \ 3)$



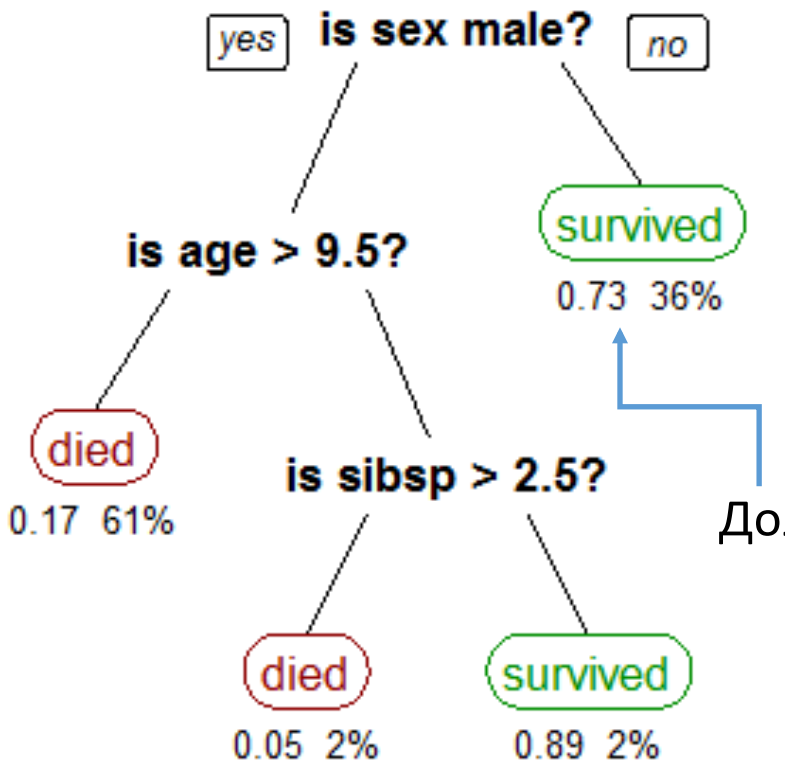
$y = died$

2. Решающие деревья в классификации и регрессии

Решающее дерево: классификация



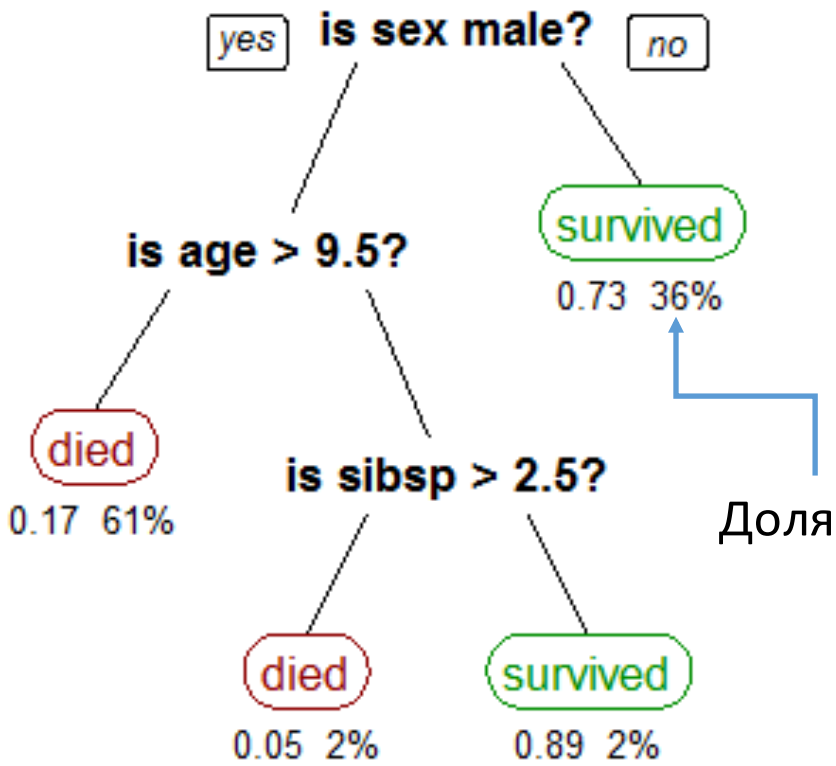
Решающее дерево: классификация



Доля выживших среди попавших в лист

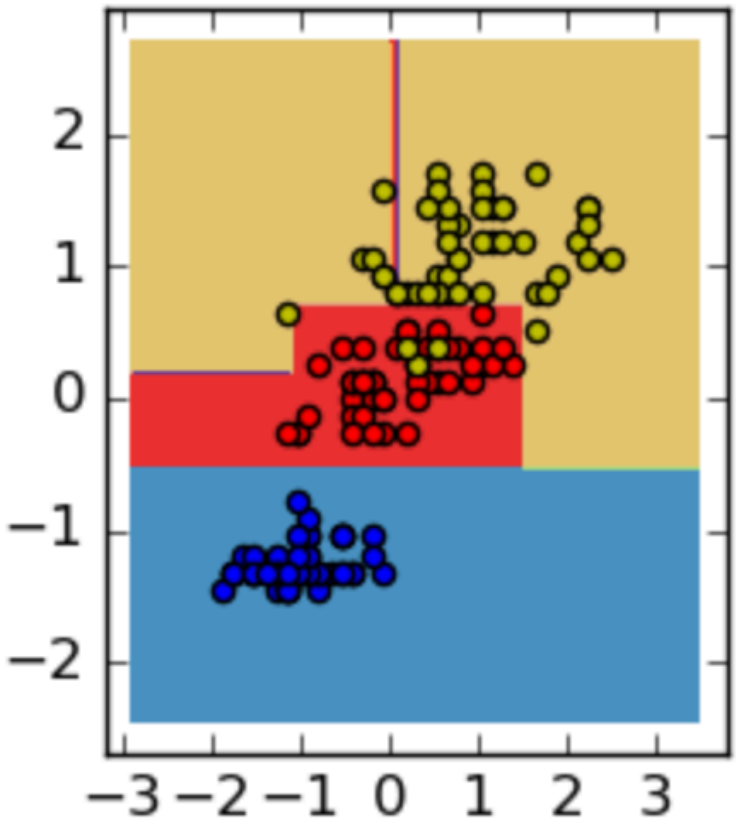
В каждом листе дерево отвечает преобладающим классом

Решающее дерево: классификация



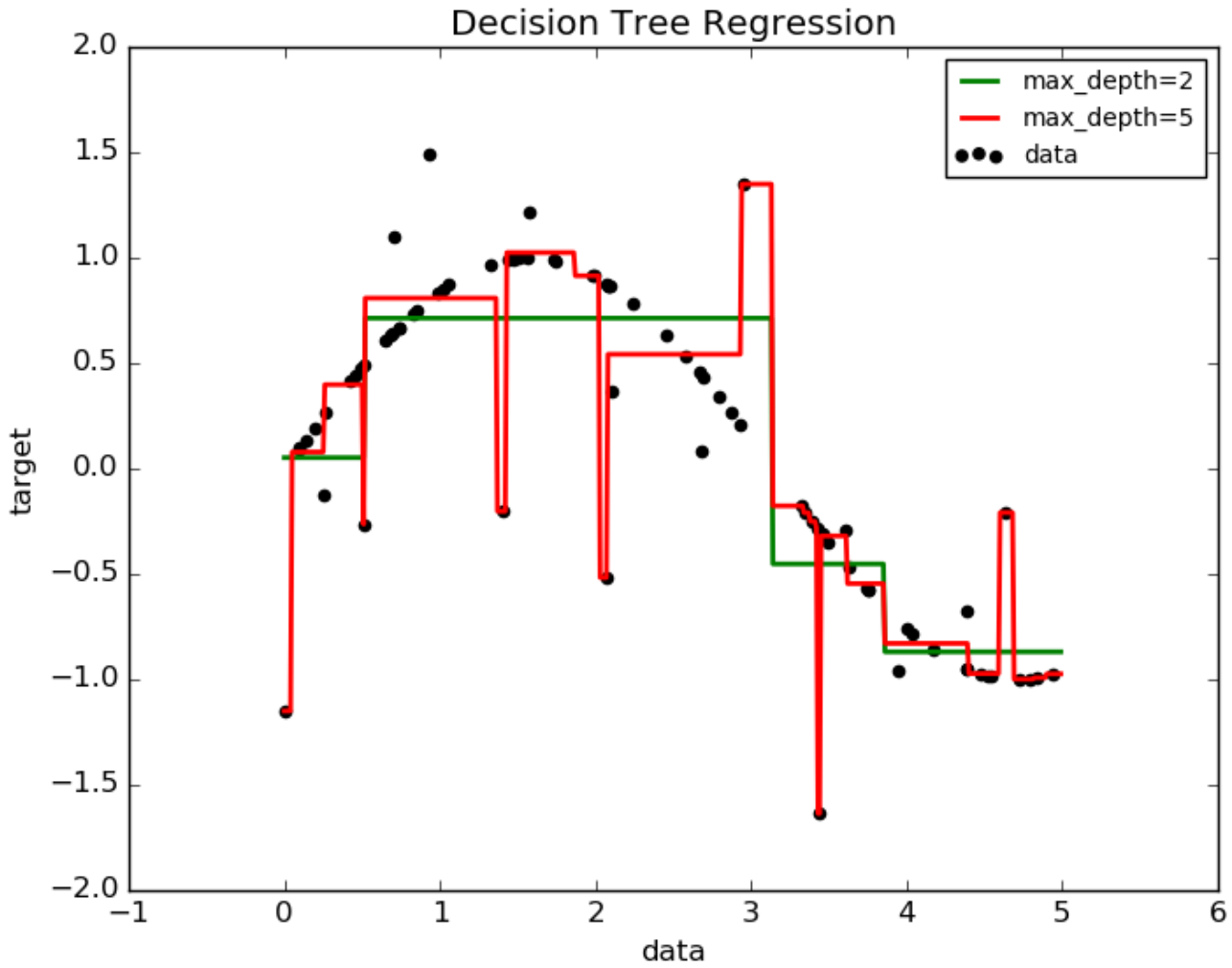
Доля объектов обучающей выборки

Решающее дерево: классификация



Пример: 3 класса и 2 признака

Решающее дерево: регрессия



Пример: восстановление зависимости y от x с помощью решающих деревьев глубины 2 и глубины 5

В каждом листе дерево отвечает некоторой константой

3. Как строить решающие деревья

Рекурсивное построение

Строим разбиение выборки по значению одного из признаков

$$x^{(j)} < t$$

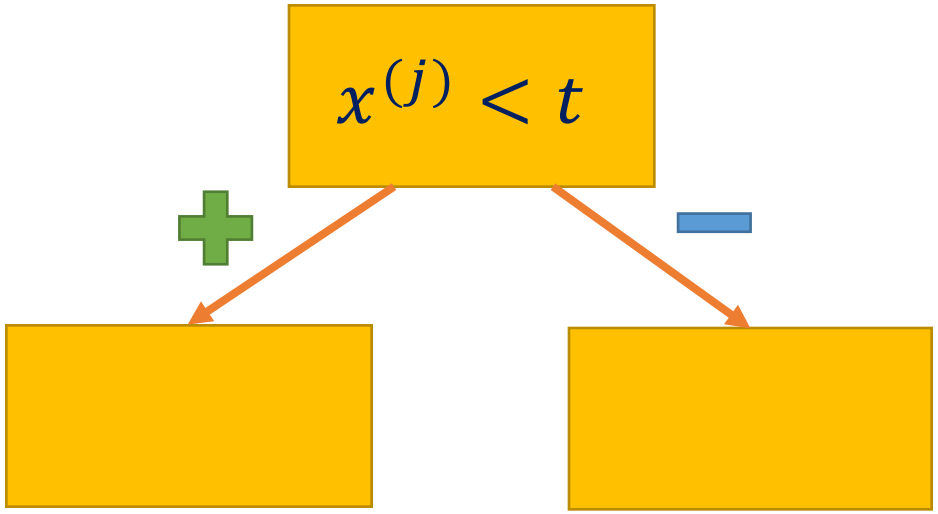
Рекурсивное построение

Строим разбиение выборки по значению одного из признаков

$$x^{(j)} < t$$

Фактически нужно только выбрать j и t наилучшим образом

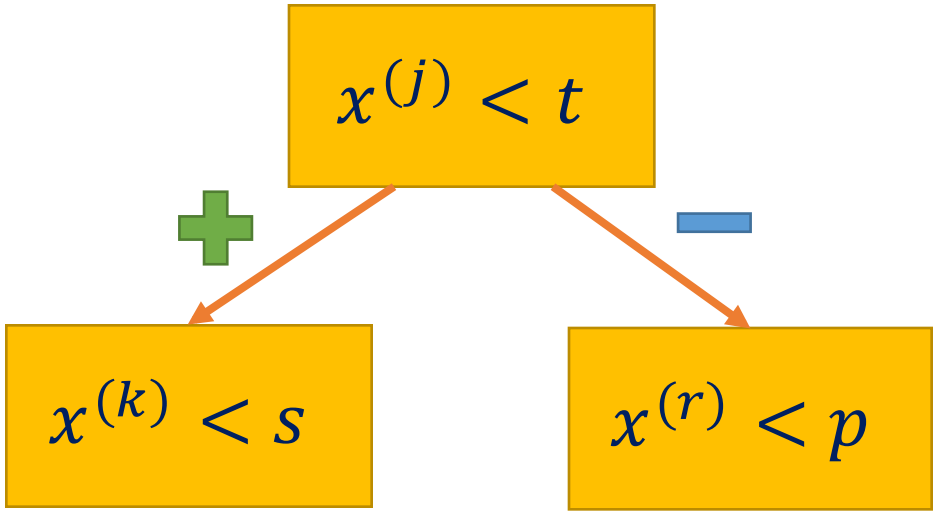
Рекурсивное построение



Выборка делится по этому условию на две части

Рекурсивное построение

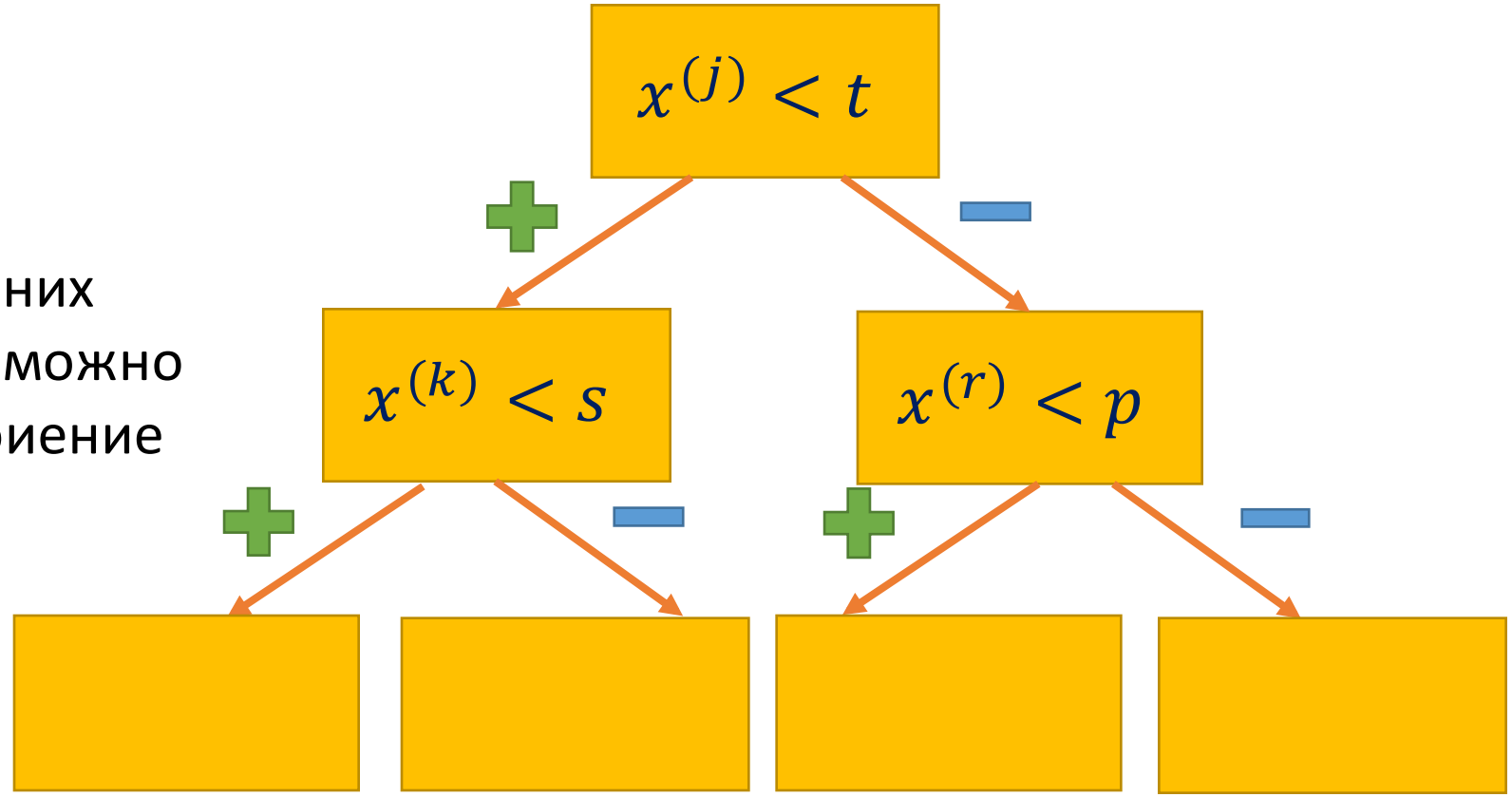
В каждой из них теперь тоже можно сделать разбиение



Выборка делится по этому условию на две части

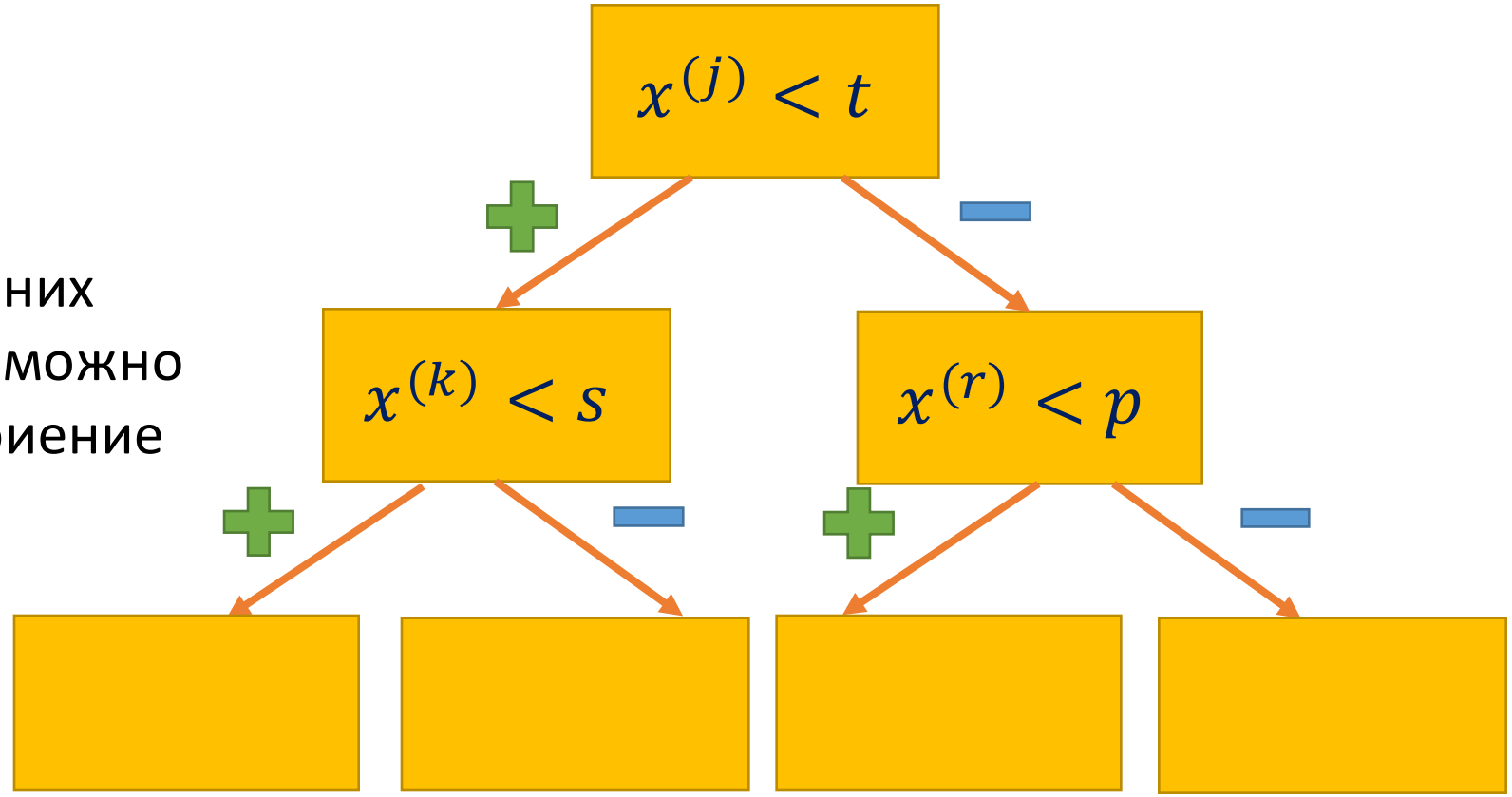
Рекурсивное построение

В каждой из них теперь тоже можно сделать разбиение



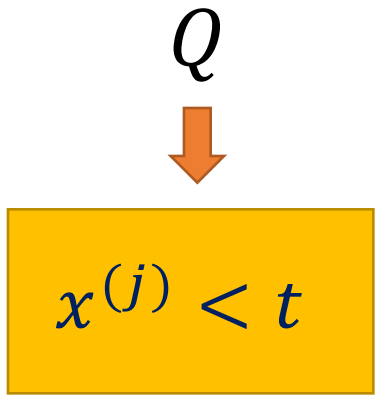
Рекурсивное построение

В каждой из них теперь тоже можно сделать разбиение

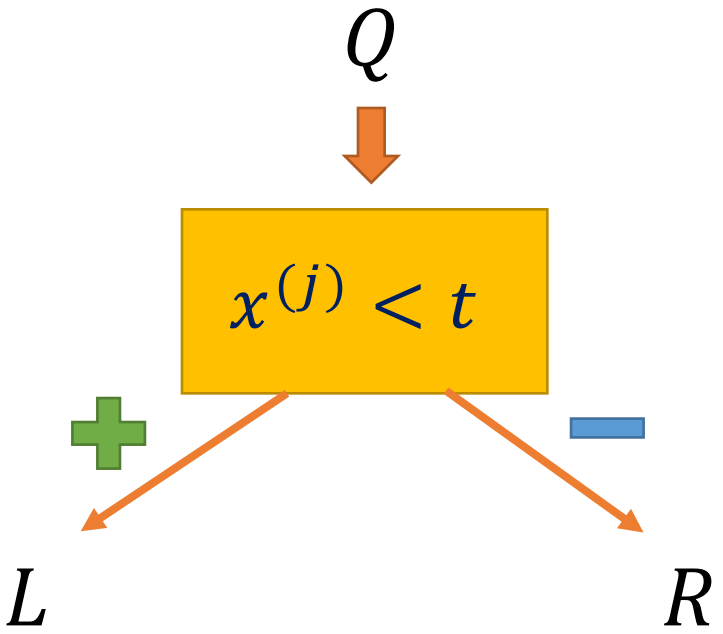


Процесс можно продолжать в тех узлах, в которые попадает достаточно много объектов

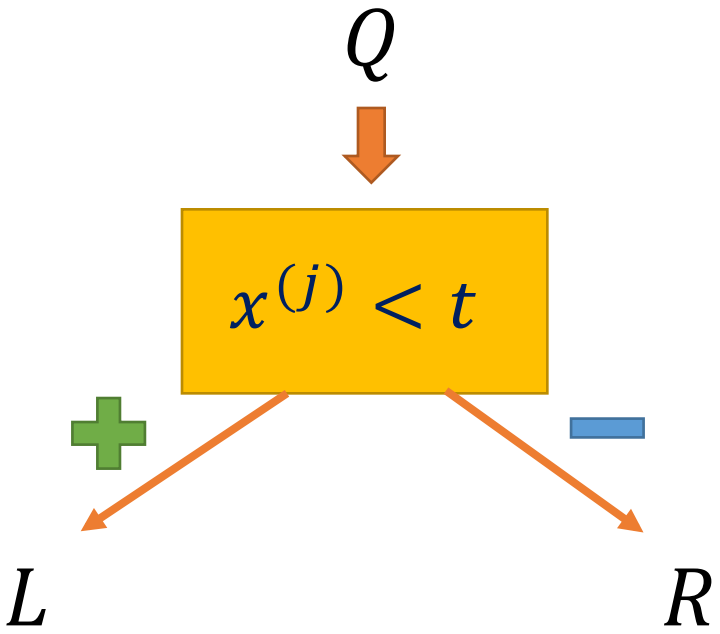
Выбор разбиения



Выбор разбиения

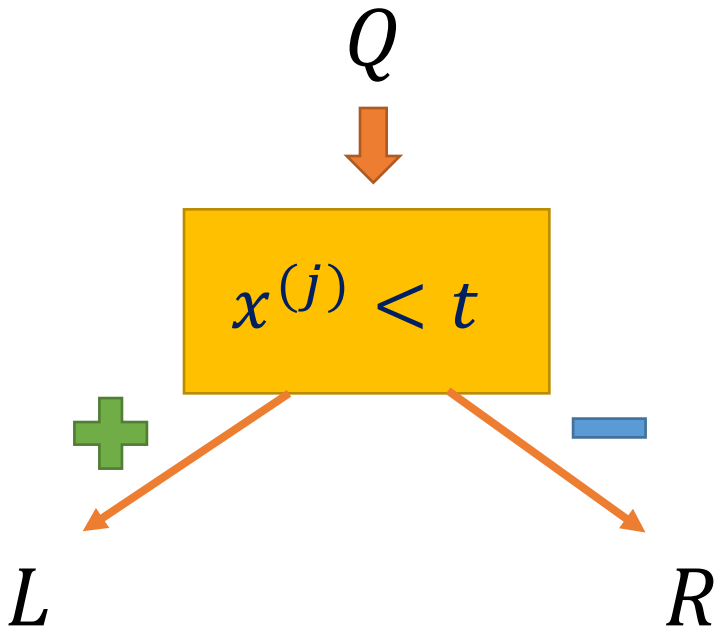


Выбор разбиения



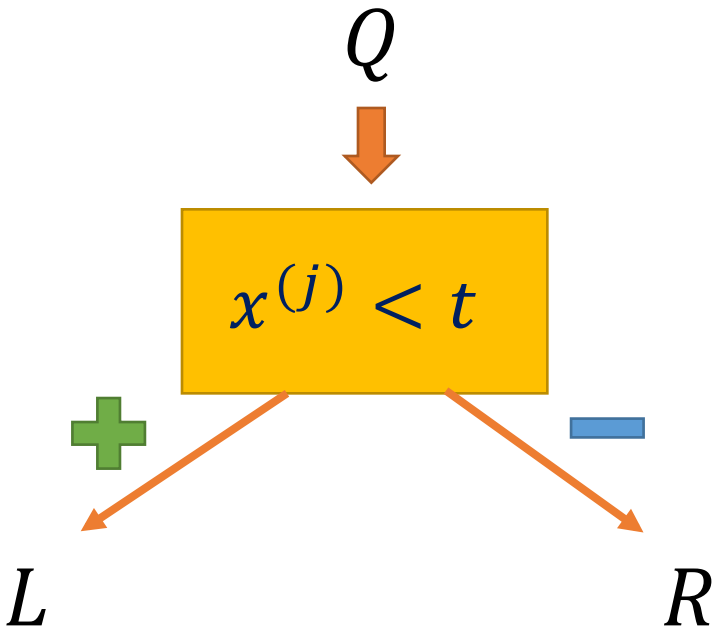
$$G(j,t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R)$$

Выбор разбиения



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

Выбор разбиения



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

$H(R)$ - мера «неоднородности» множества R

Критерии построения разбиений

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

Критерии построения разбиений

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

Пусть мы решаем задачу классификации на 2 класса,
p_0, p_1 – доли объектов классов 0 и 1 в R

1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$
2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$
3) Gini criteria: $H(R) = 1 - {p_0}^2 - {p_1}^2 = 2p_0p_1$

Критерии построения разбиений

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

Пусть мы решаем задачу классификации на К классов,
$p_1, \dots, p_K$ – доли объектов классов 1, …, К в R

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

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

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

Критерии построения разбиений

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

Чтобы решать задачу регрессии, достаточно взять среднеквадратичную ошибку в качестве $H(R)$:

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

Критерии построения разбиений

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

Чтобы решать задачу регрессии, достаточно взять среднеквадратичную ошибку в качестве $H(R)$:

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

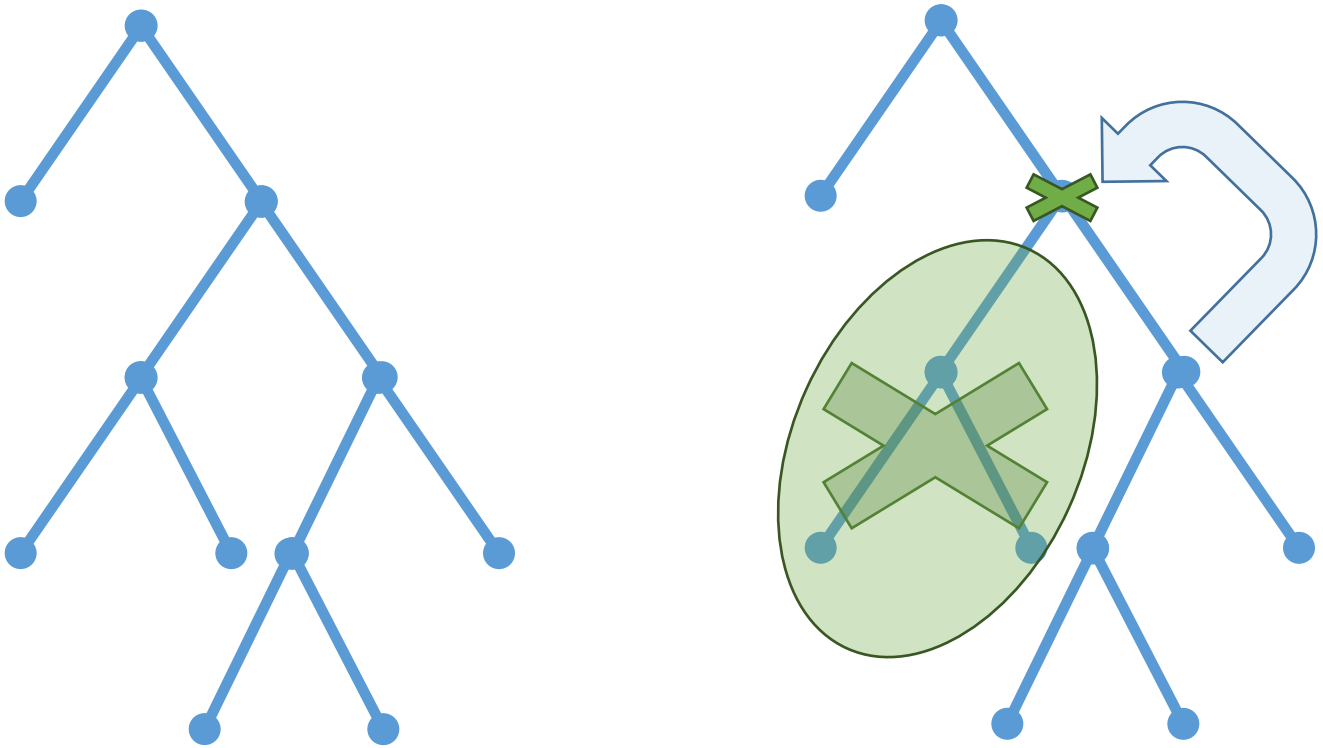
$$\bar{y} = \frac{1}{|R|} \sum_{x_i \in R} y_i$$

4. Дополнительные темы

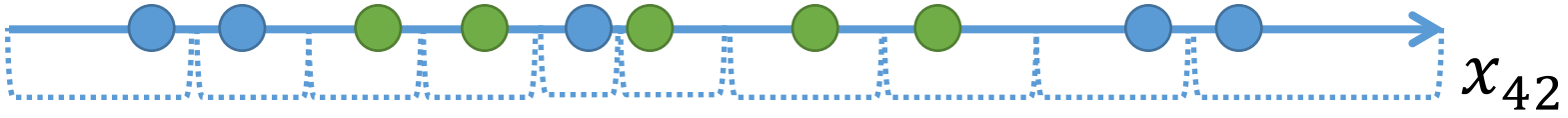
Prunning

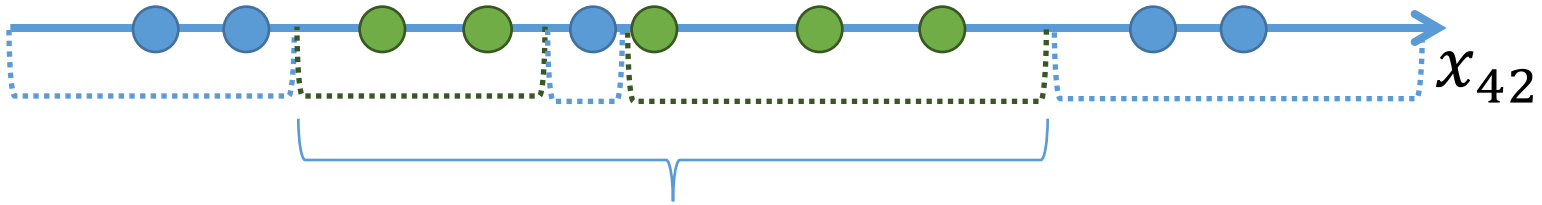
- Pre-prunning:
  - Ограничиваем рост дерева до того как оно построено
  - Если в какой-то момент информативность признаков в разбиении меньше порога – не разбиваем вершину
- Post-prunning:
  - Упрощаем дерево после того как дерево построено

Post-prunning



Бинаризация





Вариации алгоритма построения

- C4.5
- C5.0
- CART

Итог

1. Что такое решающие деревья
2. Решающие деревья в классификации и регрессии
3. Как строить решающие деревья
4. Дополнительные темы

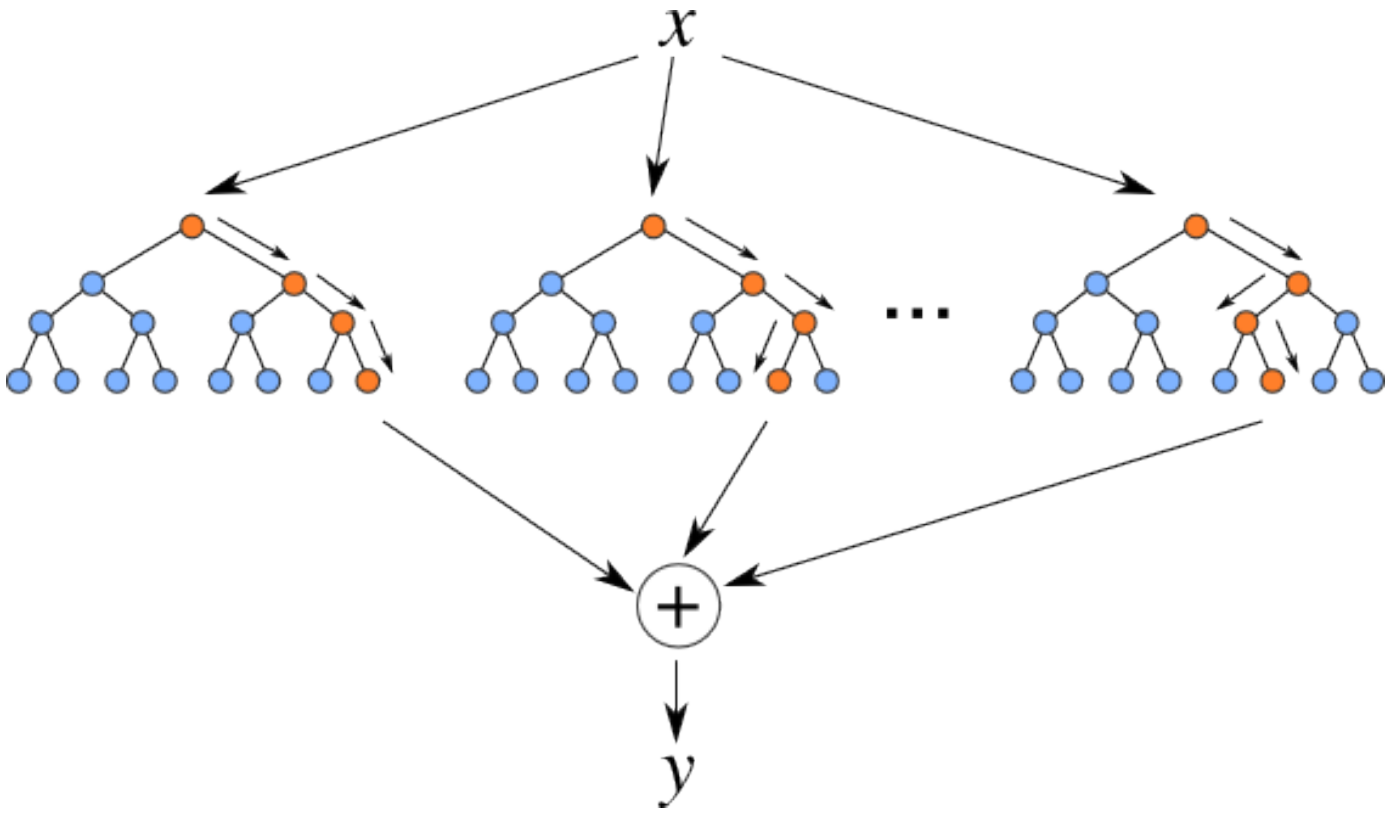
II. Ансамбли деревьев

План

1. Random Forest
2. Gradient Boosted Decision Trees (GBDT)
3. Библиотеки

Random Forest

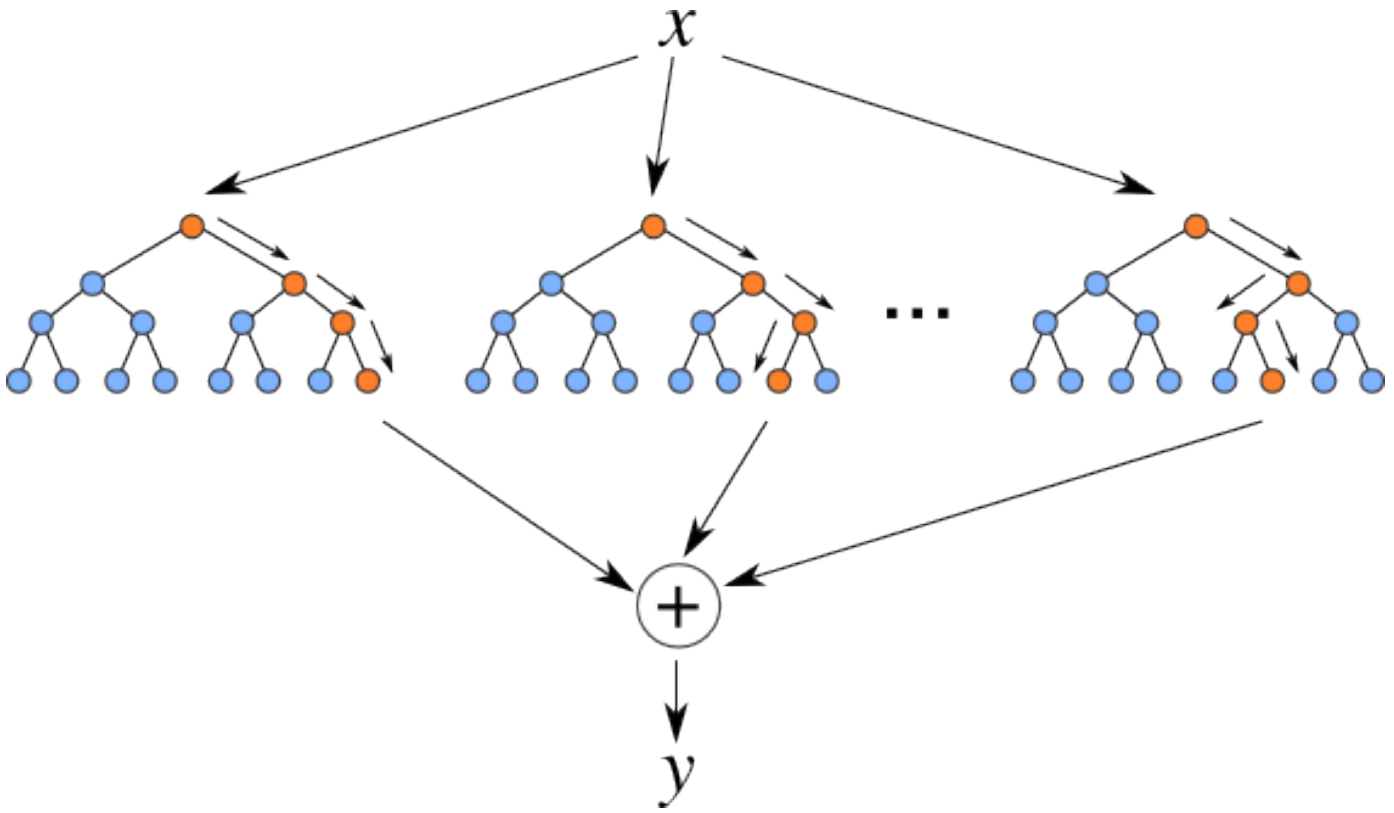
1. Генерируем M выборок на основе имеющейся



Random Forest

1. Генерируем M выборок на основе имеющейся

2. Строим на них деревья с рандомизированными разбиениями в узлах: выбираем k случайных признаков и ищем наиболее информативное разбиение по ним

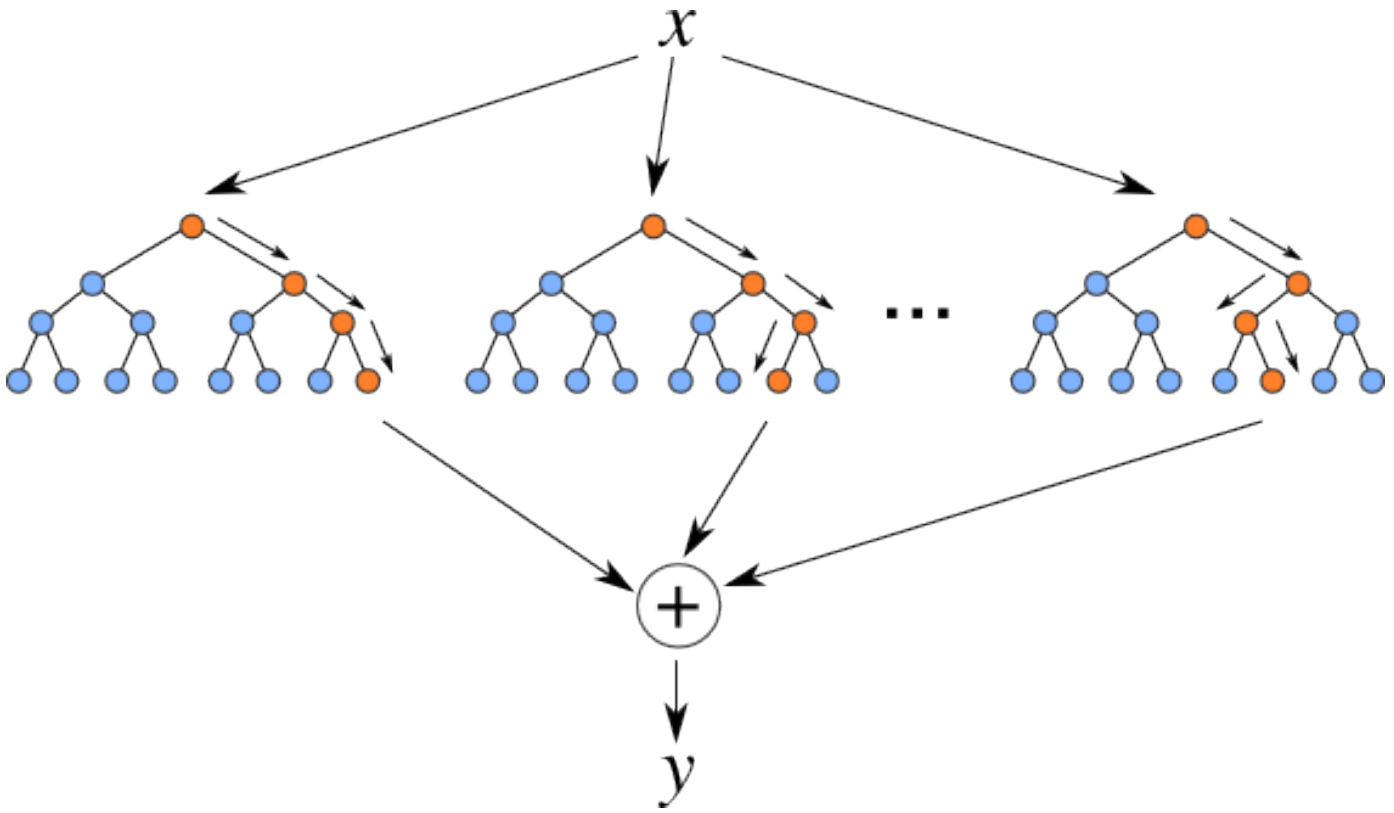


Random Forest

1. Генерируем M выборок на основе имеющейся

2. Строим на них деревья с рандомизированными разбиениями в узлах: выбираем k случайных признаков и ищем наиболее информативное разбиение по ним

3. При прогнозе усредняем ответ всех деревьев

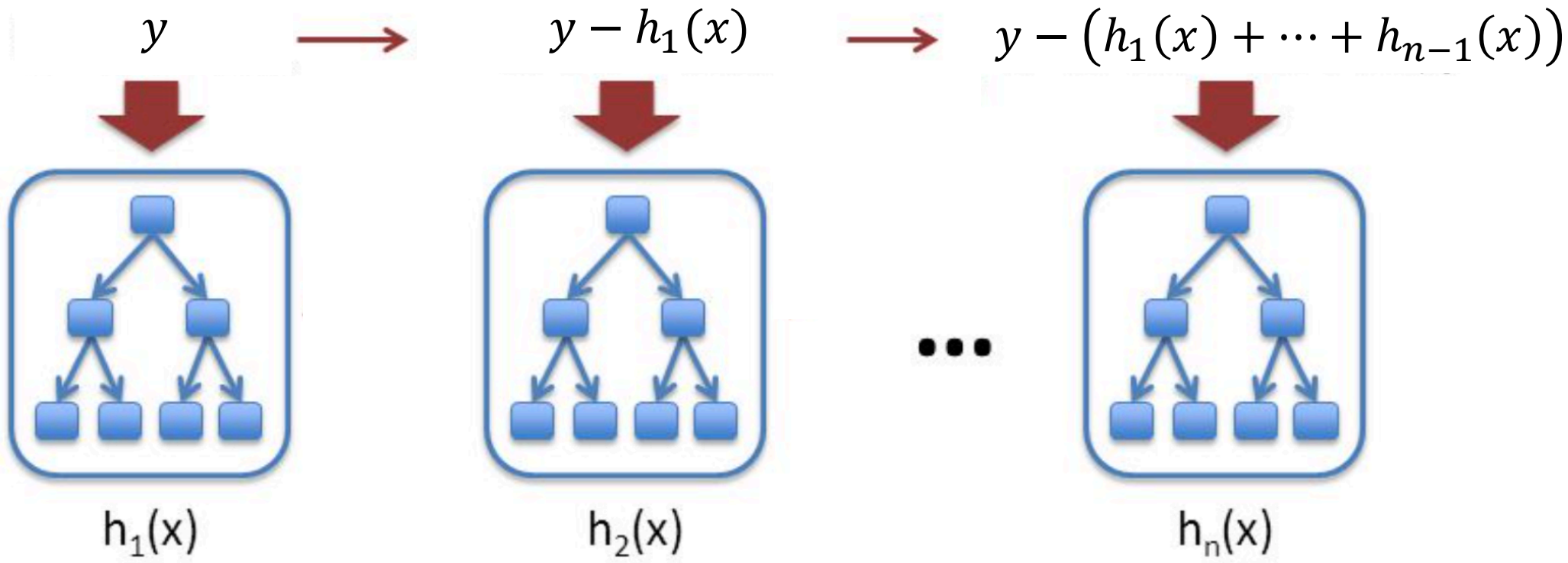


Идея Gradient Boosted Decision Trees (GBDT)

$$h(x) = h_1(x) + \cdots + h_n(x)$$

Идея Gradient Boosted Decision Trees (GBDT)

$$h(x) = h_1(x) + \cdots + h_n(x)$$



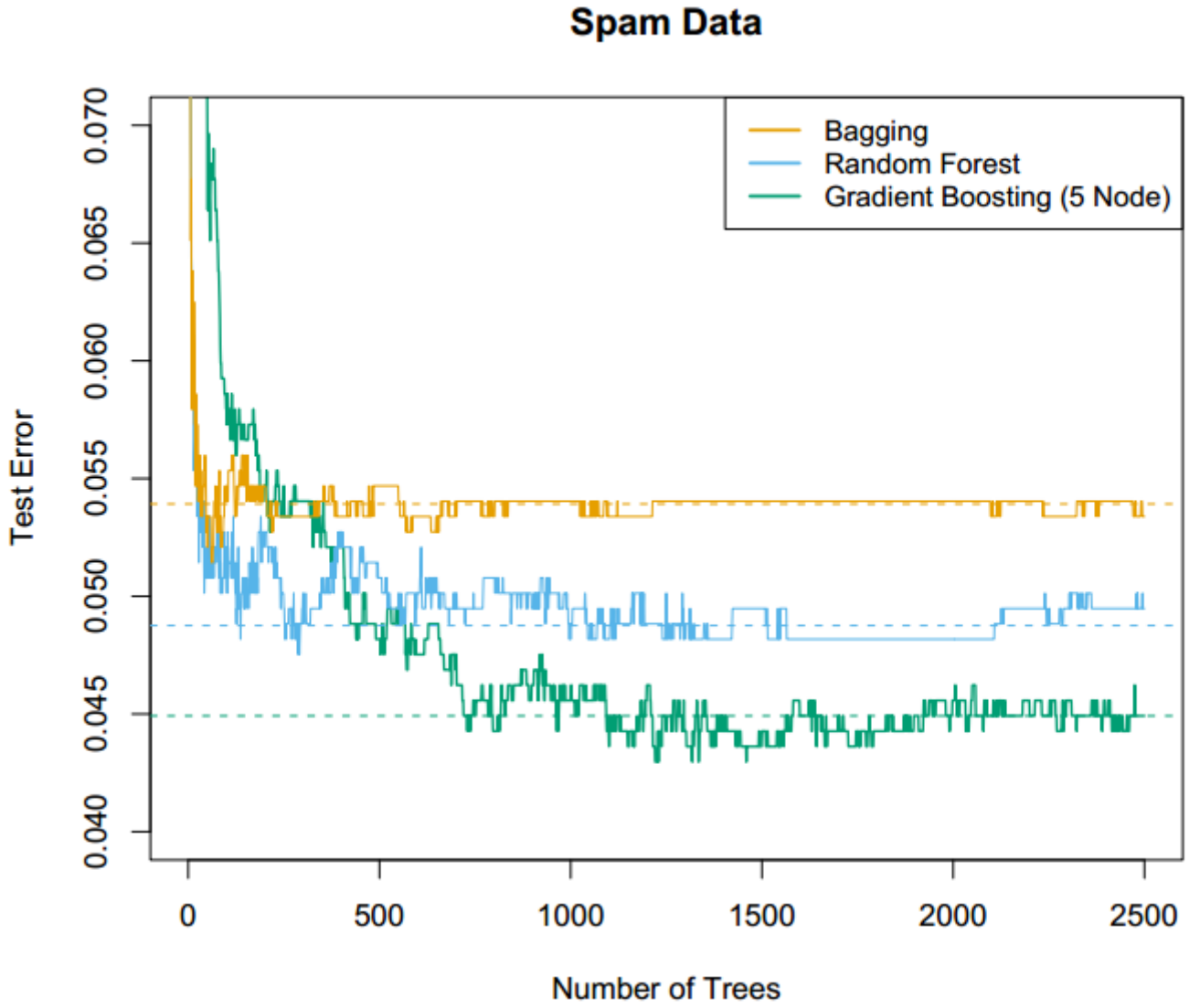
Gradient Boosted Decision Trees

- Каждое новое дерево $h_k(x)$ обучаем на ответы $y_i - h_i$

  h_i - прогноз всей композиции на i-том объекте на предыдущей итерации

- Коэффициент α_k перед новым деревом подбираем с помощью численной оптимизации ошибки

GBDT и RF


Spam Data
Bagging
Random Forest
Gradient Boosting (5 Node)
Test Error
0.040
0.045
0.050
0.055
0.060
0.065
0.070
0
500
1000
1500
2000
2500
Number of Trees

Библиотеки

- Scikit-learn:
  - sklearn.ensemble.RandomForestClassifier
  - sklearn.ensemble.RandomForestRegressor
- XGBoost

Итог

1. Random Forest
2. Gradient Boosted Decision Trees (GBDT)
3. Библиотеки

Резюме

I. Решающие деревья
II. Ансамбли решающих деревьев