Машинное обучение

Лекция 5

Решающие деревья и ансамбли деревьев: дополнительные темы

Виктор Кантор

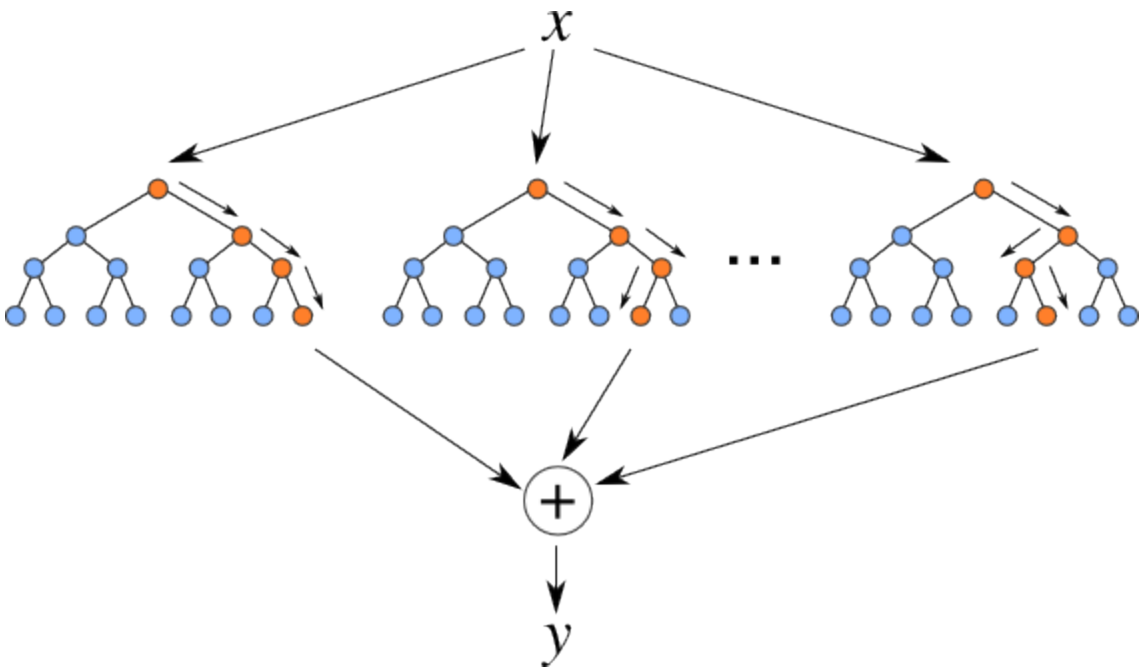
План

I. Ансамбли решающих деревьев
   a) Анализ RF и GBDT
   b) XGBoost
   c) Важность признаков

II. Решающие деревья
   a) Критерии информативности
   b) Прунниг
   c) Важность признаков
   d) Категориальные признаки
   e) Пропущенные значения
   f) ID3, C4.5, C5.0, CART

I. Ансамбли решающих деревьев: дополнительные темы

Random Forest



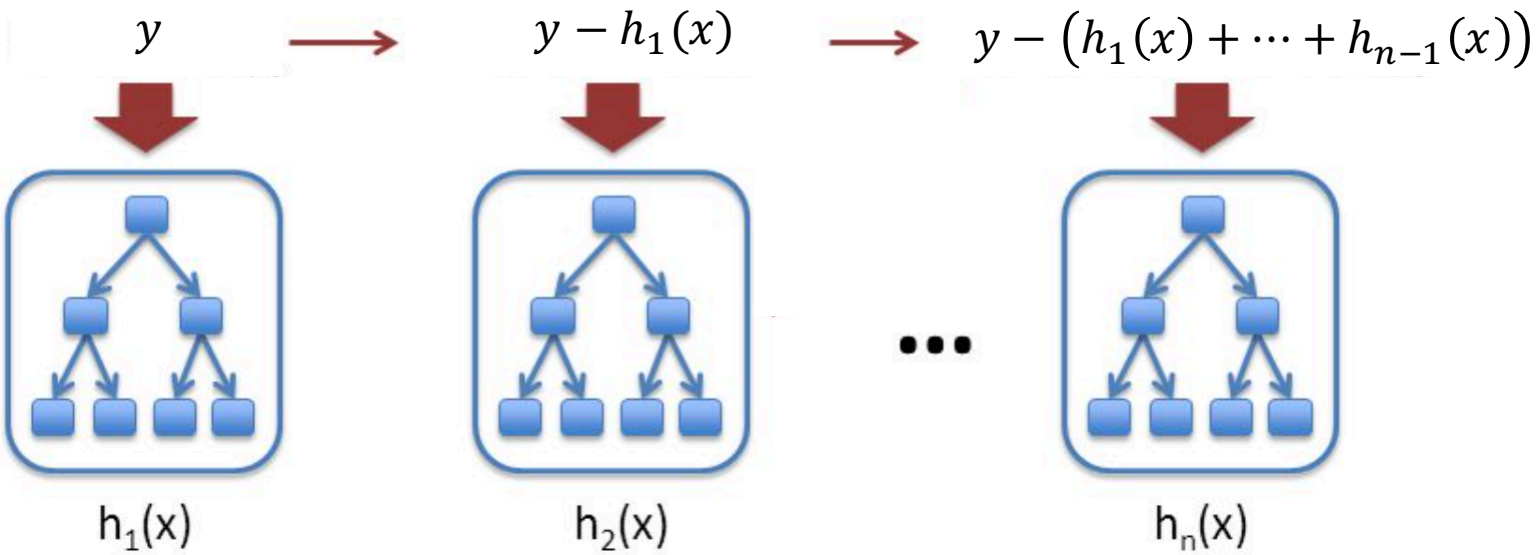
1. Бэггинг над деревьями
2. Рандомизированные разбиения в деревьях: выбираем k случайных признаков и ищем наиболее информативное разбиение по ним

Ошибка усредненной модели

Ошибка усредненной модели

Идея Gradient Boosted Decision Trees

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



GBM в наиболее общем виде

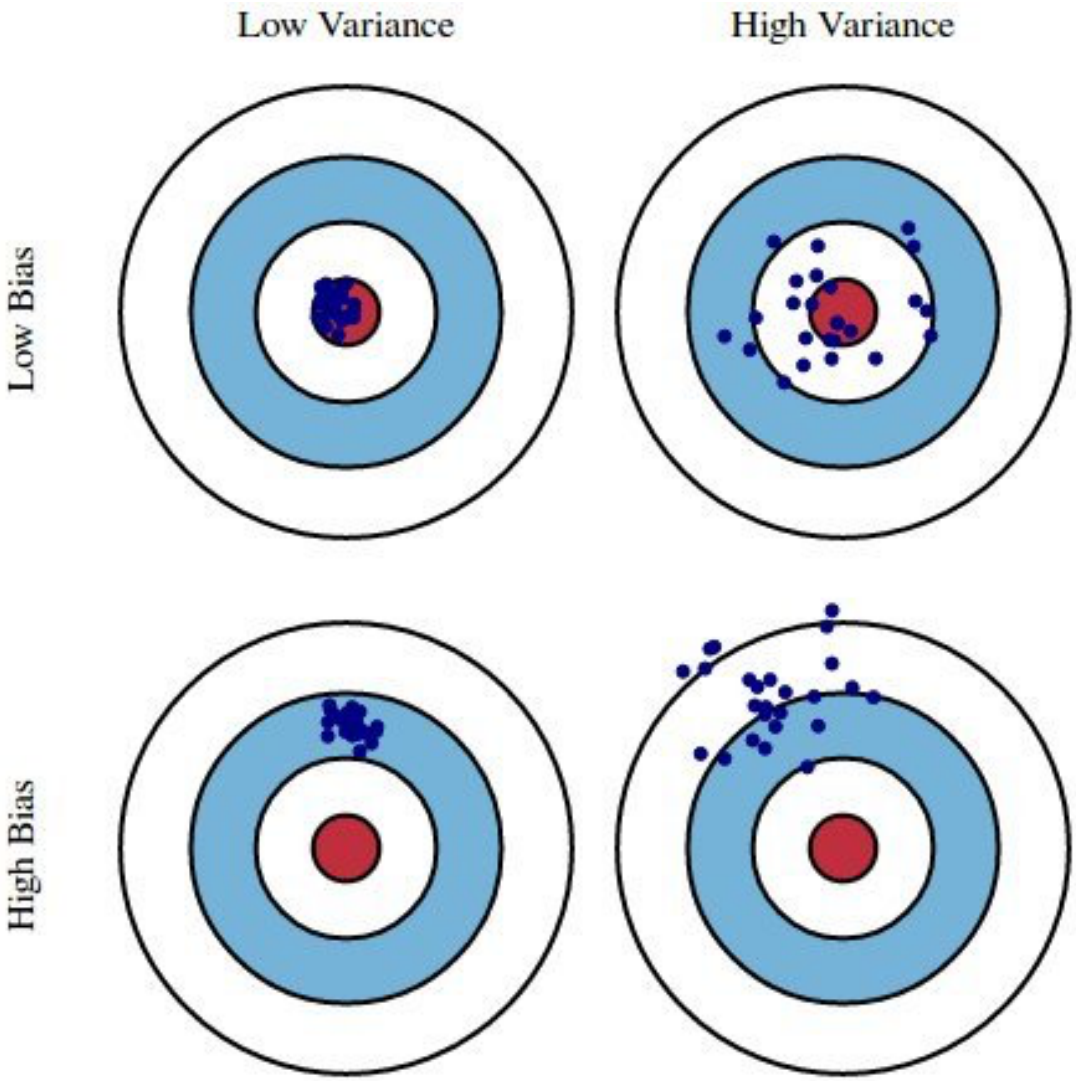
1. Обучаем первый базовый алгоритм $h_1,\ \beta_1 = 1$
2. Повторяем в цикле по t от 2 до T:

$$h_t = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{l} \tilde{L}\left(h(x_i), -\frac{\partial L(\hat{y}_i, y_i)}{\partial \hat{y}_i}\right)$$

выбираем β_t

Здесь $Q(\hat{y}, y) = \sum_{i=1}^{l} L(\hat{y}_i, y_i)$ $\qquad\qquad$ $\hat{y}_i = a_{t-1}(x_i)$

Bias-variance trade-off



Bias-variance-noize decomposition

Theorem. For the squared error loss, the bias-variance decomposition of the expected generalization error at $X = \mathbf{x}$ is

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = \text{noise}(\mathbf{x}) + \text{bias}^2(\mathbf{x}) + \text{var}(\mathbf{x})$$

where

$$\begin{aligned}
\text{noise}(\mathbf{x}) &= Err(\varphi_B(\mathbf{x})), \\
\text{bias}^2(\mathbf{x}) &= (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2, \\
\text{var}(\mathbf{x}) &= \mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}.
\end{aligned}$$

Сдвиг и разброс в бэггинге

Сдвиг и разброс в бустинге

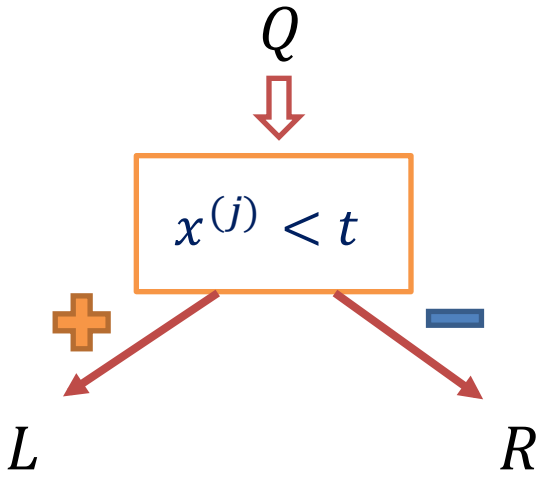
eXtreme Gradient Boosting (XGBoost)

Важность признаков

- Out-of-bag в RF
- Что в GBM??????????

II. Решающие деревья: дополнительные темы

Выбор разбиения



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

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: $H(R) = 1 - {p_0}^2 - {p_1}^2 = 2p_0 p_1$

Критерии построения разбиений

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

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

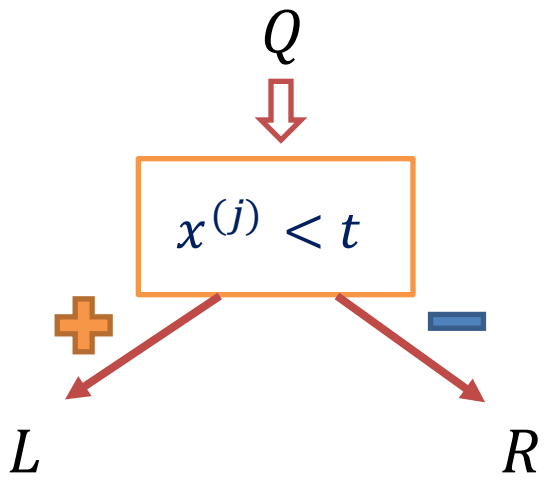
$H(R)$ – мера «неоднородности» множества R

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

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

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

Критерии информативности



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

Gini

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

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

Information gain

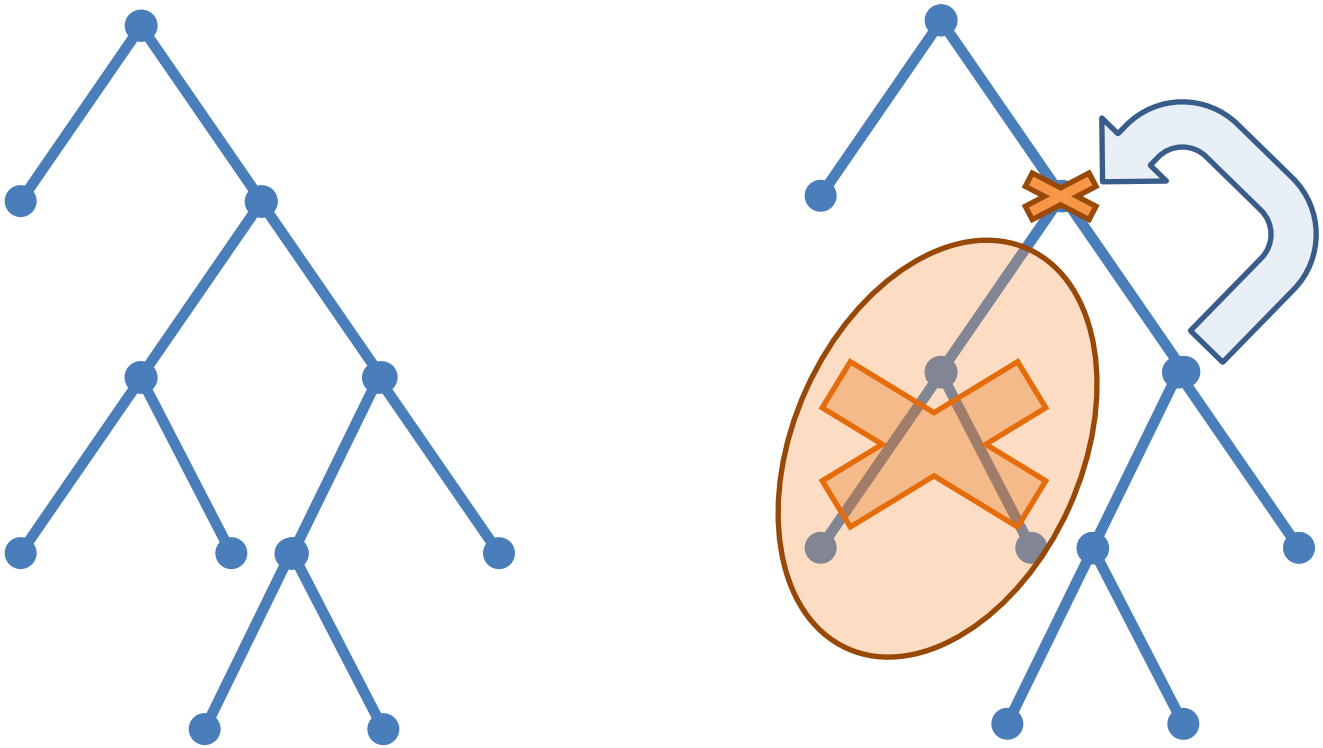
$$I(Q, j, t) = H(Q) - \frac{|L|}{|Q|} H(L) - \frac{|R|}{|Q|} H(R)$$

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

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



Важность признаков

Категориальные признаки

Пропущенные значения

ID3: Iterative Dichotomizer 3

C4.5

Information gain ratio

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

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

C5.0

CART: построение дерева

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

Для классификации:

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

Для регрессии:

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

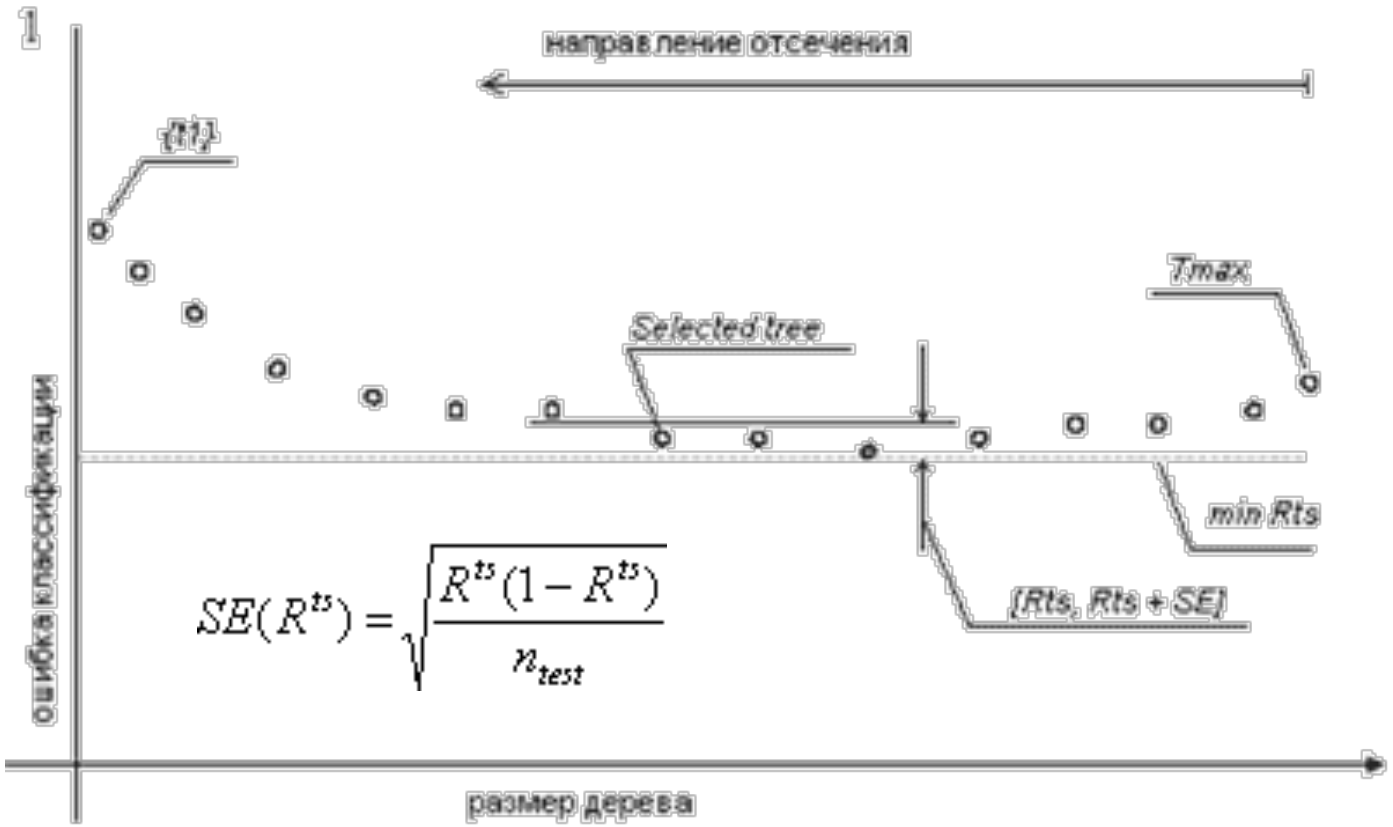
$$\bar{y} = \frac{1}{|R|} \sum_{x_i \in R} y_i$$

2 особенности CART

- Minimal cost-complexity prunning
- V-fold

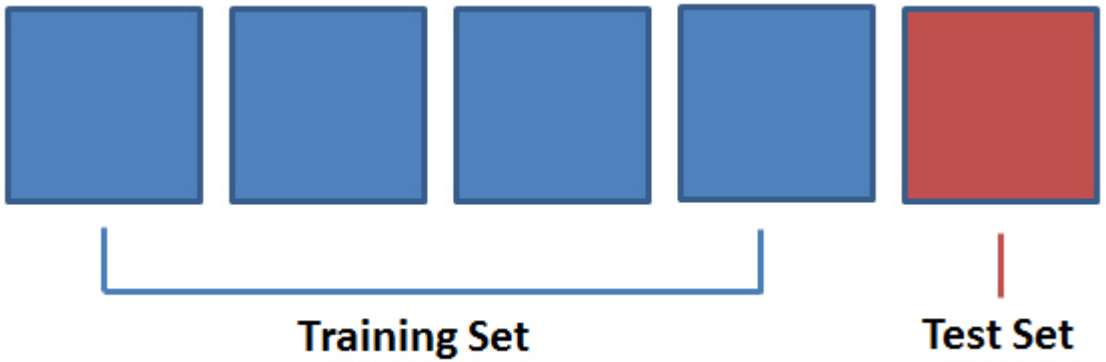
CART: cost-complexity prunning

$$C_\alpha(T) = R(T) + \alpha|T|$$

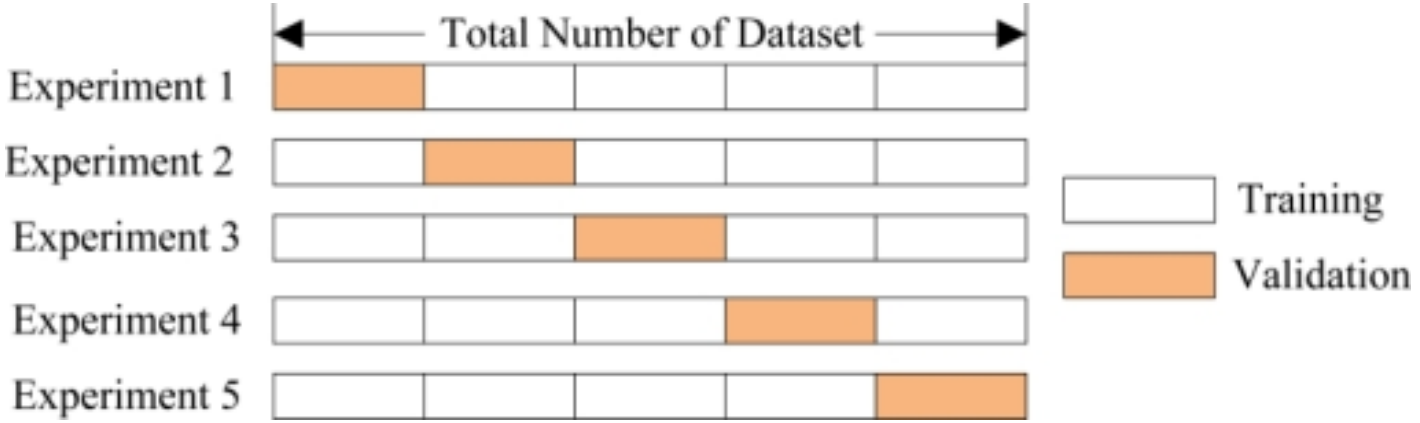


CART: V-fold cross-validation

- Как можно выбирать размер дерева:



- Как лучше выбирать размер дерева:



Резюме

I. Ансамбли решающих деревьев
   a) Анализ RF и GBDT
   b) XGBoost
   c) Важность признаков

II. Решающие деревья
   a) Критерии информативности
   b) Прунниг
   c) Важность признаков
   d) Категориальные признаки
   e) Пропущенные значения
   f) ID3, C4.5, C5.0, CART