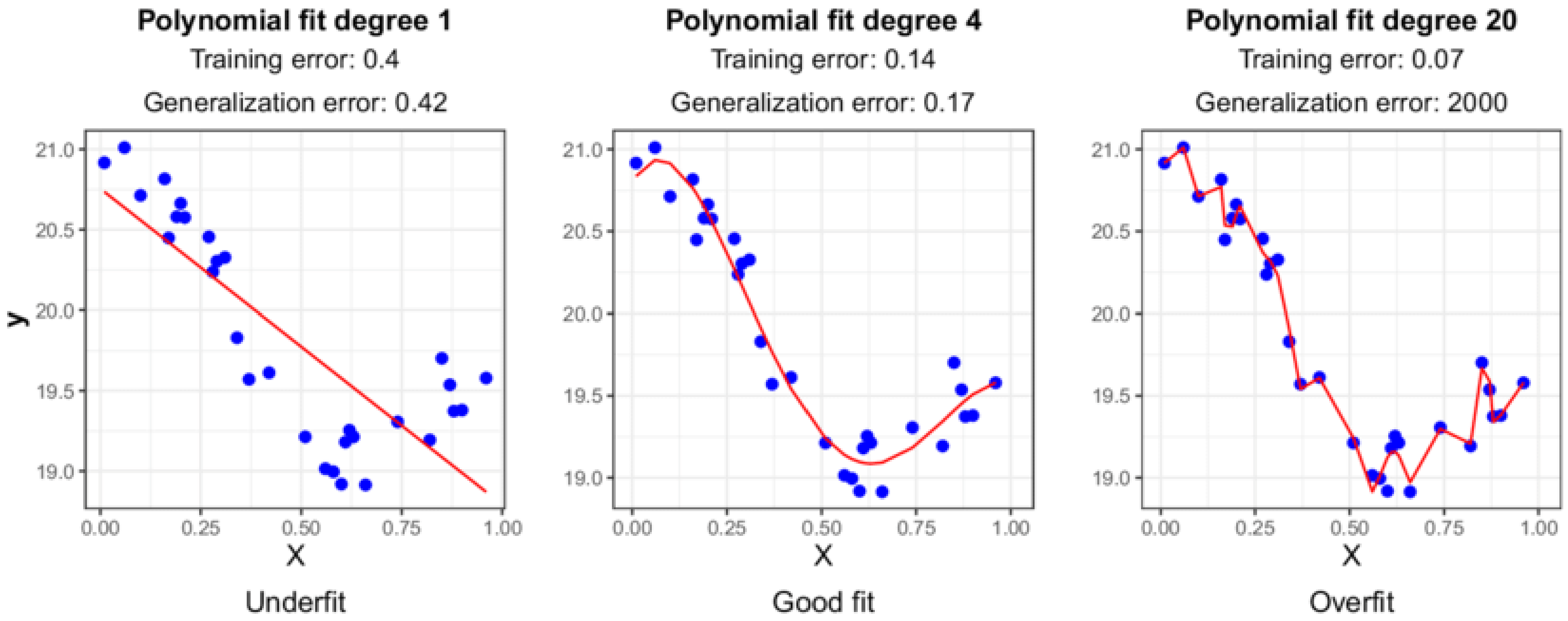
Регуляризация

2022

Мария Корлякова

Переобучение

- Overfiting
- Underfiting



Переобучение

- Cross-validation
- Hold Out
- Регуляризация

Регуляризация

- Штраф сложности

- L2

$$Q(w, X) + \lambda||w||^2 \to \min_w.$$

- L1

$$||w||_1 = \sum_{j=1}^{d} |w_j|.$$

Урок 4. Алгоритм построения дерева решений.

2021

Мария Корлякова

Деревья решений

1. ГЛАЗА : БИНОКУЛЯРНОЕ.

2. УШИ: ОСТРЫЕ.

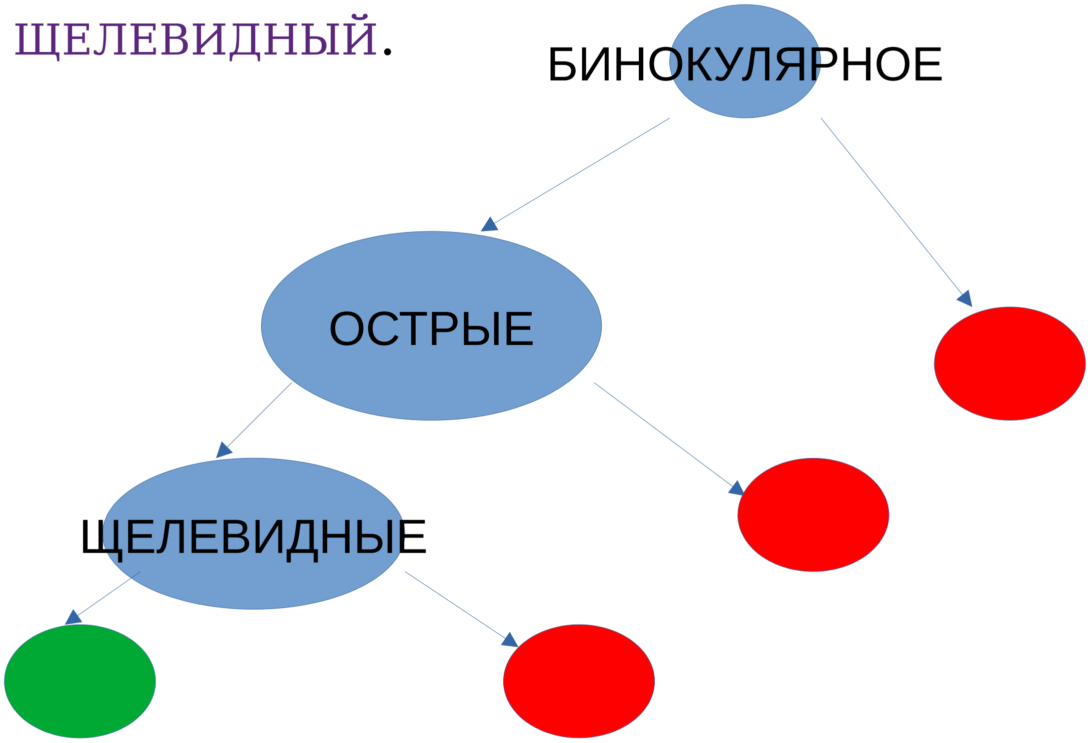
3. ЗРАЧОК: ЩЕЛЕВИДНЫЙ.

Деревья решений

1. ГЛАЗА : БИНОКУЛЯРНОЕ.

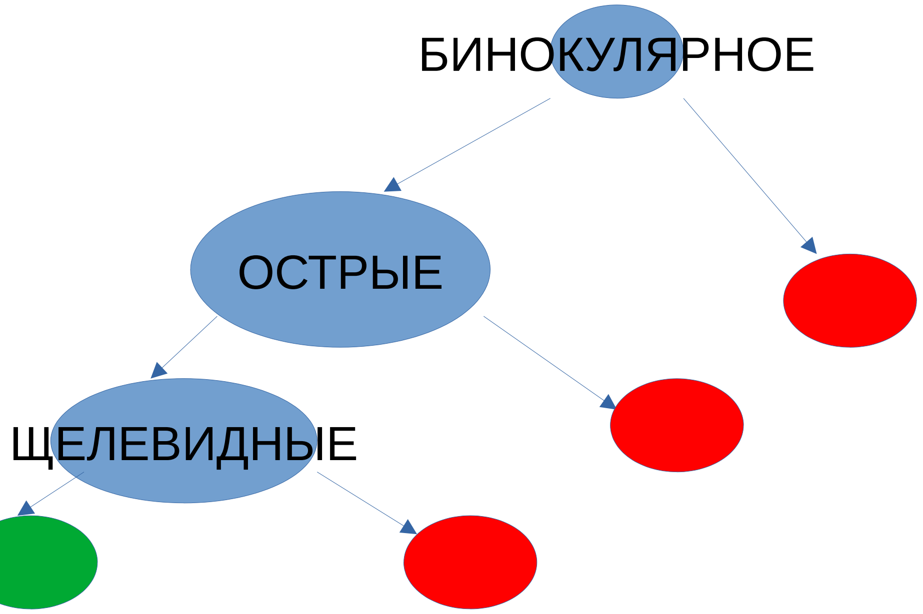
2. УШИ: ОСТРЫЕ.

3. ЗРАЧОК: ЩЕЛЕВИДНЫЙ.



Деревья решений: Алгоритм

1. Получить примеры
2. Проверить критерий останова
3.Выбрать по примерам лучший признак
4. Выбрать лучшее разделение признака
5. получить 2 ветви: для каждой ветки перейти к п.1



Классификация

ПРОБЛЕМЫ:
1) какой признак делим
2) как делим

$$[x^j \leq t].$$

3) когда остановится

$$Q(X, j, t).$$

$$a_m = \mathrm{argmax}_{y \in Y} \sum_{i \in X_m} [y_i = y]$$

4) как назначить значение в терминальном узле

$$a_{mk} = \frac{1}{|X_m|} \sum_{i \in X_m} [y_i = k].$$

Регрессия:

ПРОБЛЕМЫ:
1) какой признак делим $[x^j \leq t]$.
2) как делим

3) когда остановится $Q(X, j, t)$.

$$a_m = \mathrm{argmax}_{y \in Y} \sum_{i \in X_m} [y_i = y]$$

4) как назначить значение в терминальном узле

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i.$$

Критерий информативности

Регрессия : $$H(X) = \frac{1}{X}\sum_{i\in X}(y_i - \bar{y}(X))^2;$$

Классификация:

- Вероятность верной классификации $$p_k = \frac{1}{|X|}\sum_{i\in X}[y_i = k].$$

- Джини $$H(X) = \sum_{k=1}^{K} p_k(1 - p_k);$$

- Энтропия Шеннона $$H(X) = -\sum_{k=1}^{K} p_k \log_2 p_k.$$

Критерии останова

- Ограничение максимальной глубины дерева.

- Ограничение максимального количества листьев.

- Ограничение минимального количества объектов в листе.

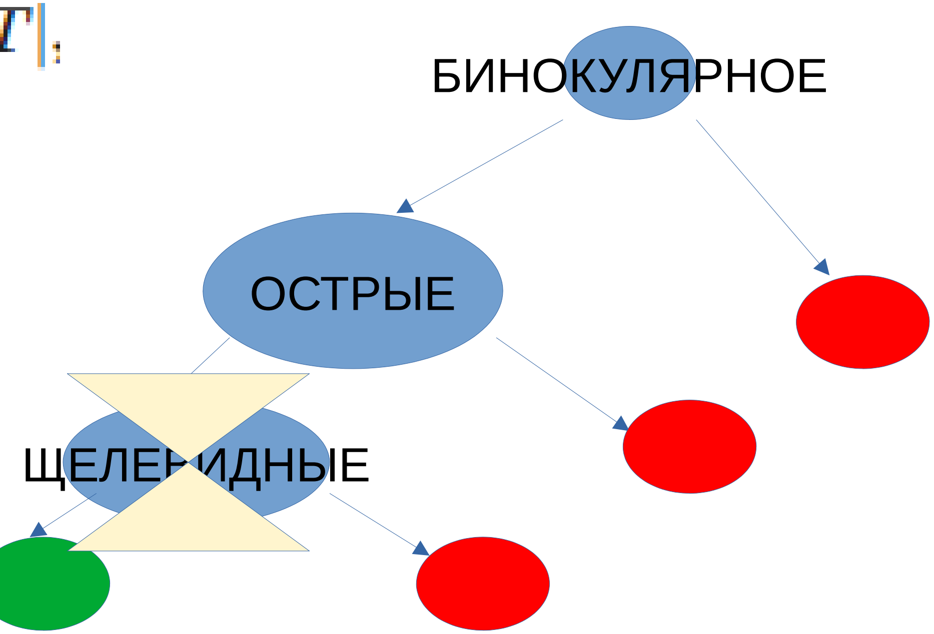
- Останов в случае, когда все объекты в листе относятся к одному классу.

- Требование улучшения функционала качества при разбиении на какую-то минимальную величину.

Обрезка

- Pruning

$$R_\alpha(T) = R(T) + \alpha|T|,$$

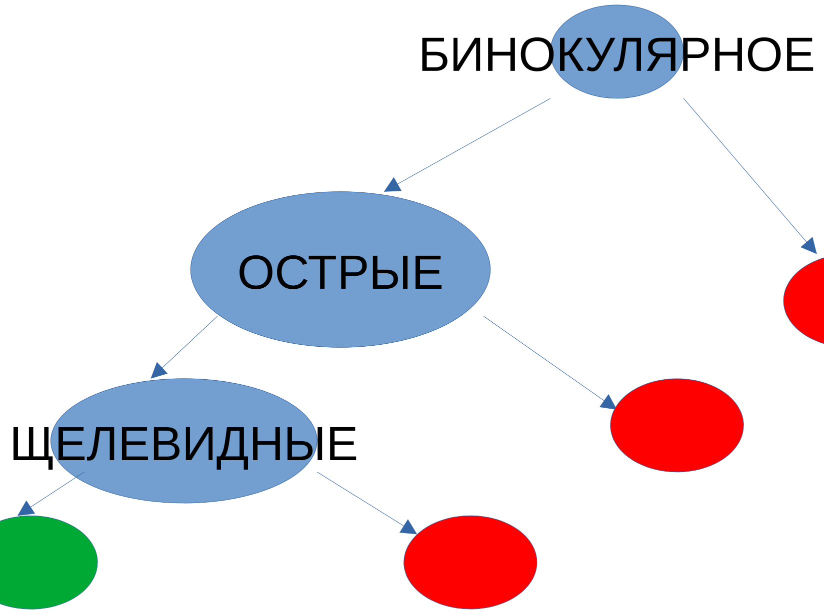


CART (Classification and regression trees)

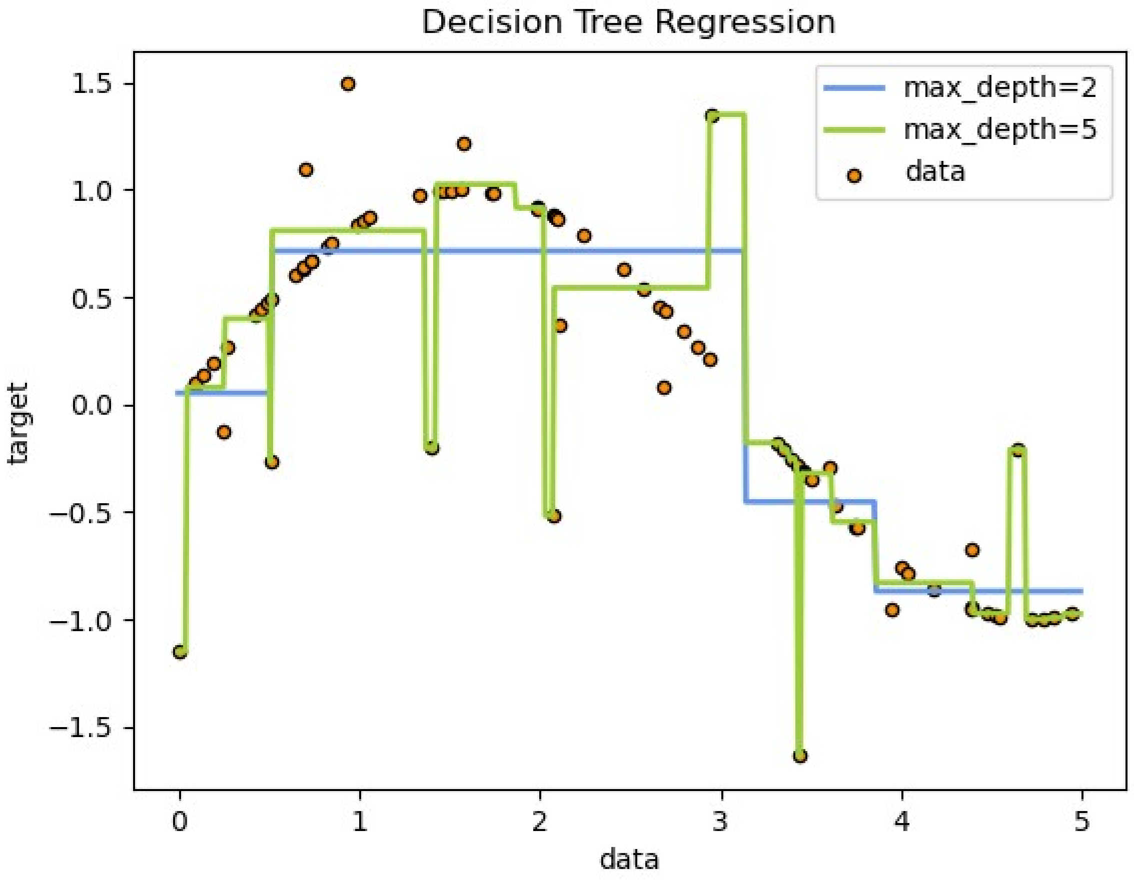
$$H(X) = \sum_{k=1}^{K} p_k(1 - p_k).$$

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

- ID3
- C4.5



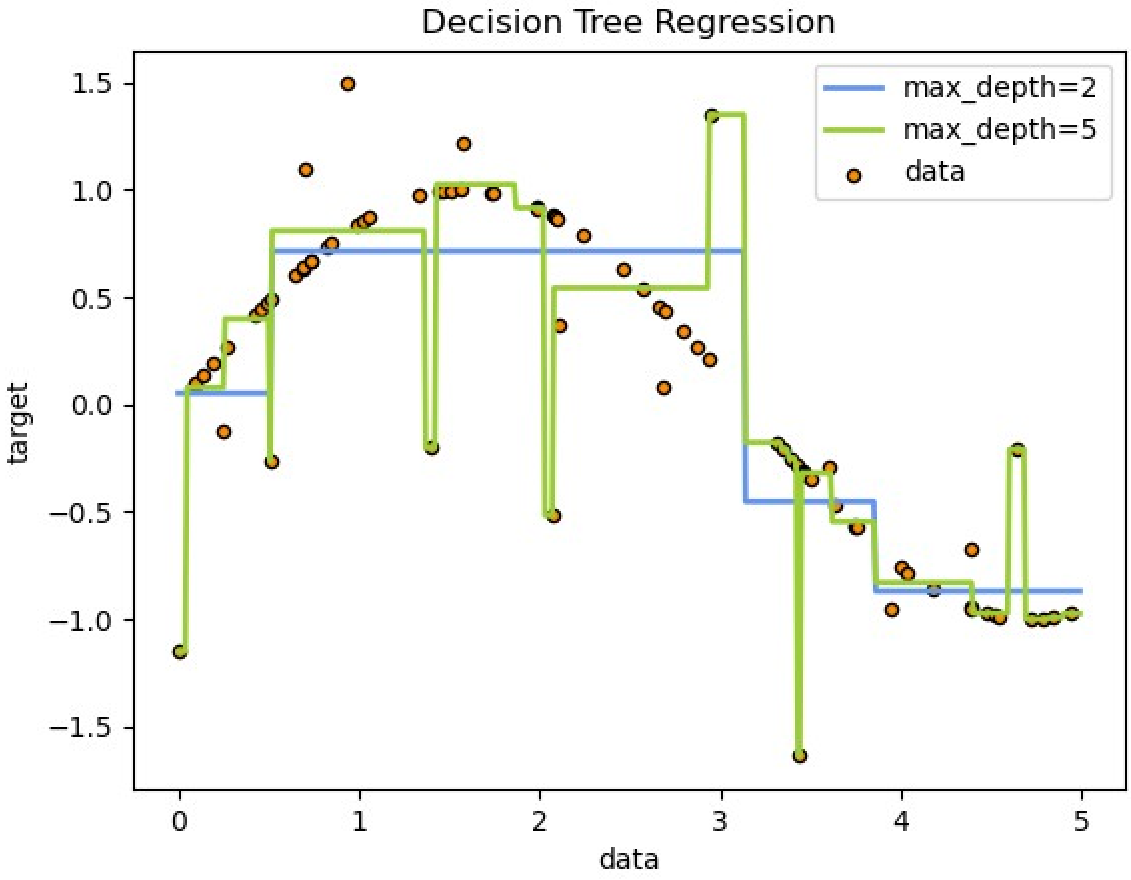
Регрессия



Decision Tree Regression
max_depth=2
max_depth=5
data
target
data
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
2
3
4
5

Регрессия

- Overfiting
- Underfiting



Регрессия и Классификация

- Деревья: Overfiting

Ансамбли

Корлякова Мария.

2022

Регрессионная модель связи X и D

- $D = f(X) + \varepsilon$
- ε - ожидаемая ошибка
  - нормально распределена
  - $M(\varepsilon)=0$
- $f(X)$ – регрессионная модель
- она есть объективно

Свойства модели

1. *Среднее значение ожидаемой ошибки ε для любой реализации X   E[ε|x] = 0, тогда*
2. *f(x) = E[D|x]*
3. *Ошибка ε не коррелирует с функцией регрессии f(X):   E[ε f(X)]=0*

E[ ] - математическое ожидание

Определение w

- Минимизация функции стоимости:

  $F(x_i,w)$ – модель (мы ее строим)

  Критерий:

  $Er(w) = ½ \Sigma (d_i – F(x_i,w))^2$,

  $Er(w) = ½ E_T[(f(x) – F(x,T))^2]$

  W зависят от выборки T={(X,d)}

  Et[ ] - среденее выборочное

Мера прогнозирования

$L_{av}(f(x), F(x,T)) = E_T[(f(x) - F(x,T))^2]$

$f(x) = E[D|x] - f(x)$ *мат. ожидание* $D|x$

$L_{av}(f(x), F(x,T)) = E_m[(E[D|X{=}x] - F(x,T))^2]$

Ошибка оценивания регрессионной функции f(X) аппроксимационной F(x,T)

$(E[D|X=x] - F(x,T)) =$

$(E[D|X=x] - E_T[F(x, T)]) + (E_T[F(x, T)] - F(x,T))$

Тогда

$L_{av}(f(x),F(x,T)) = E_T[(E[D|X=x] - F(x,T))^2] = B^2(w) + V(w) +$

$+ E_T[(E[D|X=x] - f(x))^2]$

* пренебрегаем бесконечно малыми и получим простое выражение

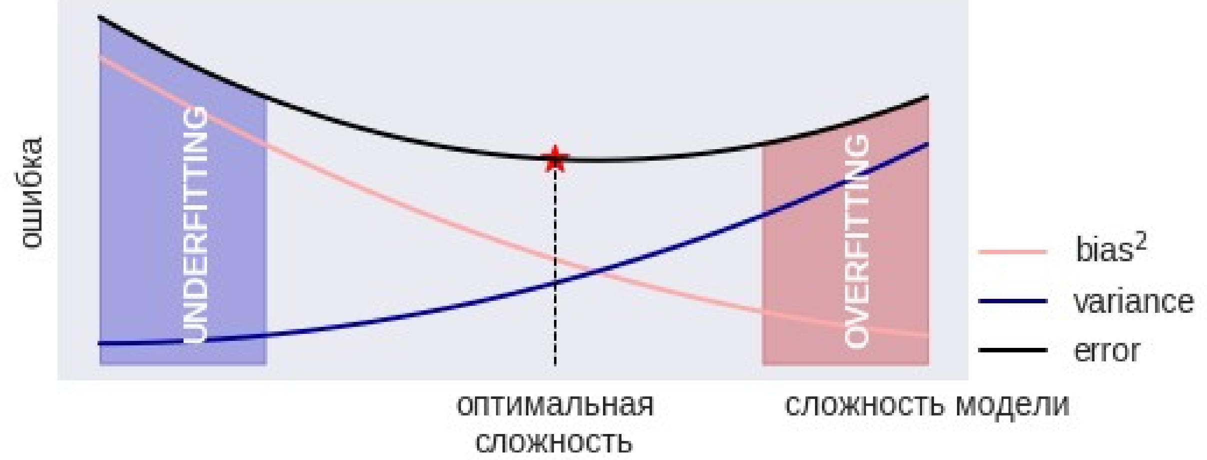
$B(w) = E_T\,[(E[D|X=x] - F(x, T))\,]$ – смещение среднего для $F(x, T)$ *относительно* $f(x) \Rightarrow$ ***ошибка аппроксимации***

$V(w) = E_T\,[\,(E_T[F(x, T)] - F(x,T))^2\,]$ – ***дисперсия*** $F(x,T)$ *на всем* T.

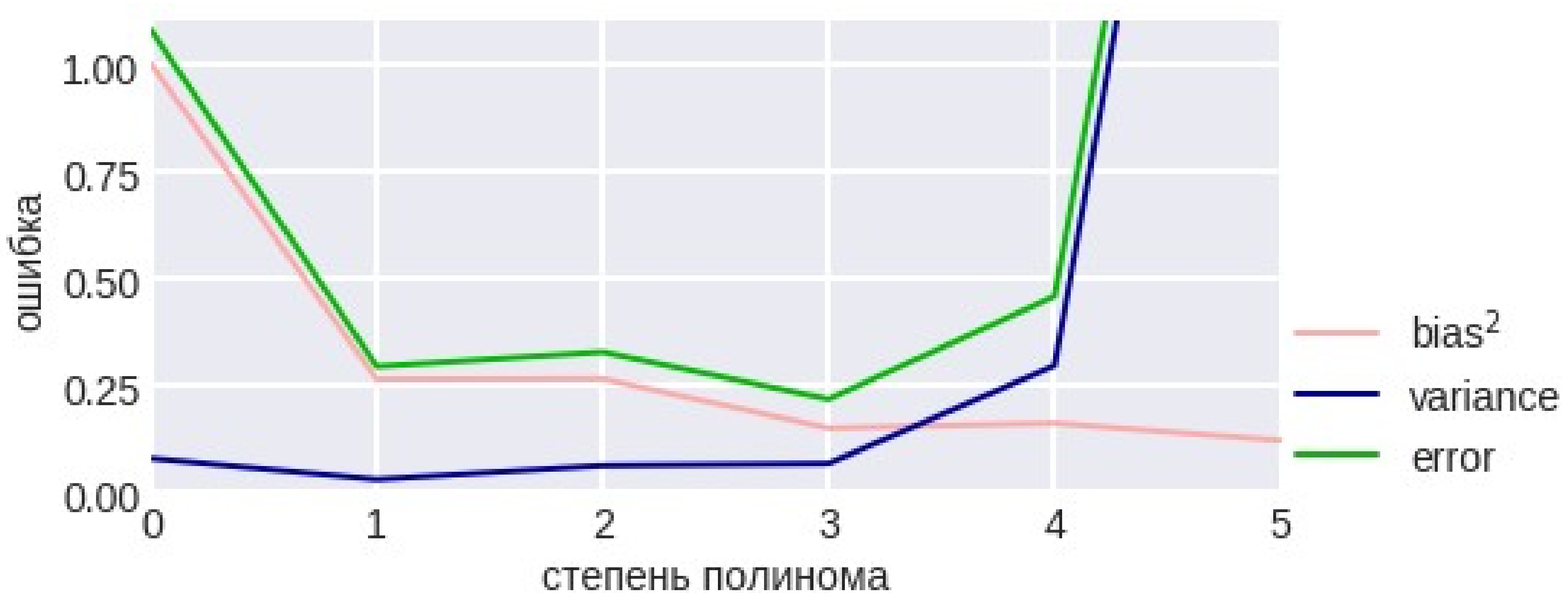
Дилемма дисперсии-смещения

Дилемма дисперсии-смещения

- Одновременно уменьшить смещение и дисперсию можно только для бесконечно большой выборки



Дилемма дисперсии-смещения



Как решить дилему:

- Композиции алгоритмов
  - Алгебраический подход к построению корректных алгоритмов
    - Области компетентности
    - Багинг - bagging
    - Бустинг - boosting

Центральная предельная теорема (теор.вер. :) !!!)

- Последовательности частичных средних, вычисленных по наборам из n независимых случайных величин, даже имеющих большую дисперсию σ, стремятся к нормальному распределению с дисперсией $D=D/\sqrt{n}$.

- Если брать среднее от значений прогнозов отдельных моделей, то неопределенность такого результата окажется ниже неопределенности отдельной модели.

Теорема Кондорсе о присяжных ?о

N - число членов жюри

p - вероятность правильного решения одного члена жюри

μ - вероятность правильного решения жюри

$\mu = \Sigma C_N^i \, p^i \, (1 - p)^{N-i}{}_{,}\ i=m,N$

Если $p > 0.5$, то $\mu > p$.

Если $N \to \infty$, то $\mu \to 1$.

C_N^i - комбинаций из N объектов по объектов i

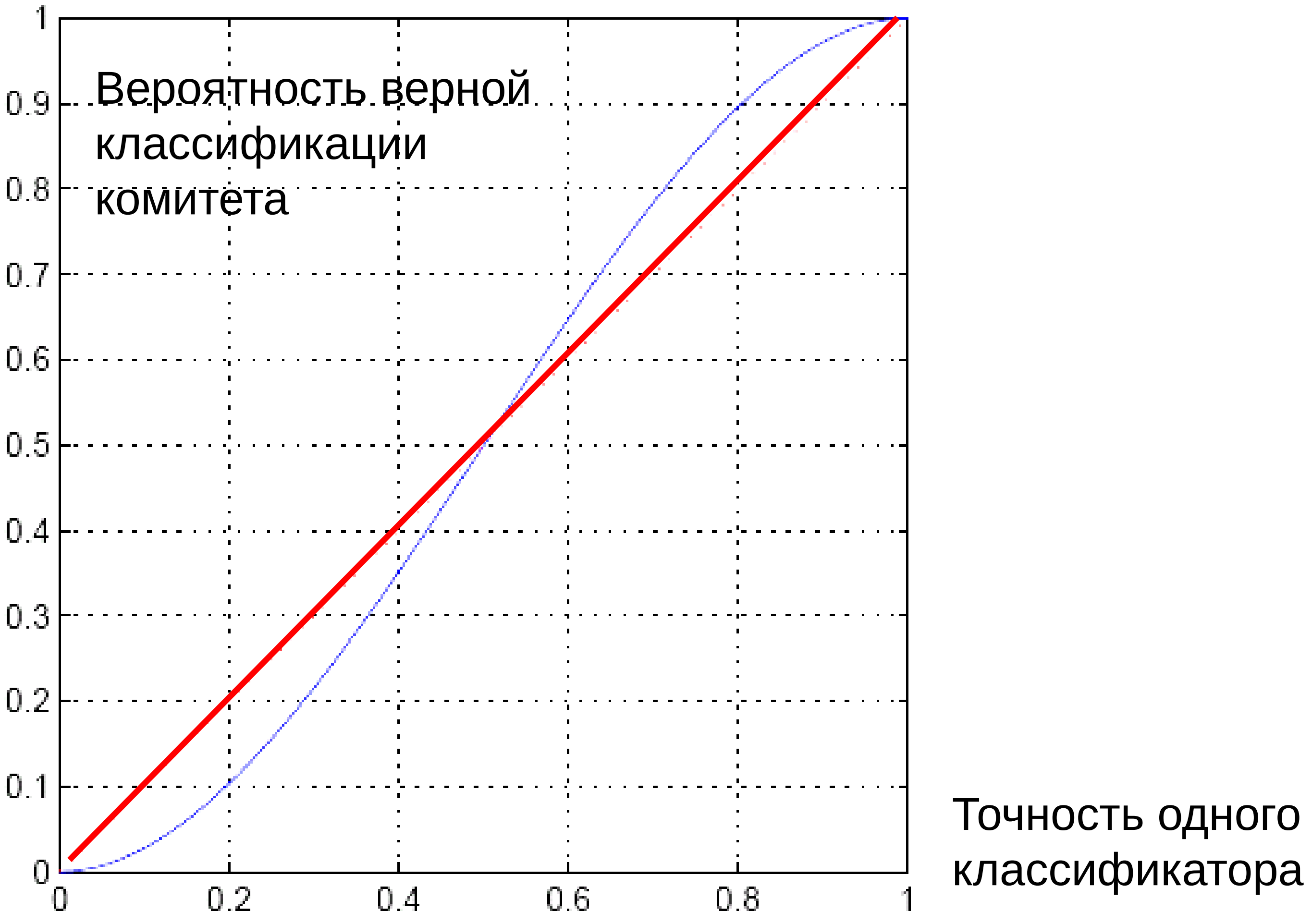
m — порог голосования

Теорема Кондорсе о присяжных

- Пусть есть три алгоритма A1, A2, A3
- Ai, решает определенную задачу бинарной классификации с вероятностью успеха **p**, независимо от остальных.
- Тогда при классификации примера X возможны 8 исходов:
  - все классификаторы выдали верный ответ, $\mathbf{p^3}$
  - два из трех не ошиблись (три варианта), $\mathbf{3p^2(1 - p)}$**,**
  - не ошибся лишь один (еще три варианта), $\mathbf{3p(1 - p)^2}$
  - ошиблись все три алгоритма одновременно, $\mathbf{(1 - p)^3}$.

- *комитет большинства*,
- вероятность благоприятного исхода (1 и 2 вариант)
- $q = p^3 + 3p^2(1 - p) = 3p^2 - 2p^3$.

Вероятность верной классификации комитетом



Усреднение по ансамблю

- Стратегия обучения
  - Уменьшение общей ошибки за счет варьирования начальных состояний.
  - Эксперты обучаются с избытком
  - Дисперсия уменьшается за счет усреднения

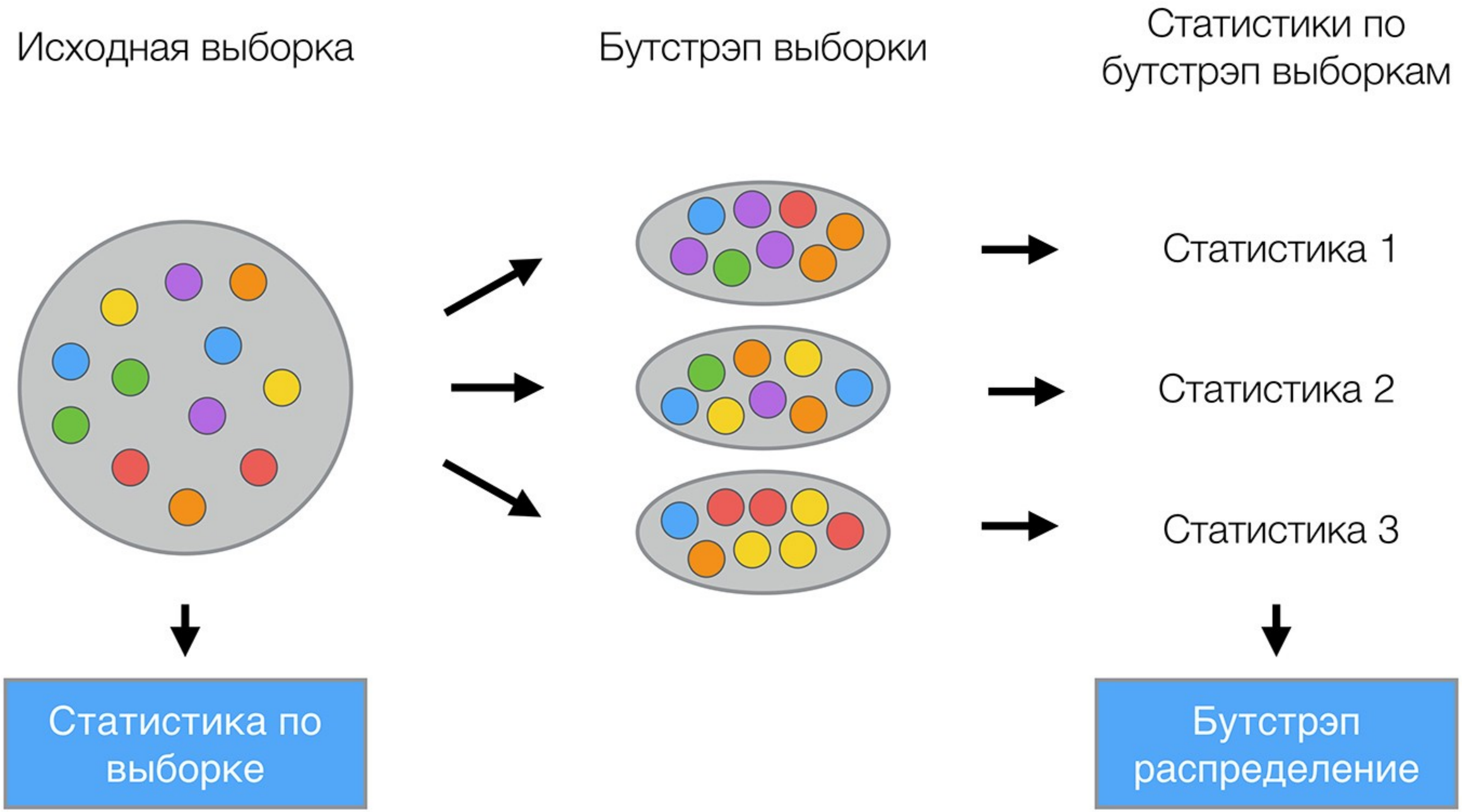
точность ансамбля моделей

- можно улучшить, если:

1) повысить точность каждой отдельной модели и, одновременно,

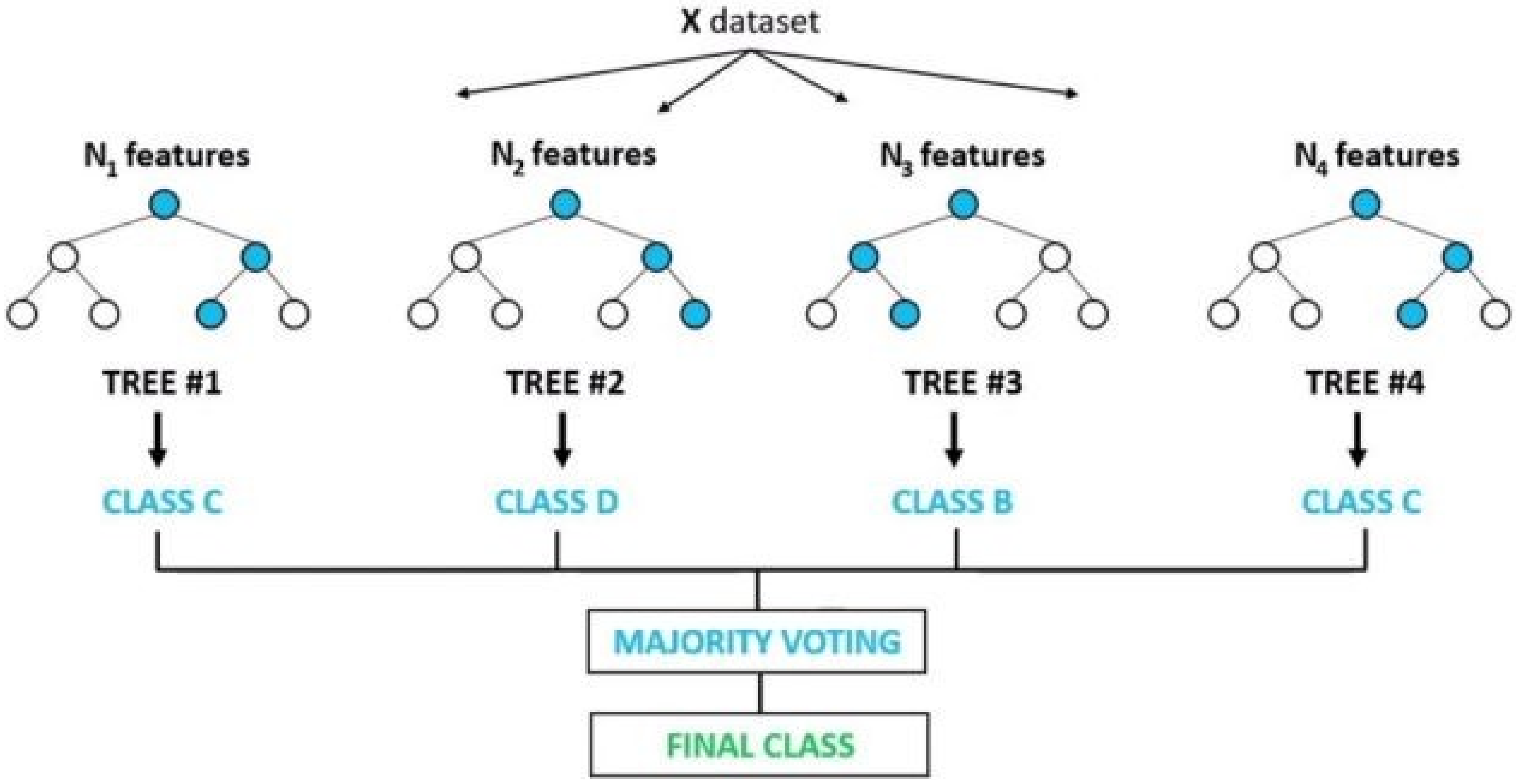
2) обеспечить статистическую независимость ошибок разных членов ансамбля.

Бэгинг

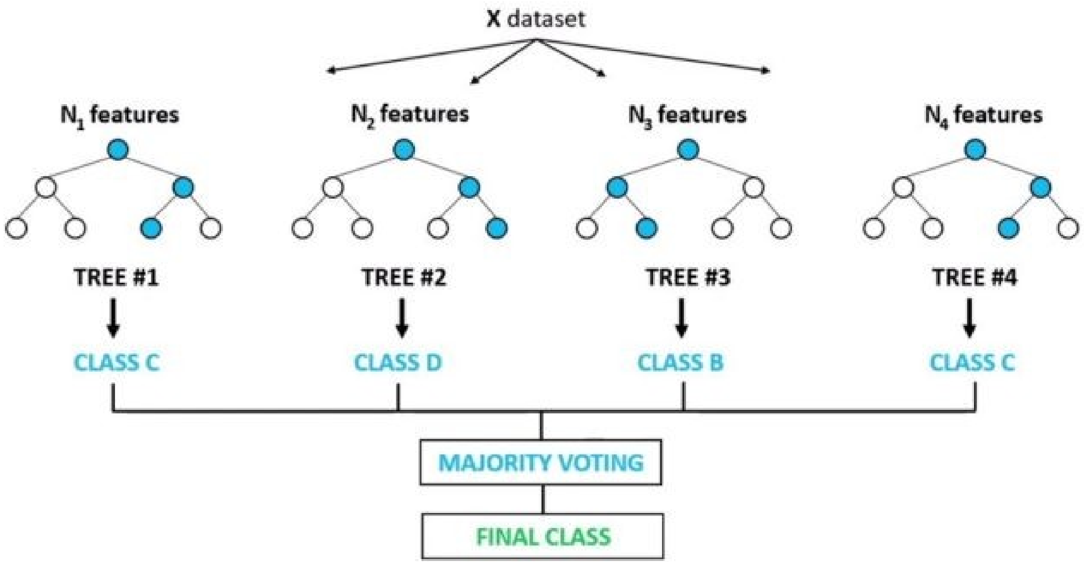


Исходная выборка
Бутстрэп выборки
Статистики по бутстрэп выборкам
Статистика 1
Статистика 2
Статистика 3
Статистика по выборке
Бутстрэп распределение

Ансамбль
Random Forest



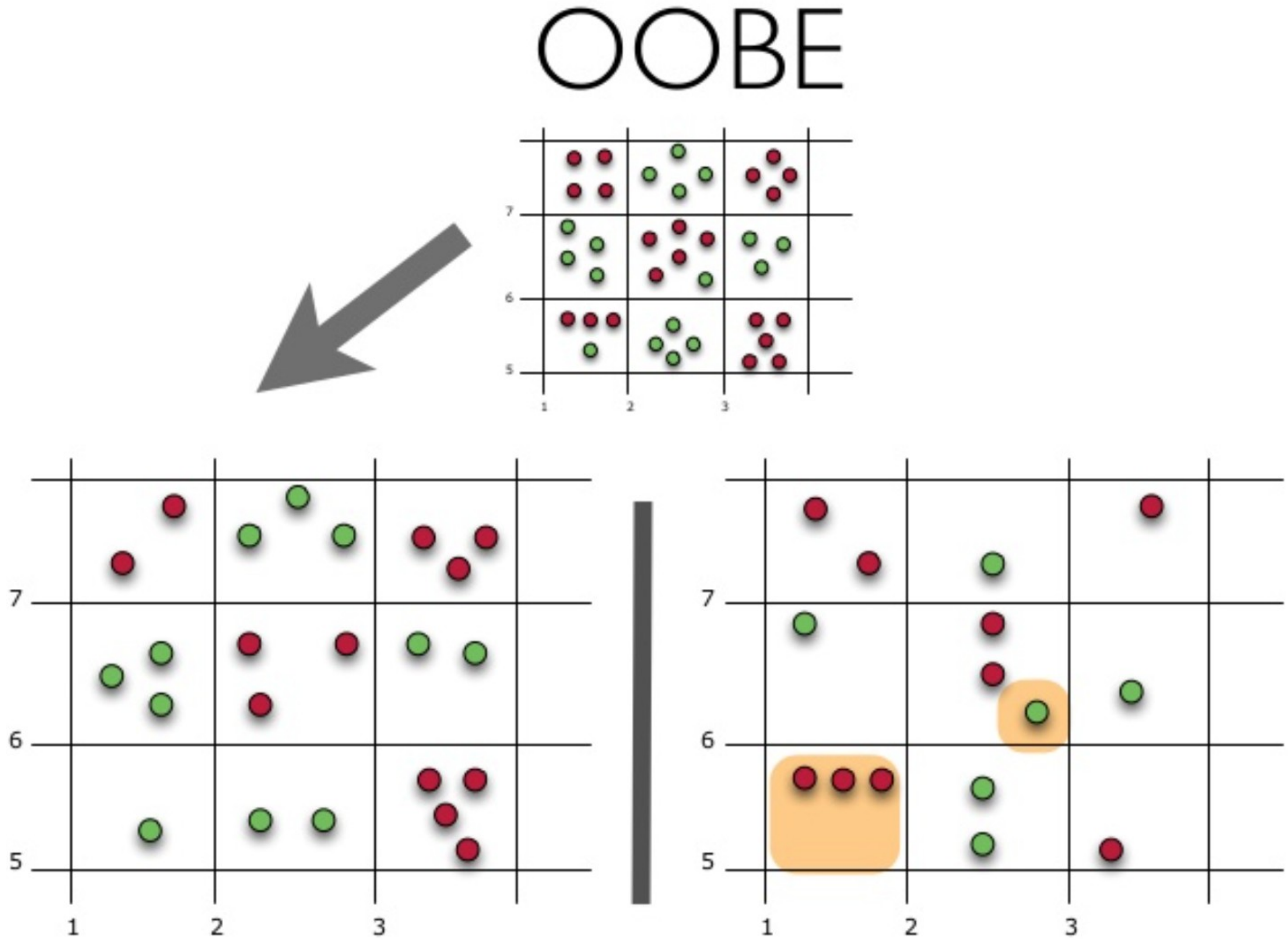
Ансамбль Random Forest



$$a_N(x) = \frac{1}{N} \sum_{n=1}^{N} b_n(x)$$

решающее дерево $b_n(x)$ по выборке $\tilde{X}_n$:

Out-Of-Bag-Error



Ансамбли Бустинг

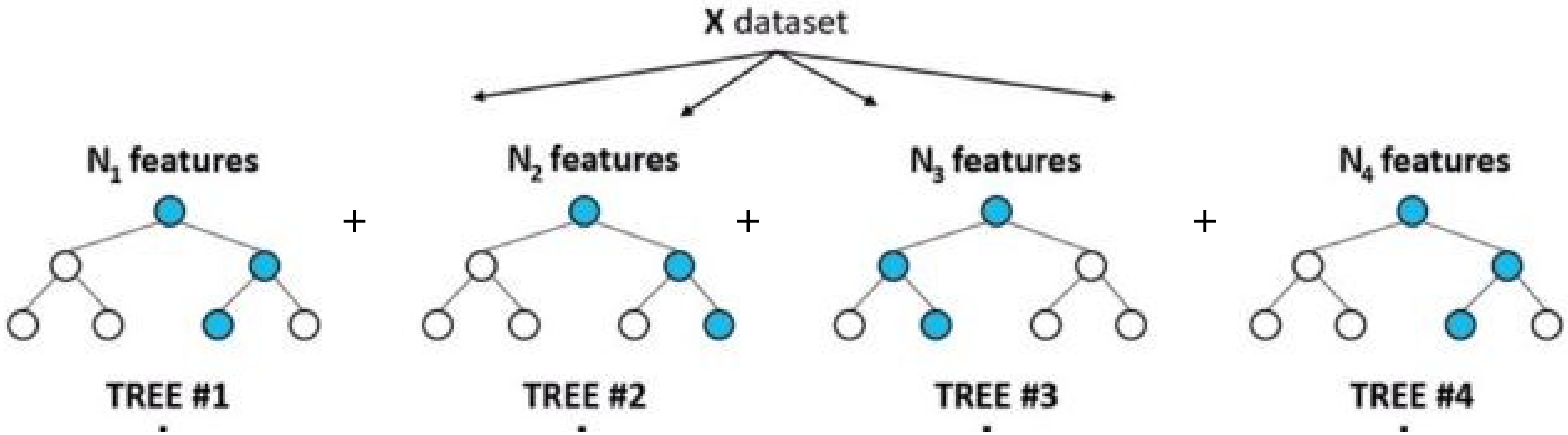
Корлякова Мария.

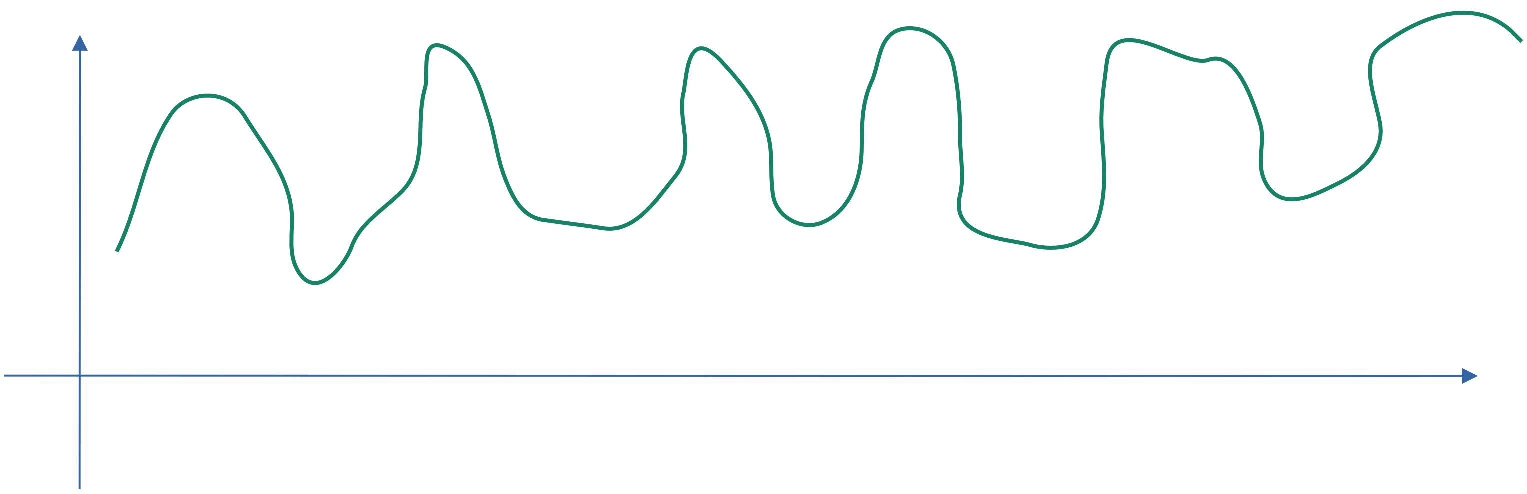
2022

Как решить дилему дисперсии-смещения:

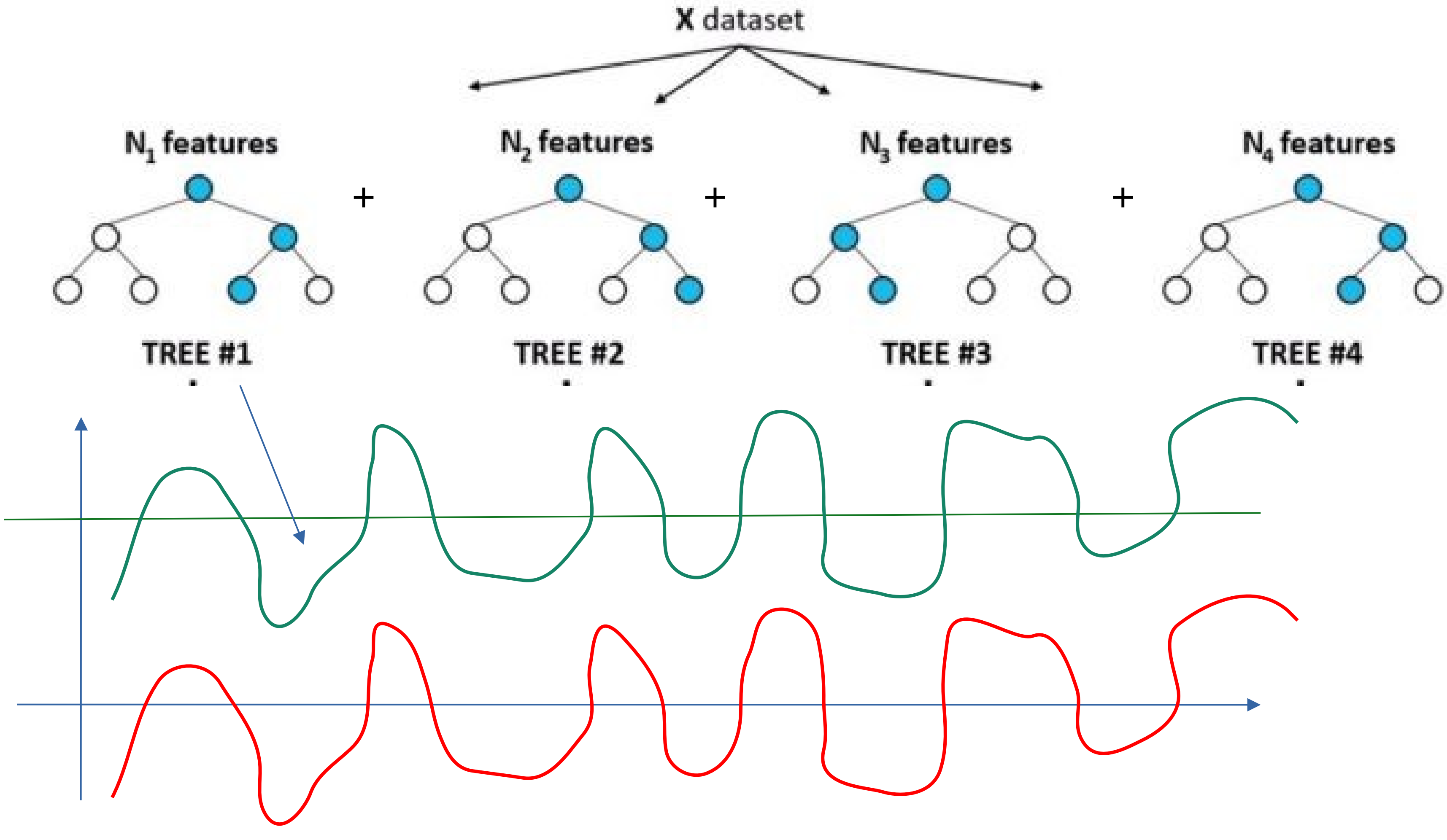
- Композиции алгоритмов
  - Алгебраический подход к построению корректных алгоритмов
    - Области компетентности
    - Багинг - bagging
    - Бустинг - boosting

Ансамбль усиления

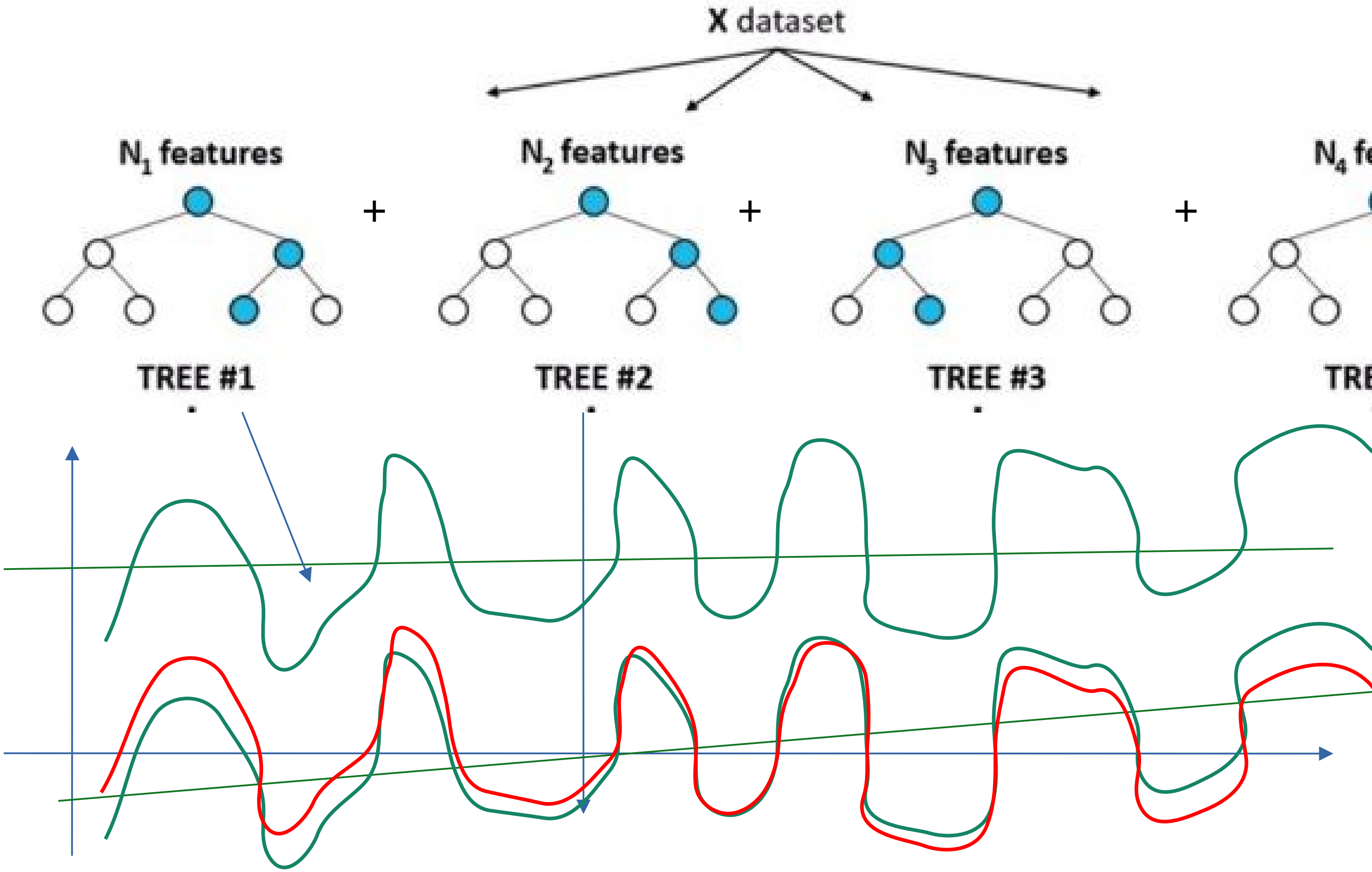




Ансамбль усиления



Ансамбль усиления



Ансамбль усиления Boosting

Потери:

$$\frac{1}{2}\sum_{i=1}^{\ell}(a(x_i)-y_i)^2 \to \min_a$$

Искомый алгоритм:

$$a_N(x)=\sum_{n=1}^{N} b_n(x),$$

b_n принадлежат некоторому семейству $\mathcal{A}$.

Ансамбль усиления Boosting

Первый шаг : $$b_1(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - y_i)^2$$

Остатки: $$s_i^{(1)} = y_i - b_1(x_i)$$

Новый шаг: $$b_2(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - s_i^{(1)})^2$$

Все следующие:

$$s_i^{(N)} = y_i - \sum_{n=1}^{N-1} b_n(x_i) = y_i - a_{N-1}(x_i), \qquad i = 1, \ldots, \ell;$$

$$b_N(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - s_i^{(N)})^2$$

Ансамбль усиления AdaBoost

Взвешивает примеры выборки X весами D:

1. D = 1/*l*

2. $$b_n = \operatorname*{argmin}_h \varepsilon_j$$

$$\varepsilon_j = \sum_{i=1}^{l} D_n(i)[y_i \neq b_j(x)]$$ пока не наступит $\varepsilon_j \geq 0.5.$

Алгоритм $$\alpha_n = \frac{1}{2}\ln\frac{1-\varepsilon_n}{\varepsilon_n}$$, Веса $$D_{n+1}(i) = \frac{D_n(i)e^{-\alpha_n y_i b_n(x_i)}}{Z_n}$$

3. $$a(x) = \operatorname{sign}\left(\sum_{n=1}^{N} \alpha_n b_n(x)\right)$$

Ансамбль усиления Boosting

Все следующие:

$$s_i^{(N)} = y_i - \sum_{n=1}^{N-1} b_n(x_i) = y_i - a_{N-1}(x_i), \qquad i = 1, \ldots, \ell;$$

$$b_N(x) := \arg\min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^{\ell} (b(x_i) - s_i^{(N)})^2$$

Остатки

$$s_i^{(N)} = y_i - a_{N-1}(x_i) = -\frac{\partial}{\partial z} \frac{1}{2}(z - y_i)^2 \Big|_{z=a_{N-1}(x_i)}$$

Ансамбль усиления Градиентный бустинг

Модель: $$a_N(x) = \sum_{n=1}^{N} \gamma_n b_n(x).$$

Инициализация b1:

- константа
- среднее (самое частое)

Цель: $$\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \to \min.$$

b1: $$b_1(x) = \operatorname{argmin} \frac{1}{l}\sum_{i=1}^{l}(b(x_i) - y_i)^2.$$

Остатки s1 $$s_i^{(1)} = y_i - b_1(x_i).$$

Ансамбль усиления
Градиентный бустинг

b2: $$b_2(x) = \underset{b}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} (b(x_i) - s_i^{(1)})^2 = \underset{b}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} (b(x_i) - (y_i - b_1(x_i)))^2.$$

bN: $$b_N(x) = \underset{b}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} (b(x_i) - s_i^{(N)})^2,$$

$$s_i^{(N)} = y_i - \sum_{n=1}^{N-1} b_n(x_i) = y_i - a_{N-1}(x_i).$$

Вектор сдвига sN: $$\sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + s_i) \to \min_{s}.$$

$$z = a_{N-1}(x_i) \qquad s_i = -\left.\frac{\partial L}{\partial z}\right|_{z=a_{N-1}(x_i)}.$$

Ансамбль усиления
Градиентный бустинг

Алгоритм bN:

$$b_N(x) = \underset{s}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} (b(x_i) - s_i)^2.$$

$$\gamma_N = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + \gamma b_N(x_i)).$$

$$a_N(x) = a_{N-1}(x) + \eta \gamma_N b_N(x).$$