

Лекция 10

Композиции

алгоритмов. Часть 1.

Юлия Конюшенко
ТГ: [@ko_iulia](https://t.me/ko_iulia)
koniushenko.iun@phystech.edu

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

- *Модель переобучена?*
- *Модель плохо предсказывает целевую переменную?*
- *В самих данных много неточностей (шумов)*

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

- $\text{Bias}(a(x))$ - средняя ошибка по всем возможным наборам данных – **смещение**.

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

- $\text{Bias}(a(x))$ - средняя ошибка по всем возможным наборам данных – **смещение**.

Смещение показывает, насколько в среднем модель хорошо предсказывает целевую переменную:

- ✓ *маленькое смещение - хорошее предсказание*
- ✓ *большое смещение – плохое предсказание*

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

- $\text{Var}(a(x))$ - дисперсия ошибки, т.е. как сильно различается ошибка при обучении на различных наборах данных – **разброс**.

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

- $\text{Var}(a(x))$ - дисперсия ошибки, т.е. как сильно различается ошибка при обучении на различных наборах данных – **разброс**.

Большой разброс означает, что ошибка очень чувствительна к изменению обучающей выборки, т.е.:

✓ *большой разброс – сильно переобученная модель*

РАЗЛОЖЕНИЕ ОШИБКИ (BIAS-VARIANCE DECOMPOSITION)

Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

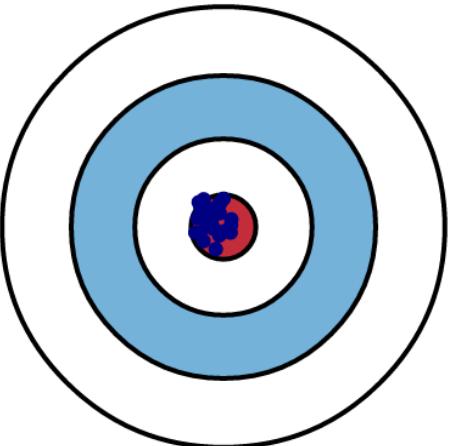
Утверждение (с док-вом): ошибку модели $a(x)$ можно представить в виде

$$\text{Err}(x) = \text{Bias}^2(a(x)) + \text{Var}(a(x)) + \sigma^2.$$

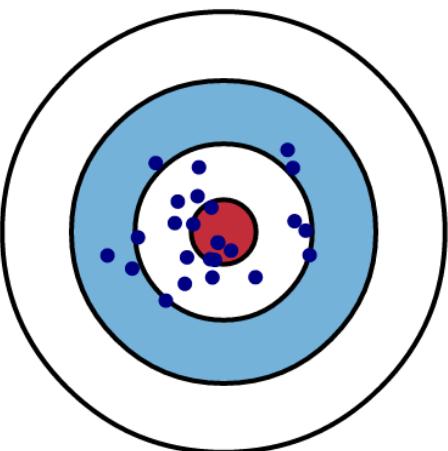
- $\text{Bias}(a(x))$ - средняя ошибка по всем возможным наборам данных – **смещение**.
- $\text{Var}(a(x))$ - дисперсия ошибки, т.е. как сильно различается ошибка при обучении на различных наборах данных – **разброс**.
- σ^2 - неустранимая ошибка – **шум**.

СМЕЩЕНИЕ И РАЗБРОС

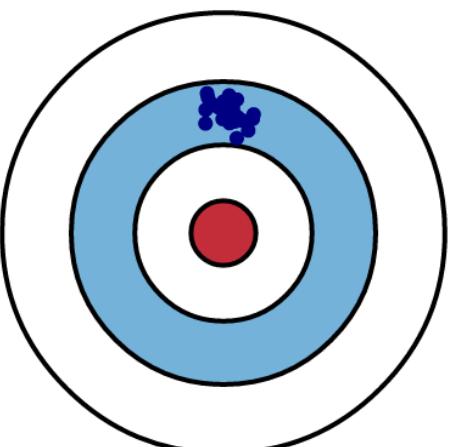
Low Variance



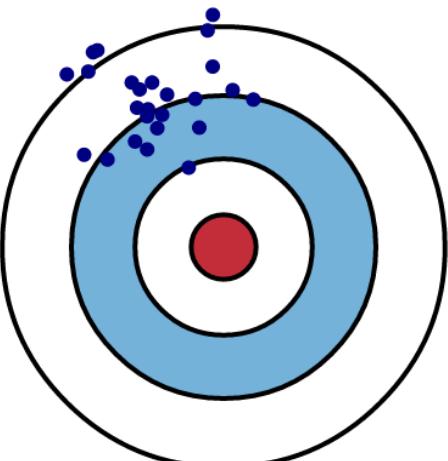
High Variance



Low Bias



High Bias



Что здесь

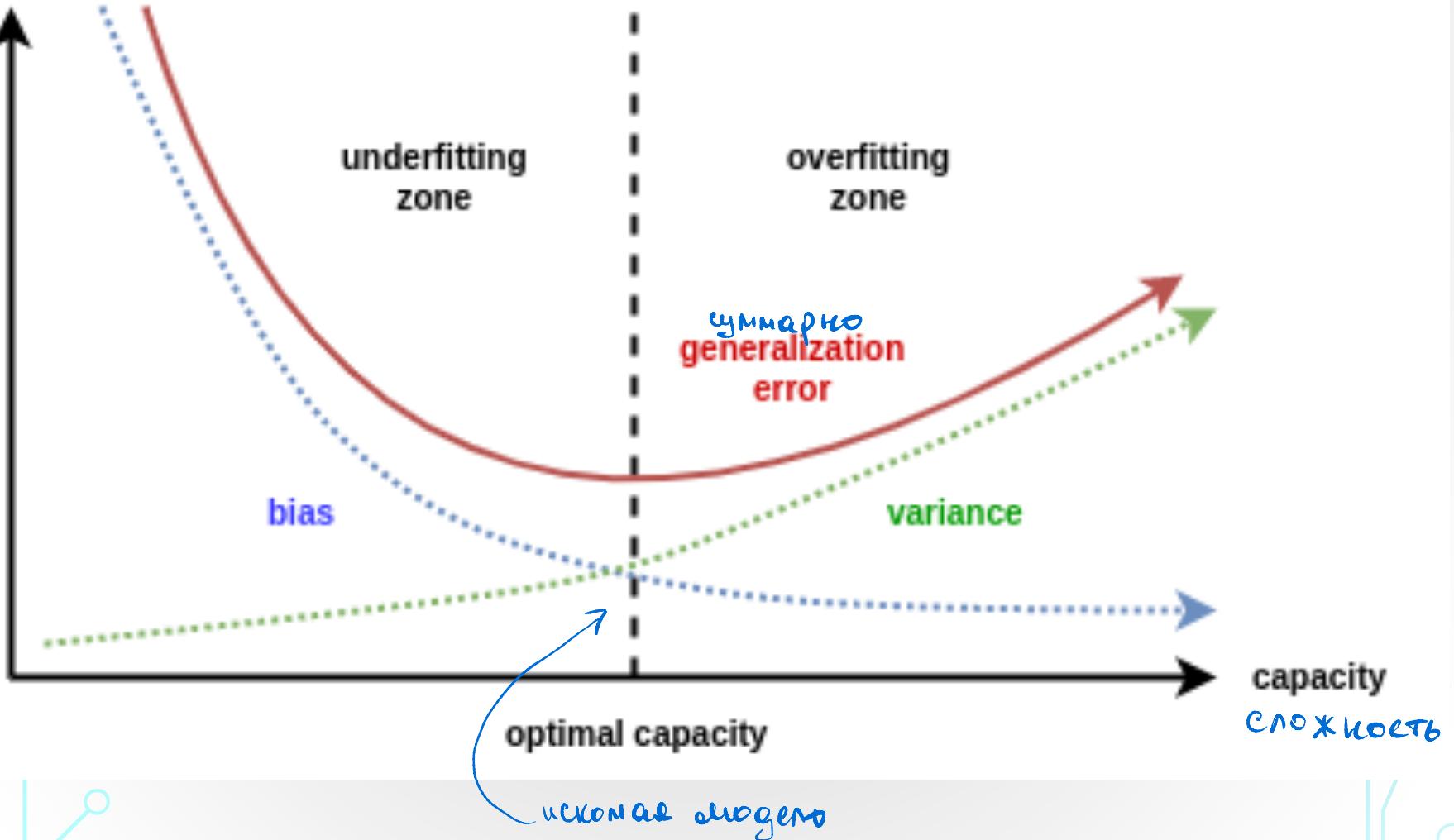
обозначено за одну точку?

Сложность модели - кол-во её весов (обучающих параметров)

У простых моделей большой bias, маленькая variance

Чем сложнее модель, тем меньше bias и больше variance

BIAS-VARIANCE TRADEOFF



Доказем теорему:

- математическая зависимость: $y = f(x) + \epsilon$, $\epsilon \sim N(0, \sigma^2)$
- модель: $x \mapsto a(x)$

Пусть решаем задачу регрессии с MSE.

Хотим понять какова средняя ошибка у нашей модели, то есть:

$$\mathbb{E} (y - a)^2 - ?$$

$$\mathbb{E} (y - a)^2 = \mathbb{E} (y^2 + a^2 - 2ya) = \underbrace{\mathbb{E} y^2 - (\mathbb{E} y)^2}_{\text{Dy}} + \underbrace{\mathbb{E} a^2 - (\mathbb{E} a)^2}_{\text{Da}} + \underbrace{(\mathbb{E} a)(\mathbb{E} y)}_{\text{no всем одинаковым объектам}}$$

$$- \mathbb{E} (2ya) =$$

$$= \text{Dy} + \text{Da} + ((\mathbb{E} y)^2 - \mathbb{E}(2ya) + (\mathbb{E} a)^2) =$$

$$= \text{Dy} + \text{Da} + (\mathbb{E} y - \mathbb{E} a)^2 =$$

m.k. $y = f(x) + \epsilon$, то $\mathbb{E} y = Ef + E\epsilon = f + 0$

m.k. const

$$= \mathbb{D}y + \mathbb{D}a + (f - Eq)^2$$

$$\begin{matrix} \parallel & \parallel & \parallel \\ \sigma^2 & Var & Bias^2 \end{matrix}$$

□

Лекциян $\mathbb{D}y$:

Пусть сеъ x_1, \dots, x_n огинақтөнде, нән y көз
разные y_1, \dots, y_n (напр. огинақтөнде
нараңжарын көзартау, нән разные стоимости)

Есеп $y_1 = \dots = y_n$, мән $\mathbb{D}y = 0$

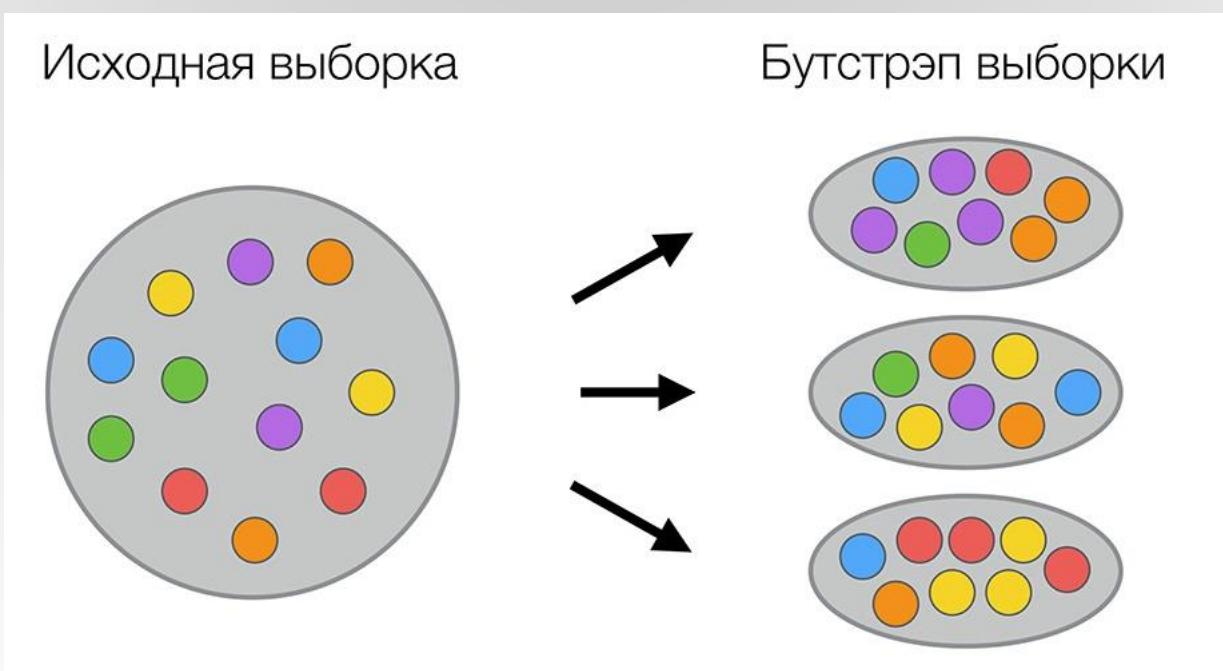
БУТСТРЭП

- метод генерации выборок

Дана выборка X .

Бутстрэп: равномерно возьмем из выборки X l объектов с возвращением (т.е. в новой выборке будут повторяющиеся объекты). Получим выборку X_1 .

- Повторяем процедуру N раз, получаем выборки X_1, \dots, X_N .



БЭГИНГ (BOOTSTRAP AGGREGATION)

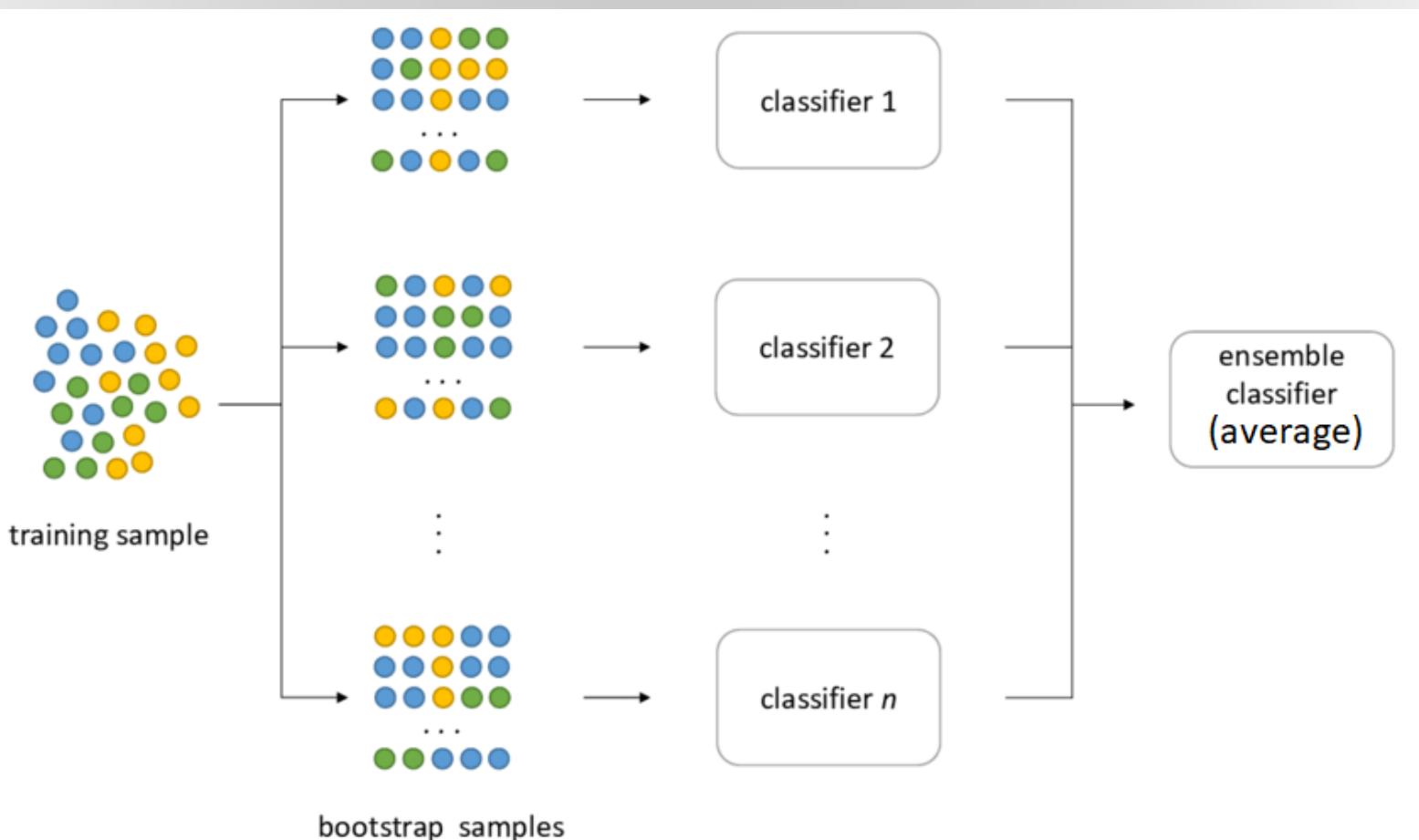
С помощью бутстрэпа мы получили выборки X_1, \dots, X_N .

- Обучим по каждой из них модель – получим базовые алгоритмы $b_1(x), \dots, b_N(x)$.
- Построим новую функцию регрессии:

$$a(x) = \frac{1}{N} \sum_{j=1}^N b_j(x)$$

БЭГГИНГ (BOOTSTRAP AGGREGATION)

$$a(x) = \frac{1}{N} \sum_{j=1}^N b_j(x)$$



СМЕЩЕНИЕ И РАЗБРОС У БЭГГИНГА

Бэггинг: $a_N(x) = \frac{1}{N} \sum_{n=1}^N b_n(x) = \frac{1}{N} \sum_{n=1}^N \tilde{\mu}(X)(x)$

(здесь $\tilde{\mu}(X) = \mu(\tilde{X})$ – алгоритм, обученный на подвыборке \tilde{X})

Утверждение (с док-вом):

- 1) *Бэггинг не ухудшает смещенность модели, т.е. смещение $a_N(x)$ равно смещению одного базового алгоритма.*
- 2) *Если базовые алгоритмы некоррелированы, то дисперсия бэггинга $a_N(x)$ в N раз меньше дисперсии отдельных базовых алгоритмов.*

Некоррелированные алгоритмы

b_1 и b_2 - алгоритмы

$$\text{test} \quad b_1(\text{test}) = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad b_2(\text{test}) = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix}$$

$\text{cor}(b_1, b_2) = ?$ even 0 , mo некоррел.

Докажем утверждение:

$$a(x) = \frac{1}{N} \sum_{i=1}^N M_i(x), \quad M_i - \text{один из алгоритмов}$$

- Следствие: $(\mathbb{E} a - f)^2 = (\mathbb{E} \left(\frac{1}{N} \sum_i M_i \right) - f)^2 = \left(\frac{1}{N} \sum_i \mathbb{E} M_i - f \right)^2 = (\mathbb{E} M - f)^2$, т.е. следствие комм-ии
работы нескольких одинаковых алгоритмов

- Рассмотрим: $\text{Var } a \stackrel{\text{def}}{=} \mathbb{E} (a - \mathbb{E} a)^2$

$$(a - \mathbb{E} a)^2 = \left(\frac{1}{N} \sum_i M_i - \mathbb{E} \left(\frac{1}{N} \sum_i M_i \right) \right)^2 = \frac{1}{N^2} \left(\sum_i [M_i - \mathbb{E} M_i]^2 \right)$$

gne gelyx ceraenix

$$((M_1 - \mathbb{E} M_1) + (M_2 - \mathbb{E} M_2))^2 = (M_1 - \mathbb{E} M_1)^2 + (M_2 - \mathbb{E} M_2)^2 + 2(M_1 - \mathbb{E} M_1)(M_2 - \mathbb{E} M_2)$$

$$\Leftrightarrow \frac{1}{N^2} \sum_i (M_{i1} - \mathbb{E} M_{i1})^2 + \frac{1}{N^2} \sum_{i_1 \neq i_2} (M_{i1} - \mathbb{E} M_{i1})(M_{i2} - \mathbb{E} M_{i2})$$

Ochmaoek neeruiaar aiat. ox. om emao bipaxenee

$$\mathbb{E} (a - \mathbb{E} a)^2 = \frac{1}{N^2} \sum_i \mathbb{E} (M_{i1} - \mathbb{E} M_{i1})^2 + \frac{1}{N^2} \sum_i \mathbb{E} (M_{i1} - \mathbb{E} M_{i1}) \times$$

~~$\times (M_{i2} - \mathbb{E} M_{i2})$~~

$\underbrace{\hspace{10em}}$ m. K $\underbrace{\hspace{10em}}$ bokopp. ait, to
 $\underbrace{\hspace{10em}}$ covar = 0.

$$= \frac{1}{N^2} \cdot N \mathbb{E} (M - \mathbb{E} M)^2 = \frac{1}{N} \text{Var } M$$

■

Как добиться некоррелир алгоритмов?

- бутстреп

- случай. лес (обучение на подвыборке признаков)

СЛУЧАЙНЫЙ ЛЕС (RANDOM FOREST)

- Возьмем в качестве базовых алгоритмов для бэггинга **решающие деревья**, т.е. каждое случайное дерево $b_i(x)$ построено по своей подвыборке X_i .
- В каждой вершине дерева будем искать **разбиение не по всем признакам, а по подмножеству признаков**.
- Дерево строится до тех пор, пока в листе не окажется n_{min} объектов.



RANDOM FOREST

Алгоритм 3.1. Random Forest

1: для $n = 1, \dots, N$

2: Сгенерировать выборку \tilde{X}_n с помощью бутстрэпа

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

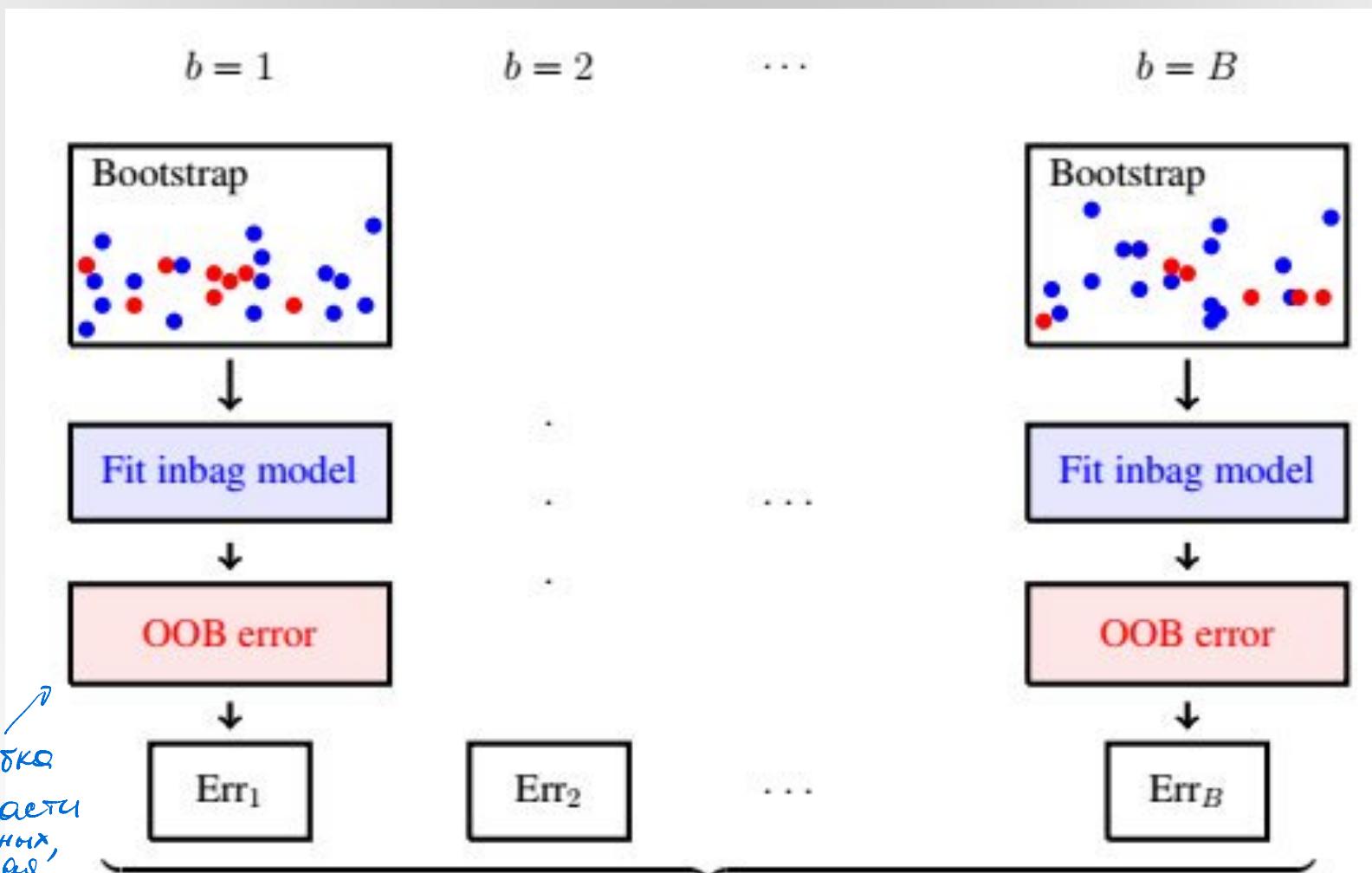
- дерево строится, пока в каждом листе не окажется не более n_{\min} объектов
- при каждом разбиении сначала выбирается t случайных признаков из p , и оптимальное разделение ищется только среди них

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

RANDOM FOREST – ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ

- Если p – количество признаков, то при классификации обычно берут $m = \lceil \sqrt{p} \rceil$, а при регрессии - $m = \lceil \frac{p}{3} \rceil$ признаков
- При классификации обычно дерево строится, пока в листе не окажется $n_{min} = 1$ объект, а при регрессии $n_{min} = 5$

OUT-OF-BAG ОШИБКА



$$\text{Err}_{\text{oob}} = \frac{\text{Err}_1 + \dots + \text{Err}_B}{B} = \frac{1}{B} \sum_{b=1}^B \text{Err}_b$$

OUT-OF-BAG ОШИБКА

- Каждое дерево в случайном лесе обучается по некоторому подмножеству объектов
- Значит, для каждого объекта есть деревья, которые на этом объекте не обучались.

Out-of-bag ошибка:

$$OOB = \sum_{i=1}^l L(y_i, \frac{\sum_{n=1}^N [x_i \notin X_n] b_n(x_i)}{\sum_{n=1}^N [x_i \notin X_n]})$$

Утверждение. При $N \rightarrow \infty$ OOB оценка стремится к leave-one-out оценке.

OOB-SCORE

По графику out-of-bag ошибки можно, например, подбирать количество деревьев в случайном лесе

