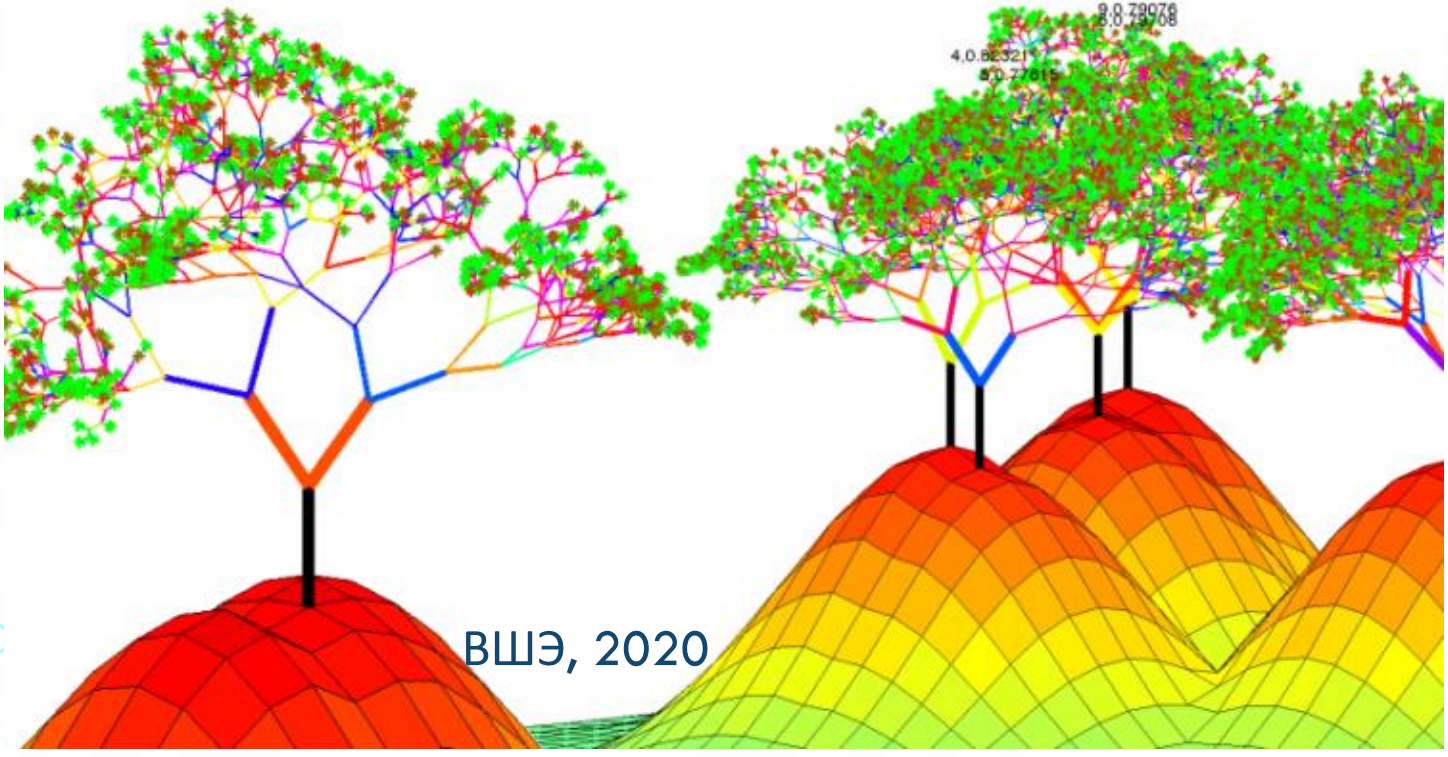
Лекция 10
Композиции алгоритмов. Часть 1.

Кантонистова Е.О.



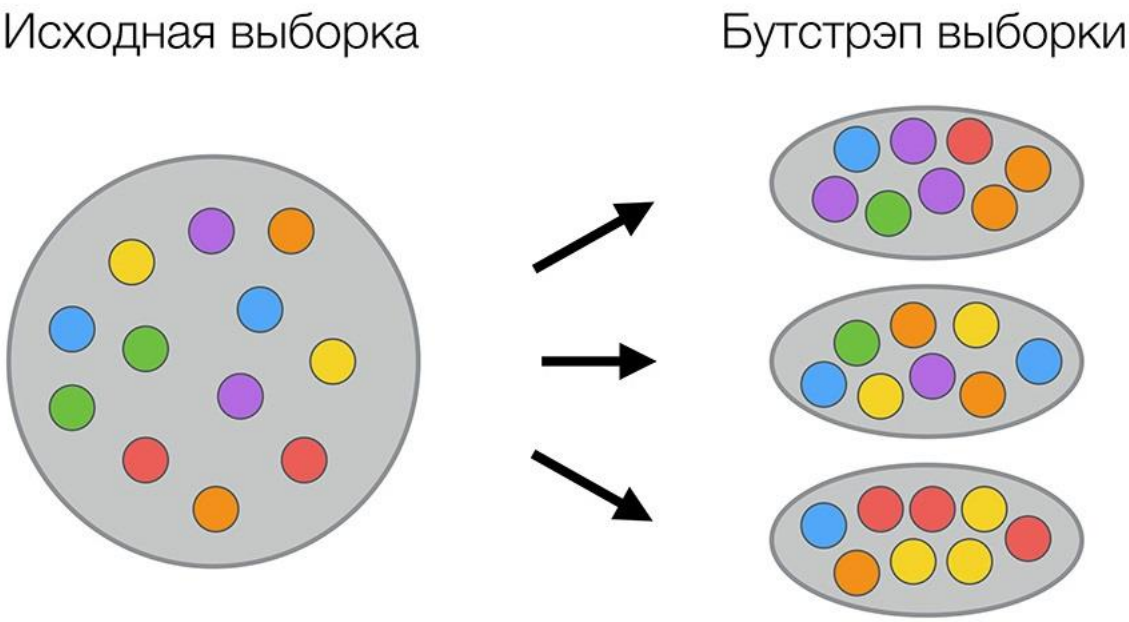
ВШЭ, 2020

БУТСТРЭП

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 $X_1, \ldots, X_N$.



БЭГГИНГ (BOOTSTRAP AGGREGATION)

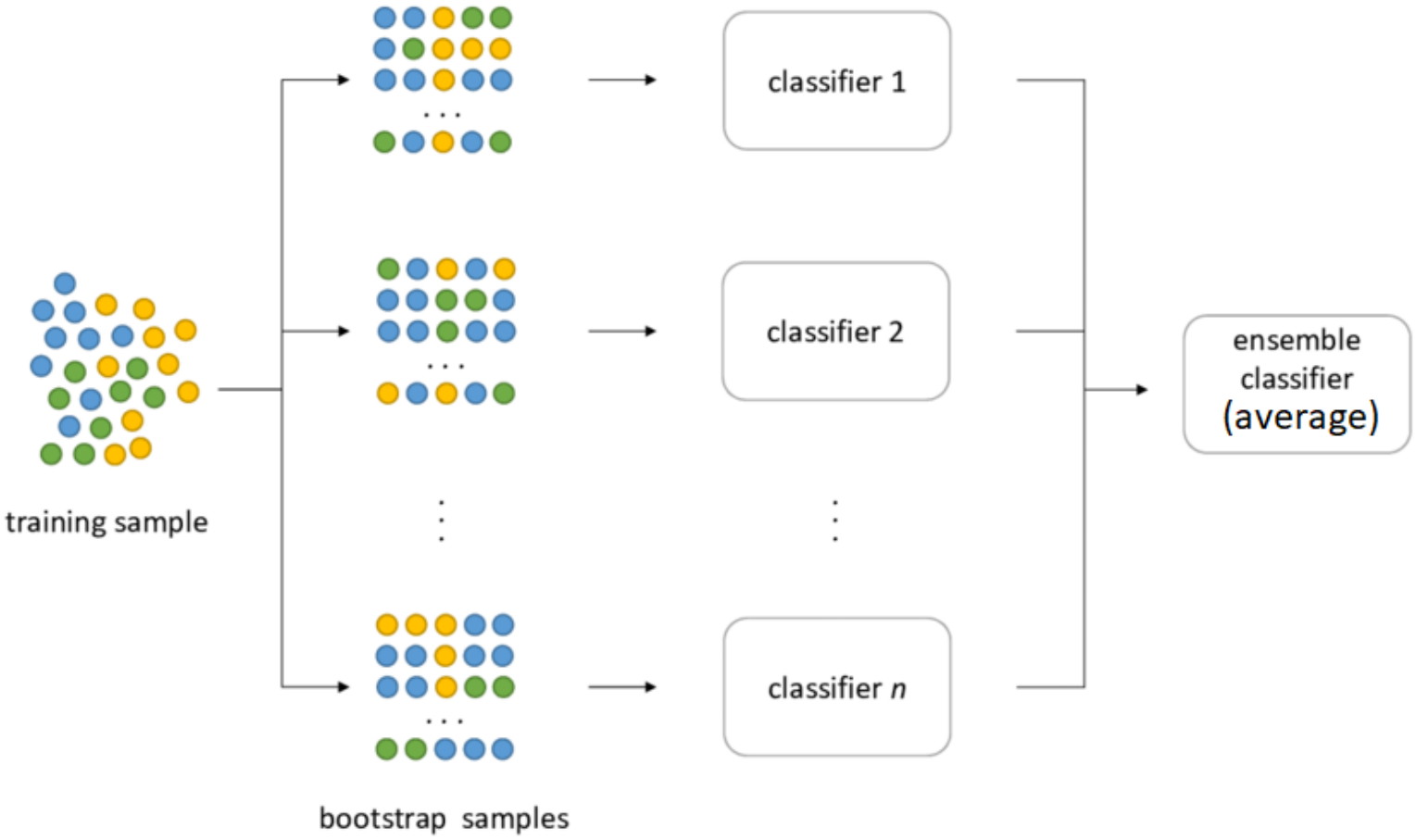
С помощью бутстрэпа мы получили выборки $X_1, \dots, X_N$.

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

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

БЭГГИНГ (BOOTSTRAP AGGREGATION)

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



БЭГГИНГ (BOOTSTRAP AGGREGATION)

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

Утверждение. *Если алгоритмы $b_1(x), \ldots, b_N(x)$ в задаче регрессии некоррелированы, то среднеквадратичная ошибка алгоритма $a(x)$, полученного при помощи бэггинга, в N раз меньше среднеквадратичной ошибки исходных алгоритмов $b_j(x)$.*

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*

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

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*

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

- $\mathbf{Bias\big(a(x)\big)}$ - 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*

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

- $\mathbf{Bias\big(a(x)\big)}$ - 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*

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

- $\mathbf{Var\big(a(x)\big)}$ - 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*

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

- $\mathbf{Var\big(a(x)\big)}$ - 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)

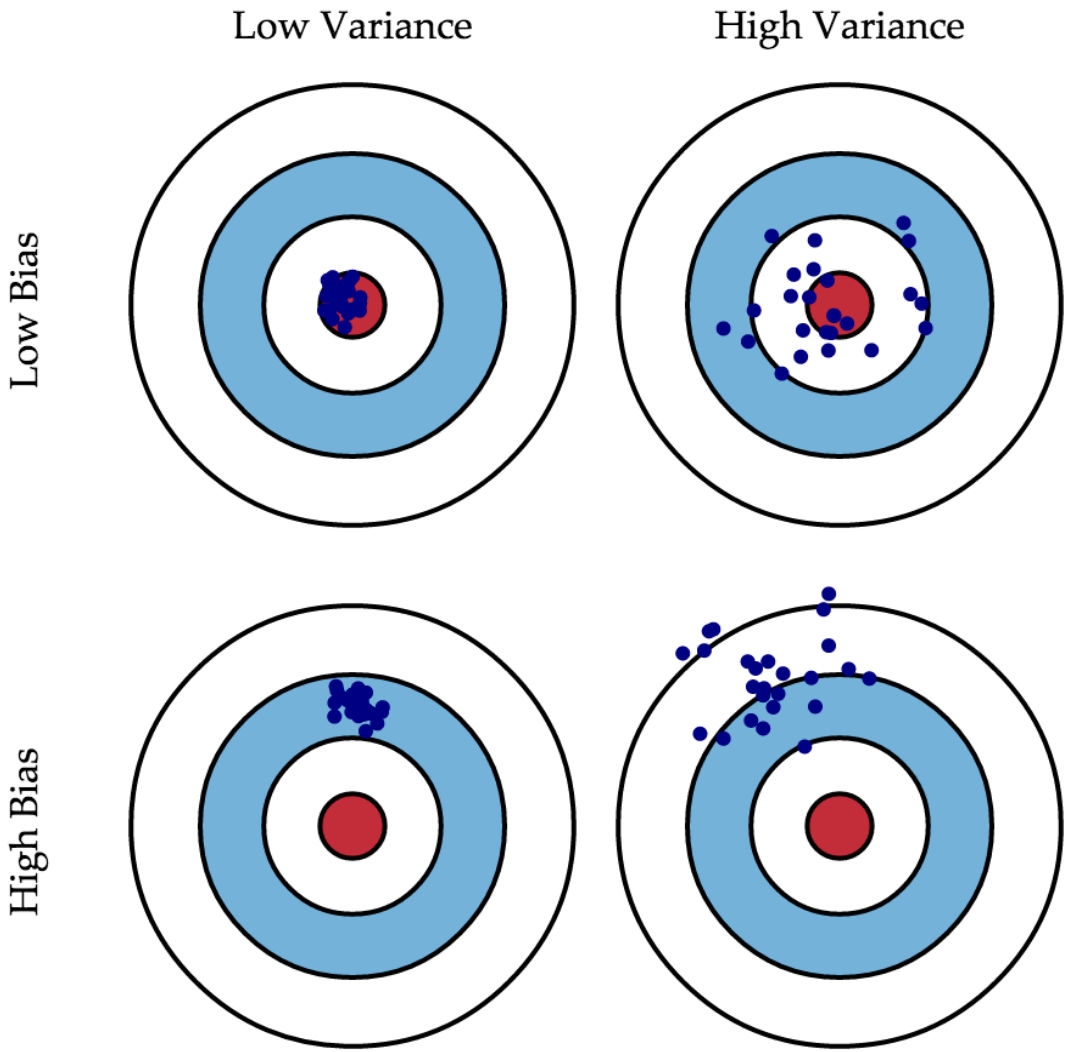
Зачастую для улучшения качества модели необходимо понять, из-за чего возникает ошибка в предсказаниях.

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

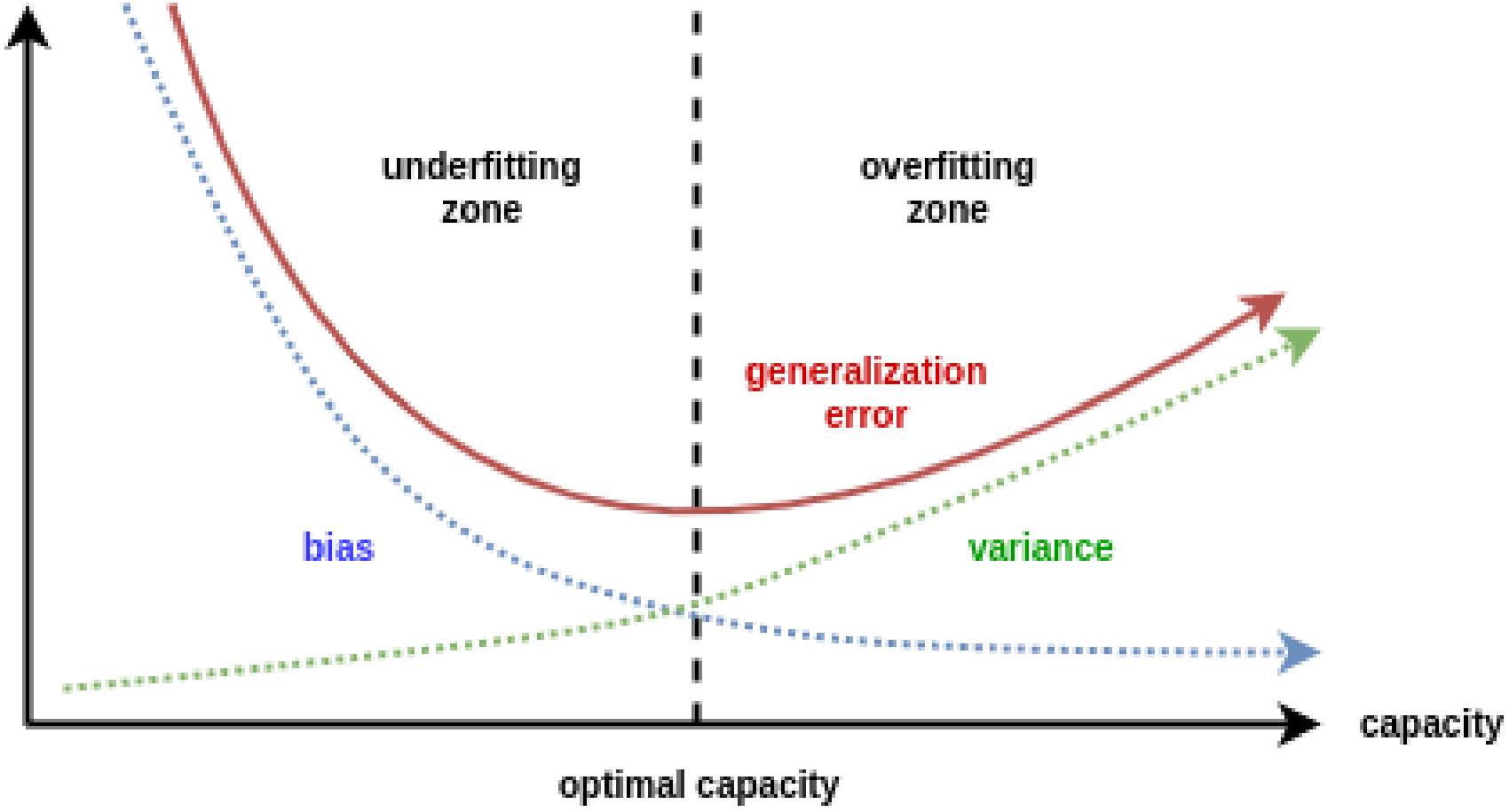
$$\mathbf{Err(x) = Bias^2\big(a(x)\big) + Var\big(a(x)\big) + \sigma^2.}$$

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

СМЕЩЕНИЕ И РАЗБРОС



BIAS-VARIANCE TRADEOFF



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 $\tilde{X}$)

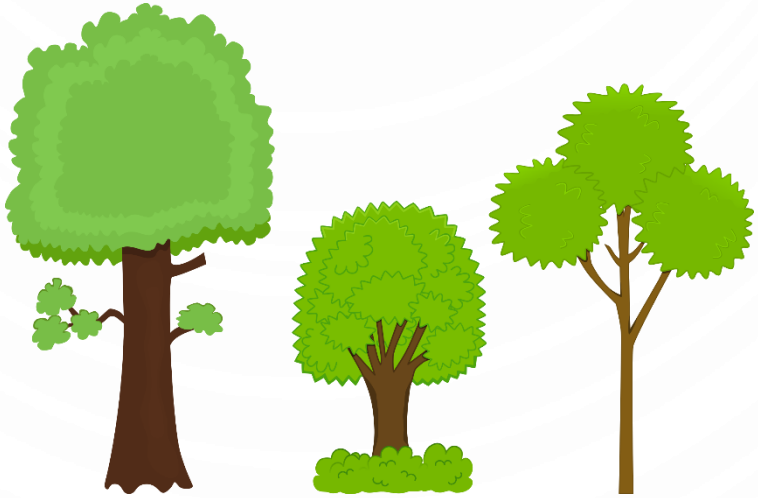
Утверждение (с док-вом):

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.***

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** $n = 1, \ldots, N$

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

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

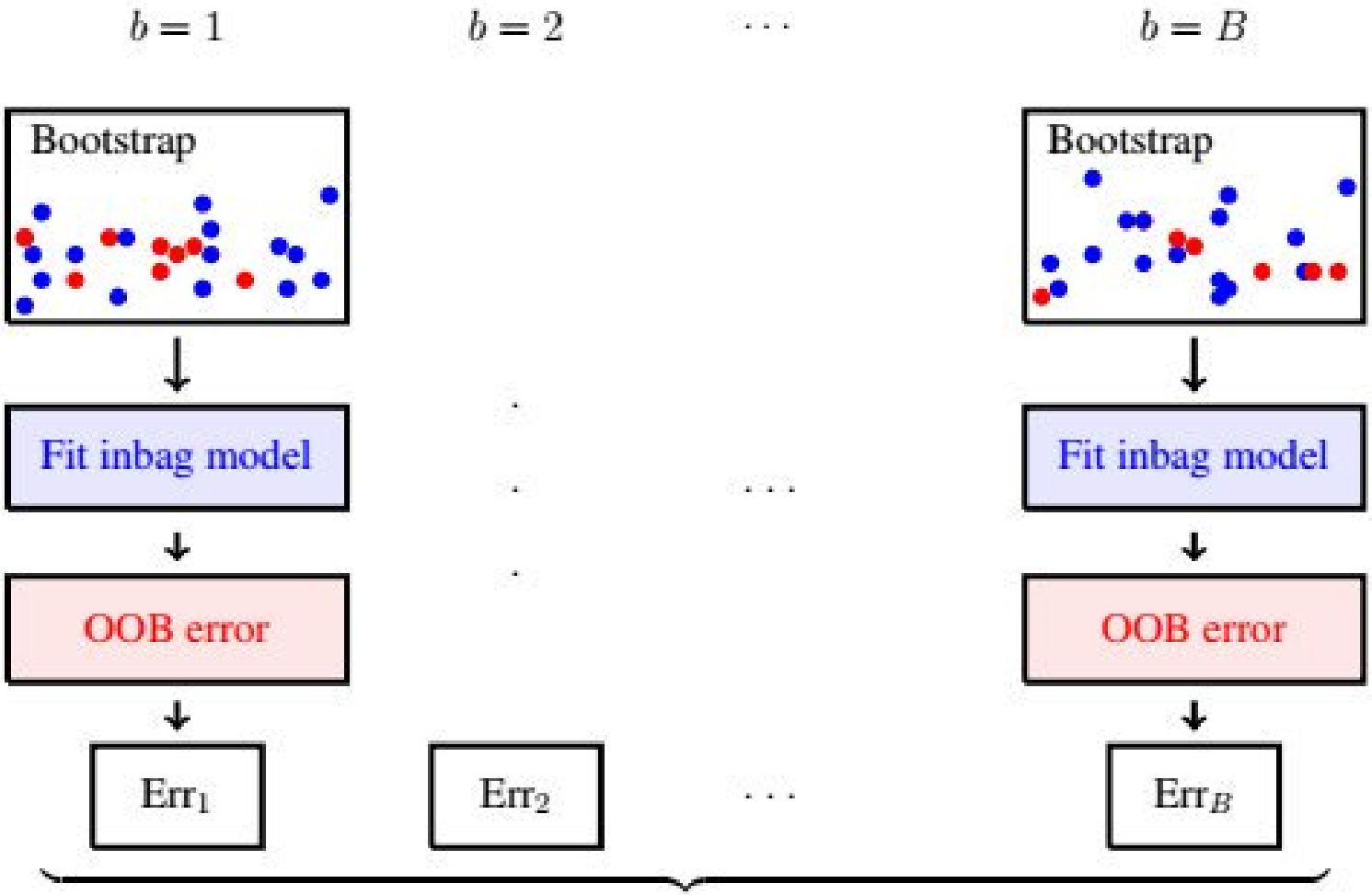
- дерево строится, пока в каждом листе не окажется не более $n_{\min}$ объектов
- при каждом разбиении сначала выбирается m случайных признаков из p, и оптимальное разделение ищется только среди них

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 $m = [\sqrt{p}]$, а при регрессии - $m = [\frac{p}{3}]$ 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



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

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.

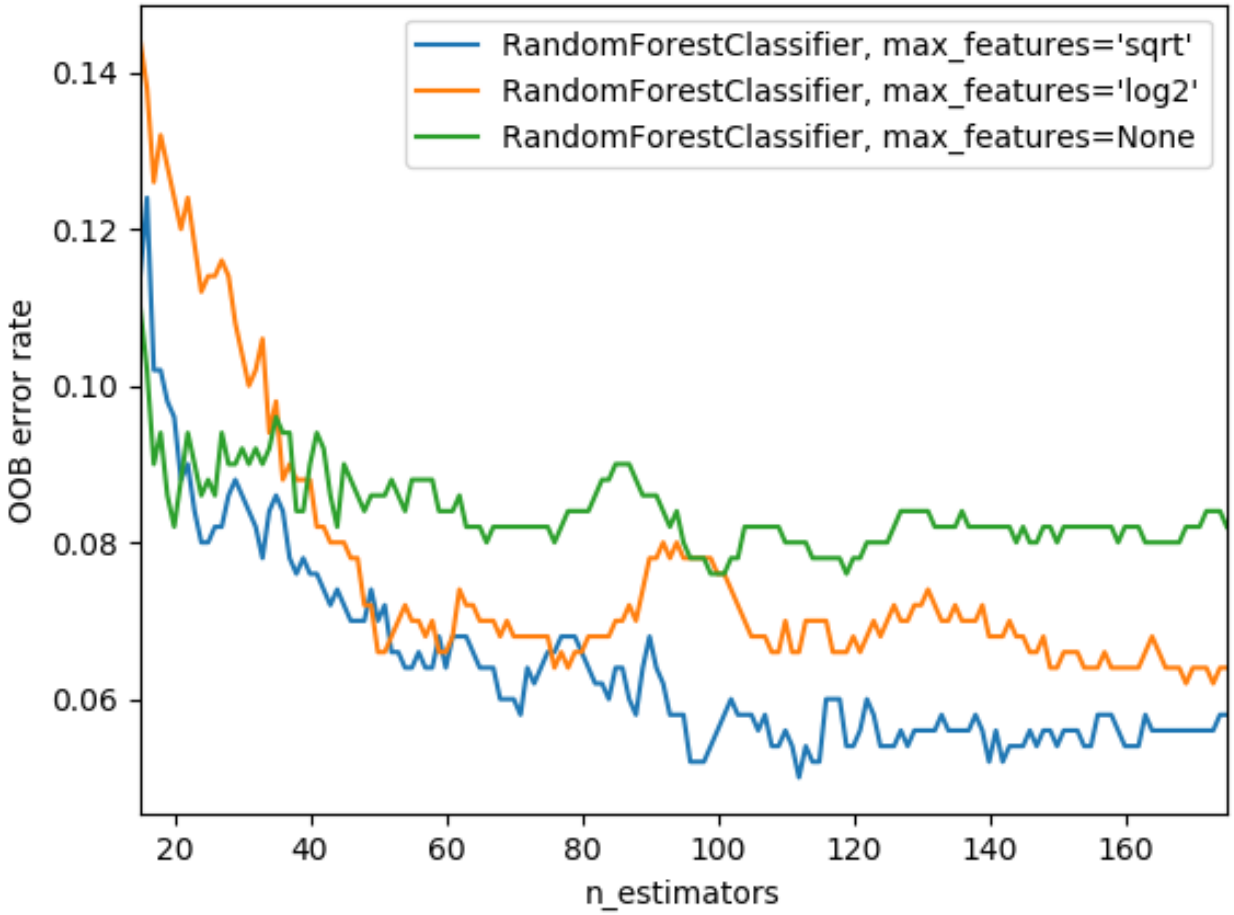
Out-of-bag ошибка:

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

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

OOB-SCORE

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



СРЕДНЕКВАДРАТИЧНЫЙ РИСК

Цель: хотим (хотя бы в теории) измерить качество модели на всех объектах, а не только на обучающей выборке.

СРЕДНЕКВАДРАТИЧНЫЙ РИСК

Цель: хотим (хотя бы в теории) измерить качество модели на всех объектах, а не только на обучающей выборке.

Определим ***среднеквадратичный риск алгоритма*** $\boldsymbol{a(x)}$, соответствующий квадратичной функции потерь:

$$R(a) = \mathrm{E}_{x,y}\left[\left(y - a(x)\right)^2\right] = \int_X \int_Y p(x,y)\left(y - a(x)\right)^2 dxdy$$

СРЕДНЕКВАДРАТИЧНЫЙ РИСК

Цель: хотим (хотя бы в теории) измерить качество модели на всех объектах, а не только на обучающей выборке.

Определим ***среднеквадратичный риск алгоритма*** $a(x)$, соответствующий квадратичной функции потерь:

$$R(a) = \mathrm{E}_{x,y}\left[\left(y - a(x)\right)^2\right] = \int_X \int_Y p(x,y)\left(y - a(x)\right)^2 dxdy$$

- Функционал усредняет ответы модели в каждой точке x пространства объектов и по каждому возможному ответу y
- Вклад пары (x, y) пропорционален вероятности получить ее в выборке $p(x, y)$.

То есть среднеквадратичный риск позволяет измерить качество модели на всех возможных объектах.

МИНИМУМ СРЕДНЕКВАДРАТИЧНОГО РИСКА

Что предскажет идеальный алгоритм $a_*(x)$ при минимизации среднеквадратичного риска?

Теорема (с док-вом):

$$\boldsymbol{a_*(x) = E[y|x].}$$

То есть алгоритм на объектах с признаковым описанием x предсказывает средневзвешенный ответ y с весами, равными вероятности встретить ответ на данном объекте.