ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

Лекция №7

01

Композиции моделей

Мудрость толпы

- Какой из этих замечательных пределов является первым?

$$\lim_{x \to inf} \frac{sin\, x}{x} \qquad \lim_{x \to inf} \left(\frac{x}{1+x}\right)^{x}$$

- Усредним ваши ответы

Мудрость толпы

- Какое расстояние от Земли до Луны

- Усредним ваши ответы

Композиция

Аналогично с моделями машинного обучения:

- Если мы возьмем N базовых алгоритмов $b_1(x),\ ...,\ b_N(x)$
- То композиция будет выглядеть следующим образом

$$a(x) = arg\,max_{y \epsilon Y} \sum_{n=1}^{N} \left[b_n(x) = y \right]$$

Композиция

Аналогично с моделями машинного обучения:

- Если мы возьмем N базовых алгоритмов $b_1(x),\ ...,\ b_N(x)$
- То композиция будет выглядеть следующим образом

$$a(x) = arg\,max_{y \in Y} \sum_{n=1}^{N} \left[b_n(x) = y \right]$$

- Либо же для регрессии

$$a(x) = \frac{1}{N} \sum_{n=1}^{N} \left(b_n(x) \right)$$

Композиция

- Если каждая базовая модель хоть немного лучше угадывания, то рост количества моделей устремляет ответ к истинному

Композиция

- Если каждая базовая модель хоть немного лучше угадывания, то рост количества моделей устремляет ответ к истинному
- Теорема о “жюри присяжных”:

N — количество присяжных
p — вероятность правильного решения присяжного
μ — вероятность правильного решения всего жюри
m — минимальное большинство членов жюри, $m = \text{floor}(N/2) + 1$
C_N^i — число сочетаний из N по i

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

Если $p > 0.5$, то $\mu > p$
Если $N \to \infty$, то $\mu \to 1$

02

Случайный лес

Бэггинг

- $b_1(x), \dots, b_N(x)$ — базовые модели
- Как на одной выборке построить N различных моделей?

Бэггинг

- $b_1(x), \ldots, b_N(x)$ — базовые модели
- Как на одной выборке построить N различных моделей?

Обучим их независимо на разных подвыборках

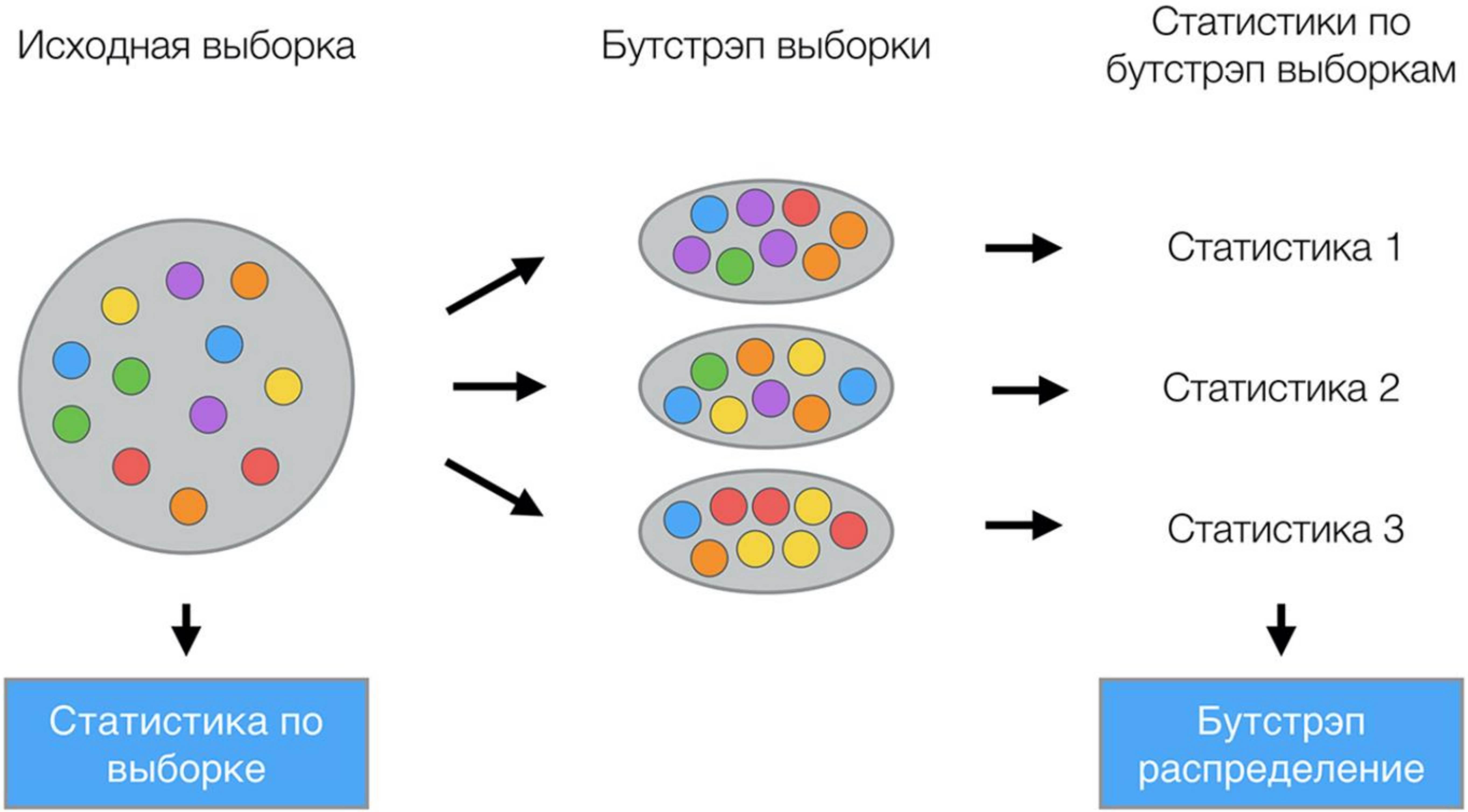
Бэггинг

- Bagging (bootstrap aggregating)
- Базовые модели обучаются независимо
- Подмножество выбирается с помощью бутстрапа

Бутстрап

- Выборка с возвращением
- Берём ℓ элементов из X
- Пример: $\{x_1, x_2, x_3, x_4\} \to \{x_1, x_2, x_2, x_4\}$
- В подвыборке будет ℓ объектов, из них около 63.2% уникальных
- Если объект входит в выборку несколько раз, то мы как бы повышаем его вес

Бутстрап



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

Случайные подпространства

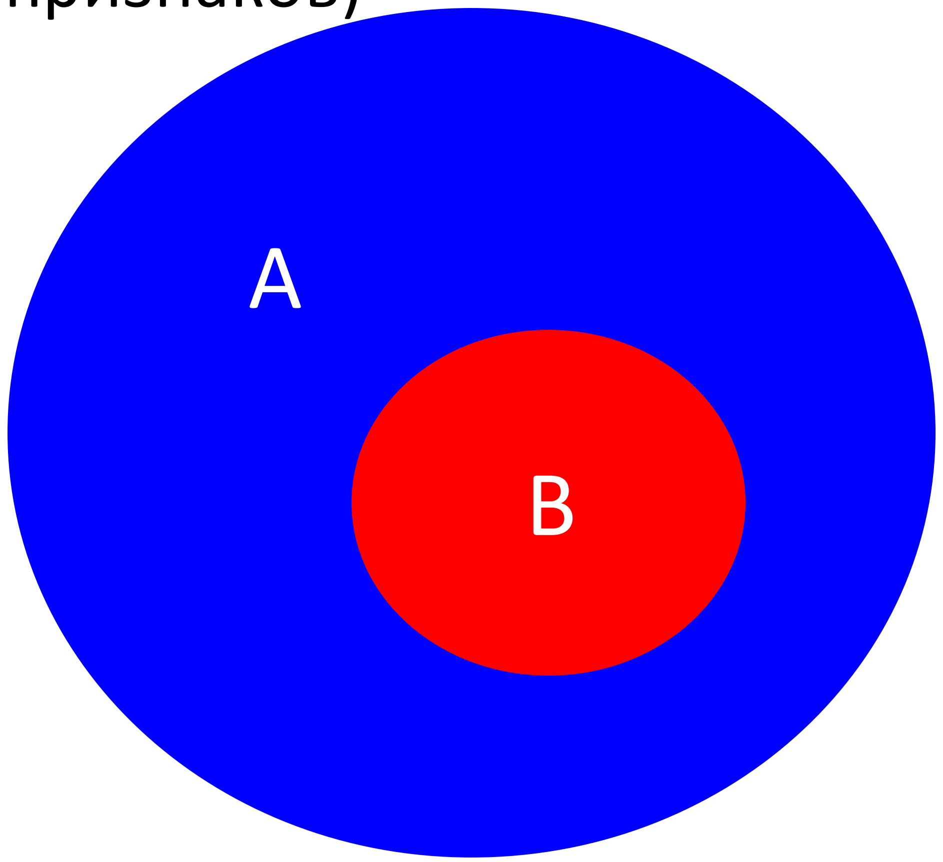
- Выбираем случайное подмножество признаков
- Обучаем модель только на них

Случайные подпространства

- Выбираем случайное подмножество признаков
- Обучаем модель только на них
- Может быть плохо, если в подмножество не попадут важные признаки

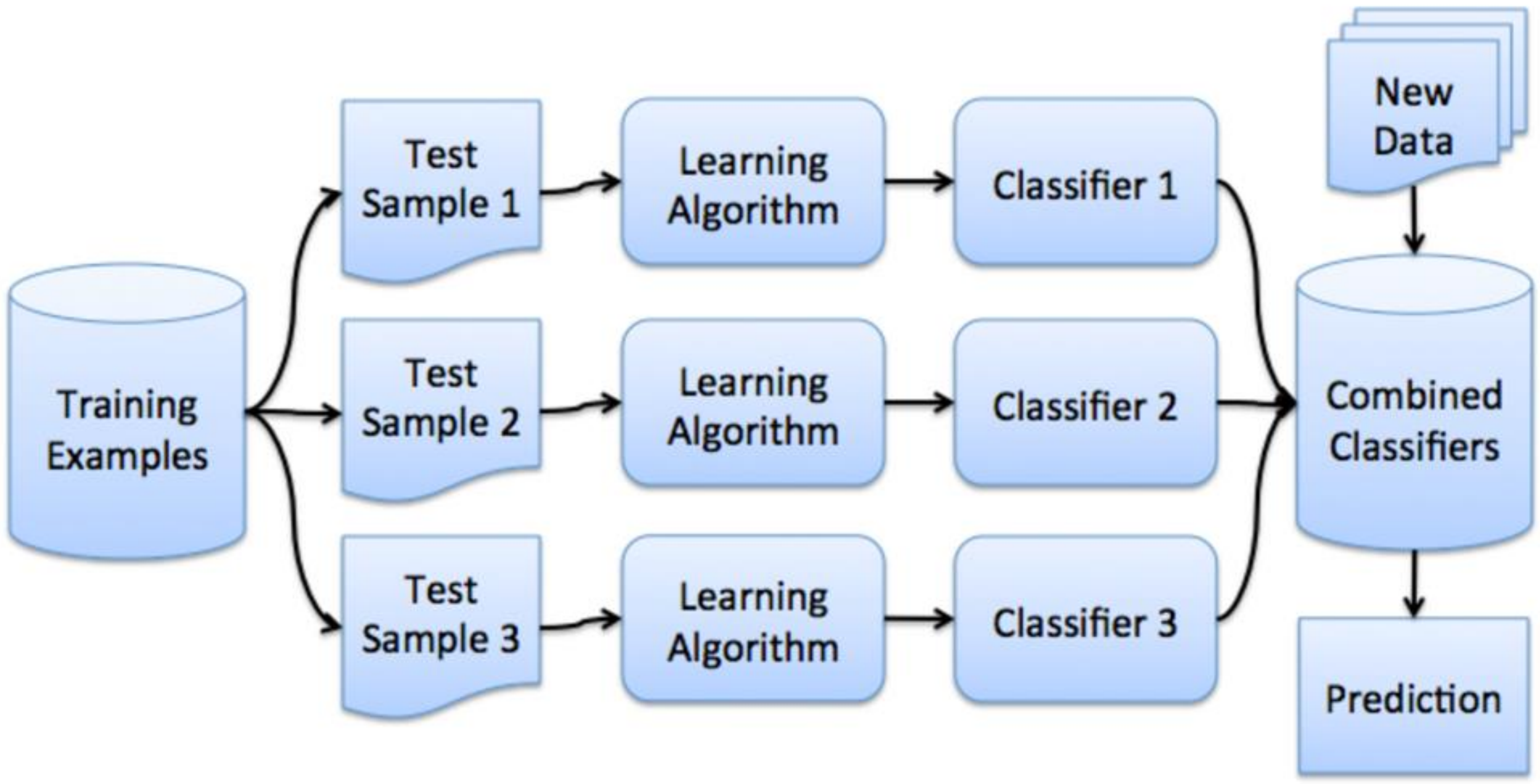
Виды рандомизации

- Бэггинг (случайные подвыборки)
- Случайные подпространства (случайное множество признаков)



Резюме

- Будем объединять модели в композиции через усреднение или голосование большинством
- Бэггинг - композиция моделей, обученных независимо на случайных подмножествах объектов
- Рандомизируем признаки, на которых обучаются модели



03

Разложение на смещение и разброс

Bias-variance decomposition

- Разберем на уровне идеи
- Ошибка модели складывается из трех компонент:

Bias-variance decomposition

- Разберем на уровне идеи
- Ошибка модели складывается из трех компонент:
- Шум (noise) - характеристика сложности и противоречивости данных

Bias-variance decomposition

- Разберем на уровне идеи
- Ошибка модели складывается из трех компонент:
- Шум (noise) - характеристика сложности и противоречивости данных
- Смещение (bias) - способность модели приблизить лучшую среди всех возможных моделей

Bias-variance decomposition

- Разберем на уровне идеи
- Ошибка модели складывается из трех компонент:
- Шум (noise) - характеристика сложности и противоречивости данных
- Смещение (bias) - способность модели приблизить лучшую среди всех возможных моделей
- Разброс (variance) - устойчивость модели к изменениям в обучающей выборке

Bias-variance decomposition

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{шум}} +$$

$$+ \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{смещение}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X) - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right]}_{\text{разброс}}$$

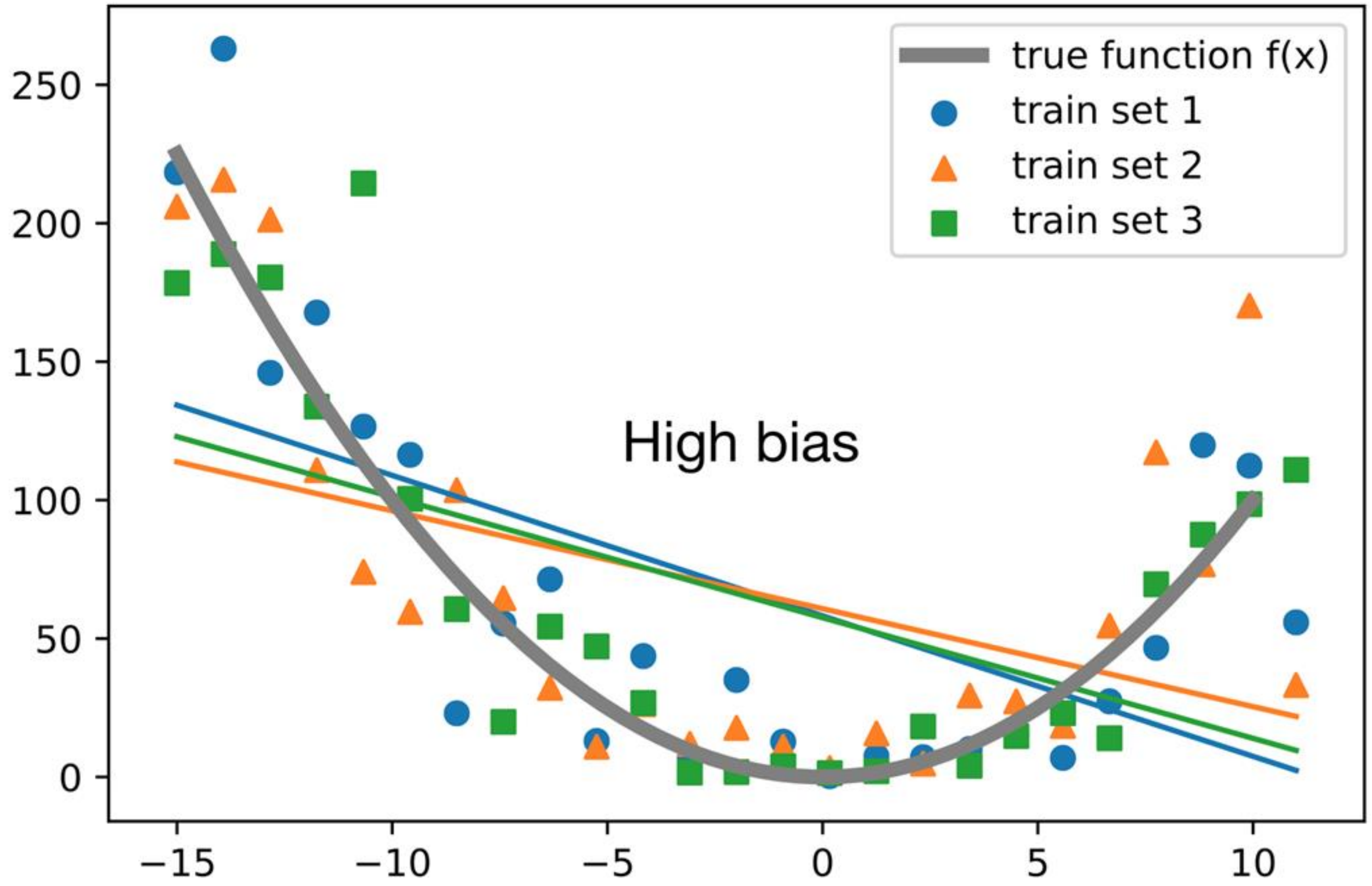
$$\text{bias} := \mathbb{E}(\hat{y}) - y.$$

$$\text{variance} := \mathbb{E}[\mathbb{E}(\hat{y}) - \hat{y}]^2$$

$$\text{noise} := \mathbb{E}[y - \mathbb{E}(y)]^2$$

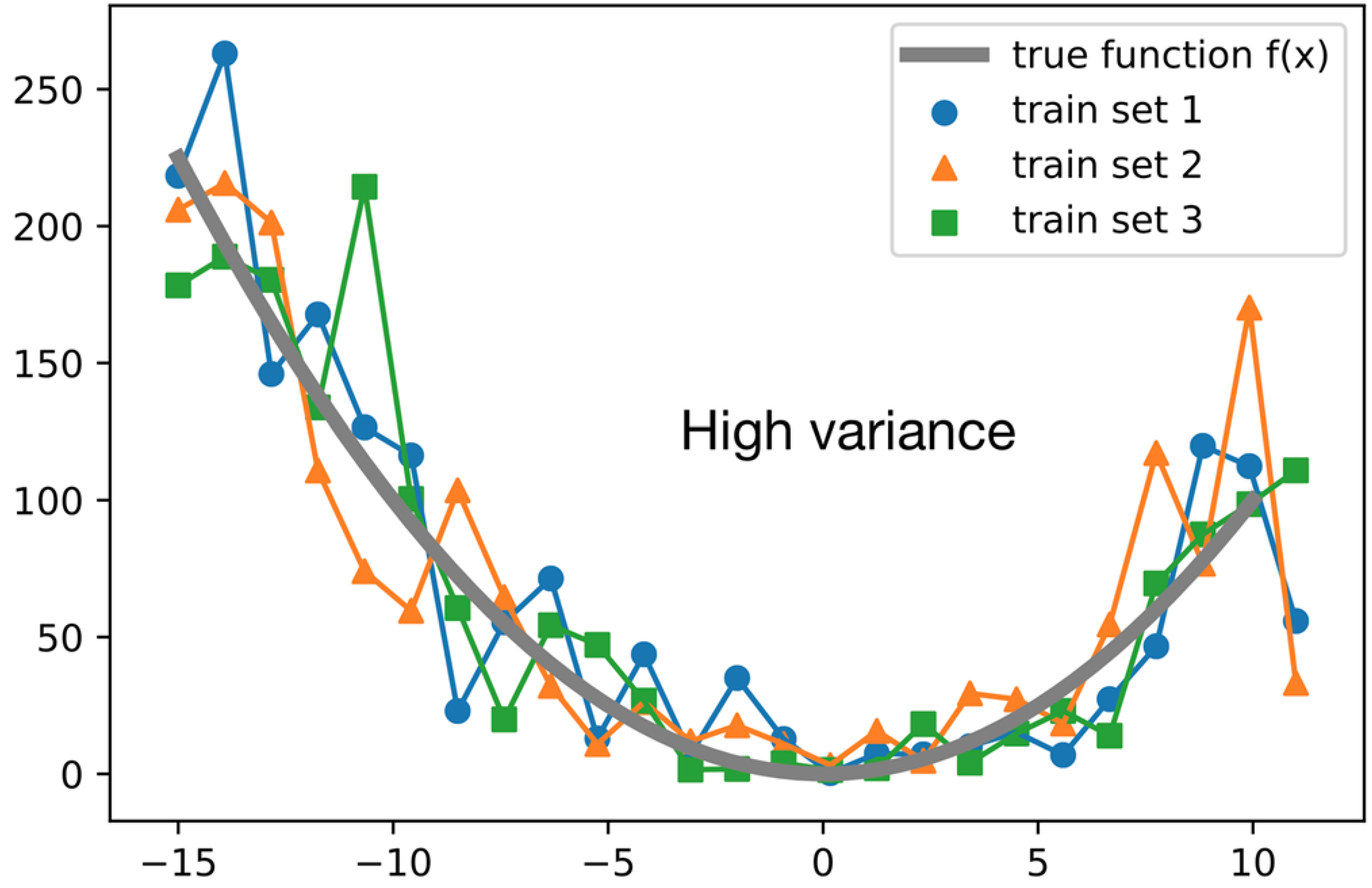
Bias-variance decomposition

- Пример высокого смещение у линейной модели



Bias-variance decomposition

- Пример высокого разброса у более сложной модели

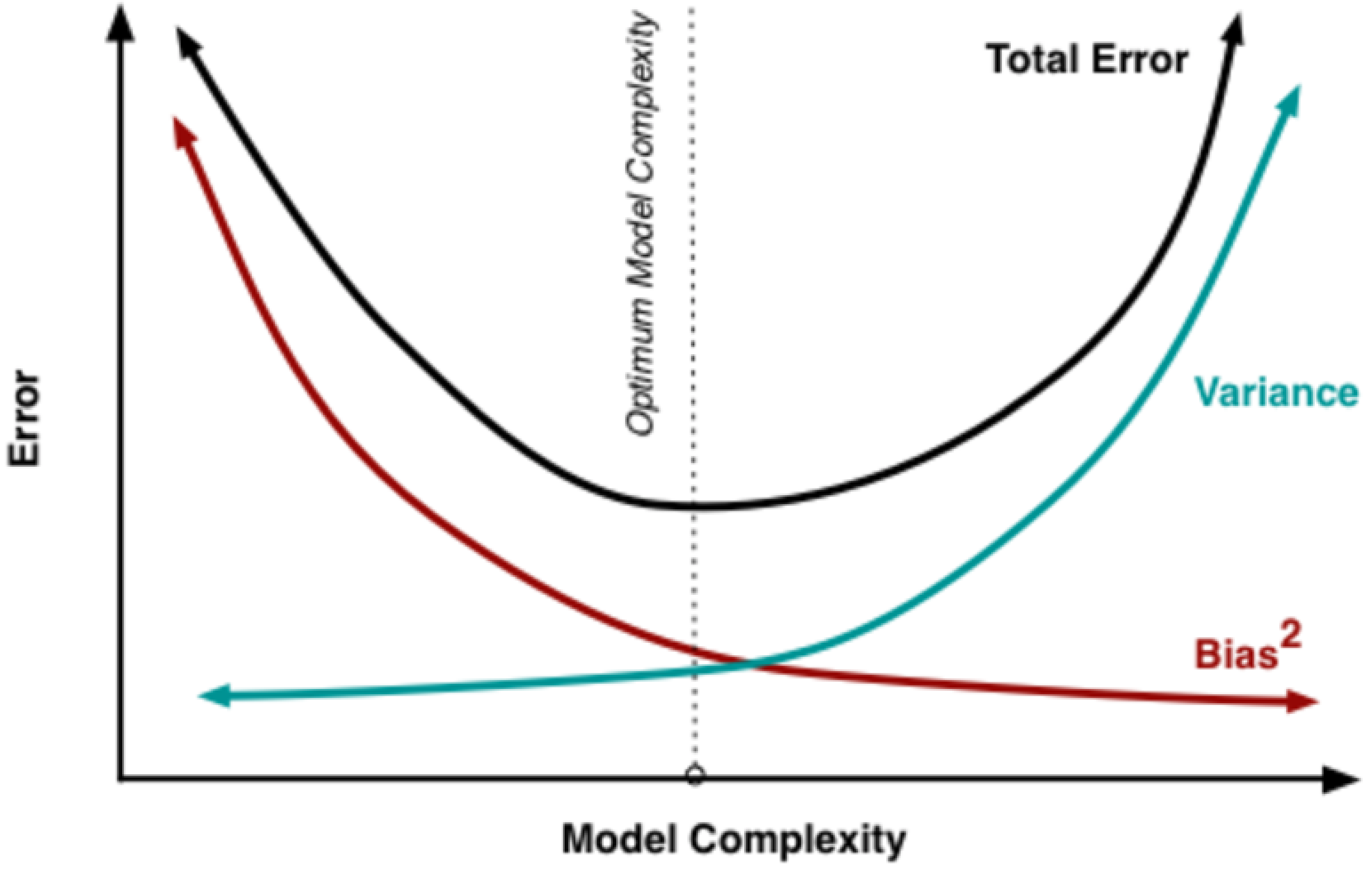


Bias-variance tradeoff

- Недообученная модель имеет низкий разброс, но высокое смещение
- Переобученная модель имеет высокий разброс, но низкое смещение

Bias-variance tradeoff

- Недообученная модель имеет низкий разброс, но высокое смещение
- Переобученная модель имеет высокий разброс, но низкое смещение
- Необходимо искать золотую середину



Bias-variance + бэггинг

- Смещение $a_N(x)$ такое же, как у $b_n(x)$
- Разброс $a_N(x)$:

$$\frac{1}{N}\big(\text{разброс } b_n(x)\big) + \text{ковариация}\big(b_n(x), b_m(x)\big)$$

- Если базовые модели независимы, то разброс уменьшается в N раз!
- Чем более похожи выходы базовых моделей, тем меньше эффект от построения композиции

04

Алгоритм случайного леса

Алгоритм

Для n = 1, ..., N:

1. Генерируем выборку **X**ʹ с помощью бутстрапа
2. Строим решающее дерево $b_n(x)$ по выборке **X**ʹ
3. Строим дерево, пока не выполнится критерий остановки (обычно пока не достигнет n_{min} объектов в листах)
4. Оптимальное разбиение ищется среди *q* случайных признаков, в каждом узле (не дереве) обновляется набор признаков

Алгоритм

Для n = 1, ..., N:

1. Генерируем выборку **X**′ с помощью бутстрапа
2. Строим решающее дерево $b_n(x)$ по выборке **X**′
3. Строим дерево, пока не выполнится критерий остановки (обычно пока не достигнет n_{min} объектов в листах)
4. Оптимальное разбиение ищется среди *q* случайных признаков, в каждом узле (не дереве) обновляется набор признаков

Выбор предиката

4. Оптимальное разбиение ищется среди q случайных признаков, в каждом узле (не дереве) обновляется набор признаков

$$j, t = \arg\min_{\mathrm{j,t}} Q(R_m, j, t)$$

Будемискатьлучшийпредикатсредислучайногоподмножества признаков размера q

Алгоритм

Рекомендации для q:

- Регрессия:

$$q = \frac{d}{3}$$

- Классификация:

$$q = \sqrt{d}$$

05

Особенности применения

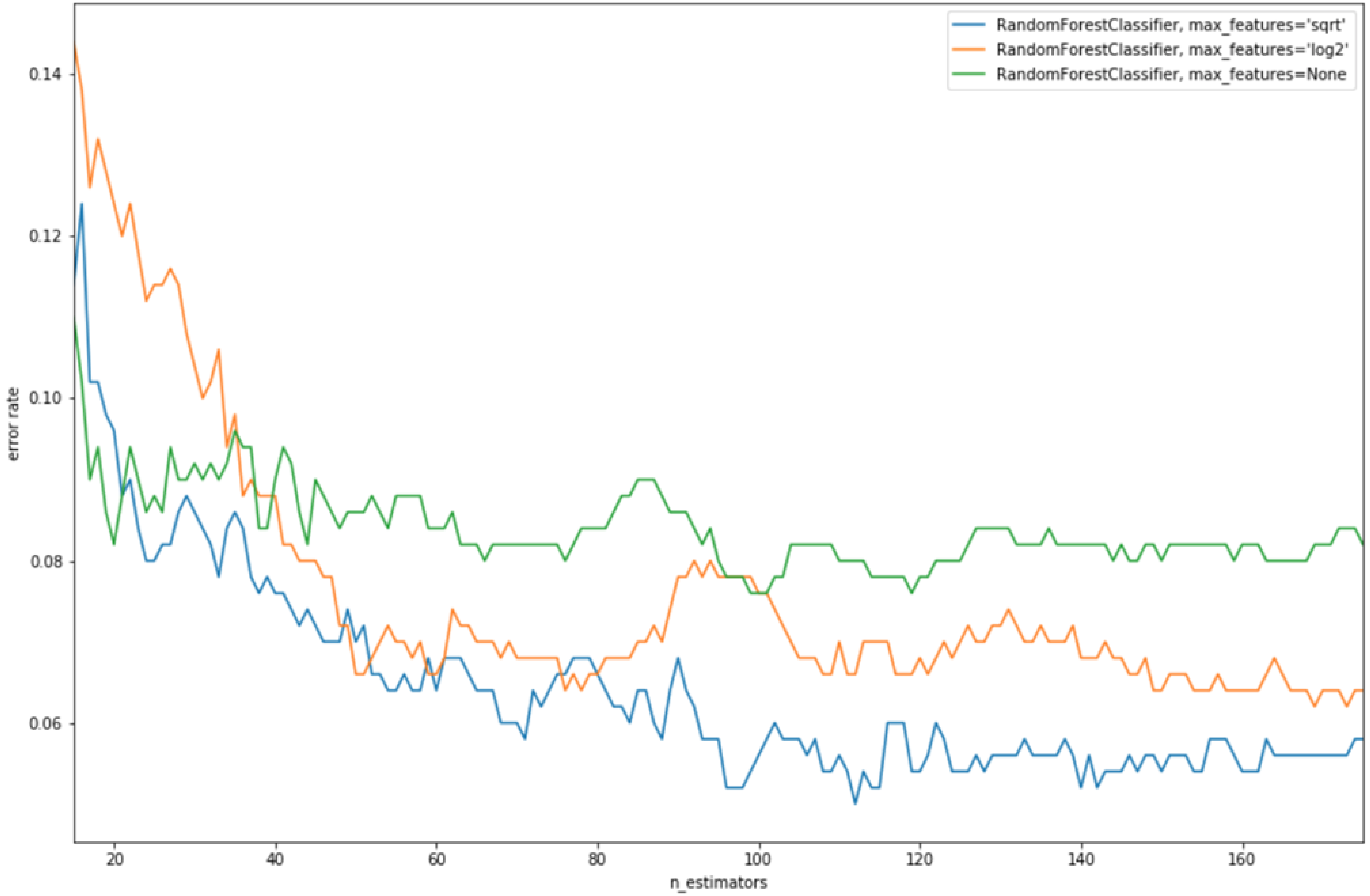
Переобучение

- Мы уже узнали, что увеличение количества базовых моделей приводит к уменьшению разброса

Получается мы можем брать неограниченное количество базовых моделей и уменьшать ошибку?

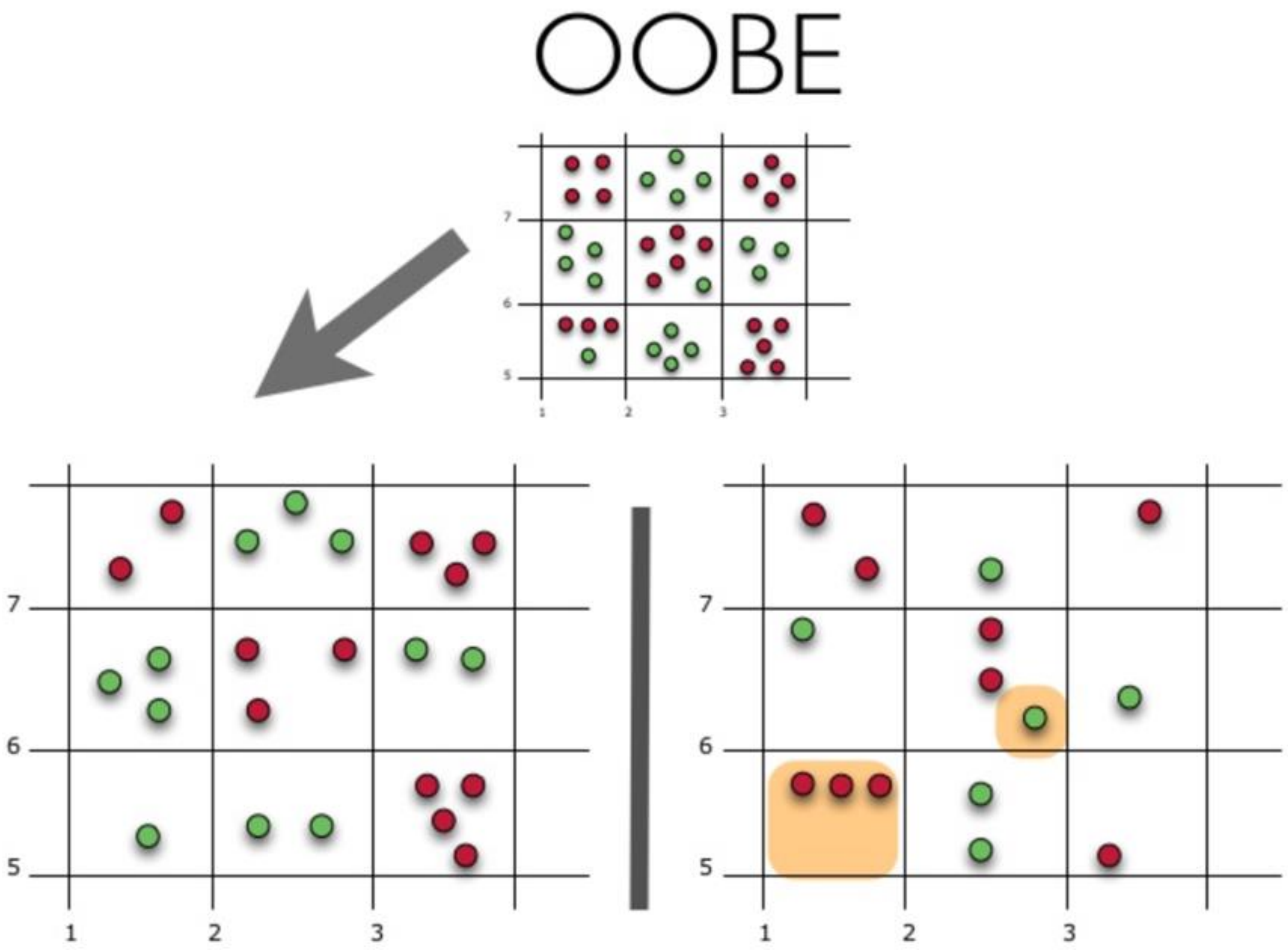
Переобучение

- Ошибка сначала убывает, а затем выходит на один уровень
- Случайный лес не переобучается при росте N



Out-of-bag error

- Благодаря особенности построения случайного леса, потребность в кросс-валидации отсутствует



Out-of-bag error

- Благодаря особенности построения случайного леса, потребность в кросс-валидации отсутствует
- Каждое дерево обучается примерно на 63% данных
- Остальные объекты — как бы тестовая выборка для дерева
- X_n — обучающая выборка для $b_\mathrm{n}(x)$
- Можно оценить ошибку на новых данных:

Out-of-bag error

- Благодаря особенности построения случайного леса, потребность в кросс-валидации отсутствует
- Каждое дерево обучается примерно на 63% данных
- Остальные объекты — как бы тестовая выборка для дерева
- X_n — обучающая выборка для $b_n(x)$
- Можно оценить ошибку на новых данных:

$$Q_{test} = \frac{1}{\ell}\sum_{i=1}^{\ell} L\left(y_i, \frac{1}{\sum_{n=1}^{N}[x_i \notin X_n]}\sum_{n=1}^{N}[x_i \notin X_n] b_n(x_i)\right)$$

Важность признаков

- Перестановочный метод для проверки важности j-го признака
- Перемешиваем соответствующий столбец в матрице «объекты-признаки» для тестовой выборки
- Измеряем качество модели
- Чем сильнее оно упало, тем важнее признак

Объект / Признак	Возраст	Вес	Рост
0	17	60	165
1	24	86	193
2	28	98	185

→

Объект / Признак	Возраст	Вес	Рост
0	17	86	165
1	24	98	193
2	28	60	185

Резюме

- Случайный лес — метод на основе бэггинга, в котором делается попытка повысить разнообразие деревьев
- Метод практически без гиперпараметров
- Можно оценить обобщающую способность без тестовой выборки

Место для ваших вопросов