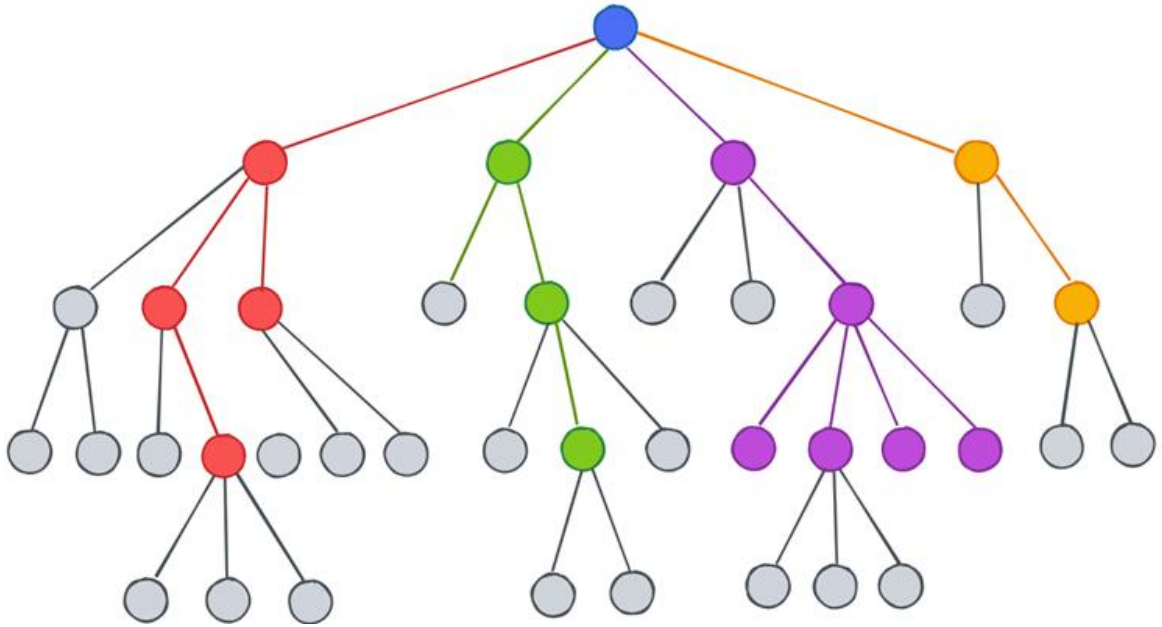
Композиции моделей



Разложение ошибки на смещение и разброс

Ошибка модели складывается из трёх компонент:

- Шум (noise) – характеристика сложности и противоречивости данных

$$noise \coloneqq \mathbb{E}[y - \mathbb{E}(y)]^2$$

- Смещение (bias) – способность модели приблизить лучшую среди всех возможных моделей

$$bias \coloneqq \mathbb{E}(\hat{y}) - y$$

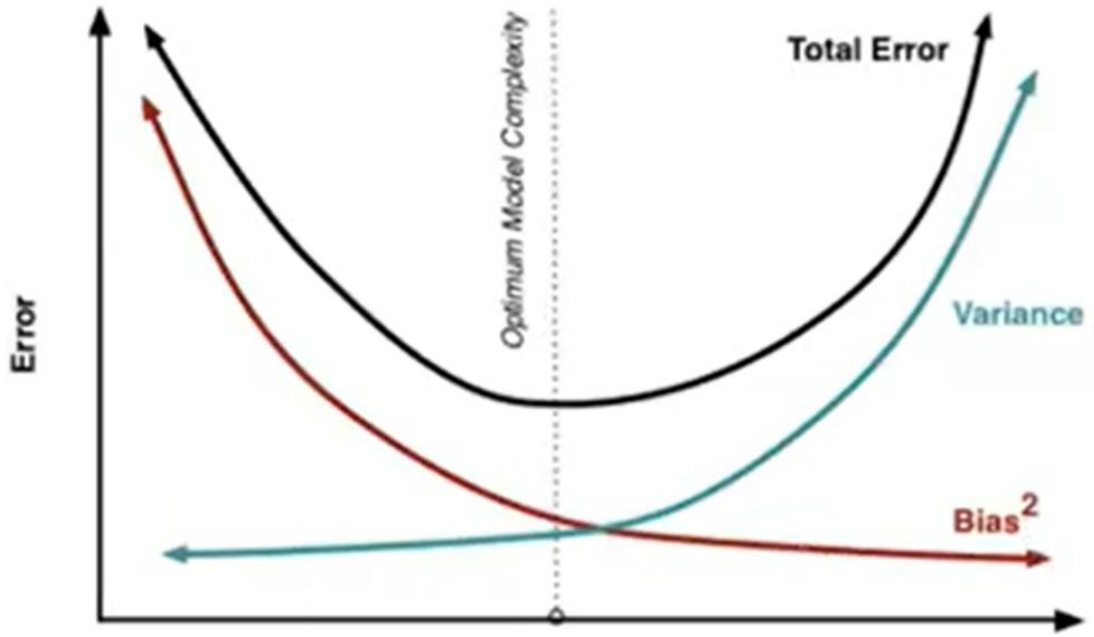
- Разброс (variance) – устойчивость модели к изменениям в обучающей выборке

$$variance \coloneqq \mathbb{E}[\mathbb{E}(\hat{y}) - y\,]^2$$

Среднеквадратичская ошибка:

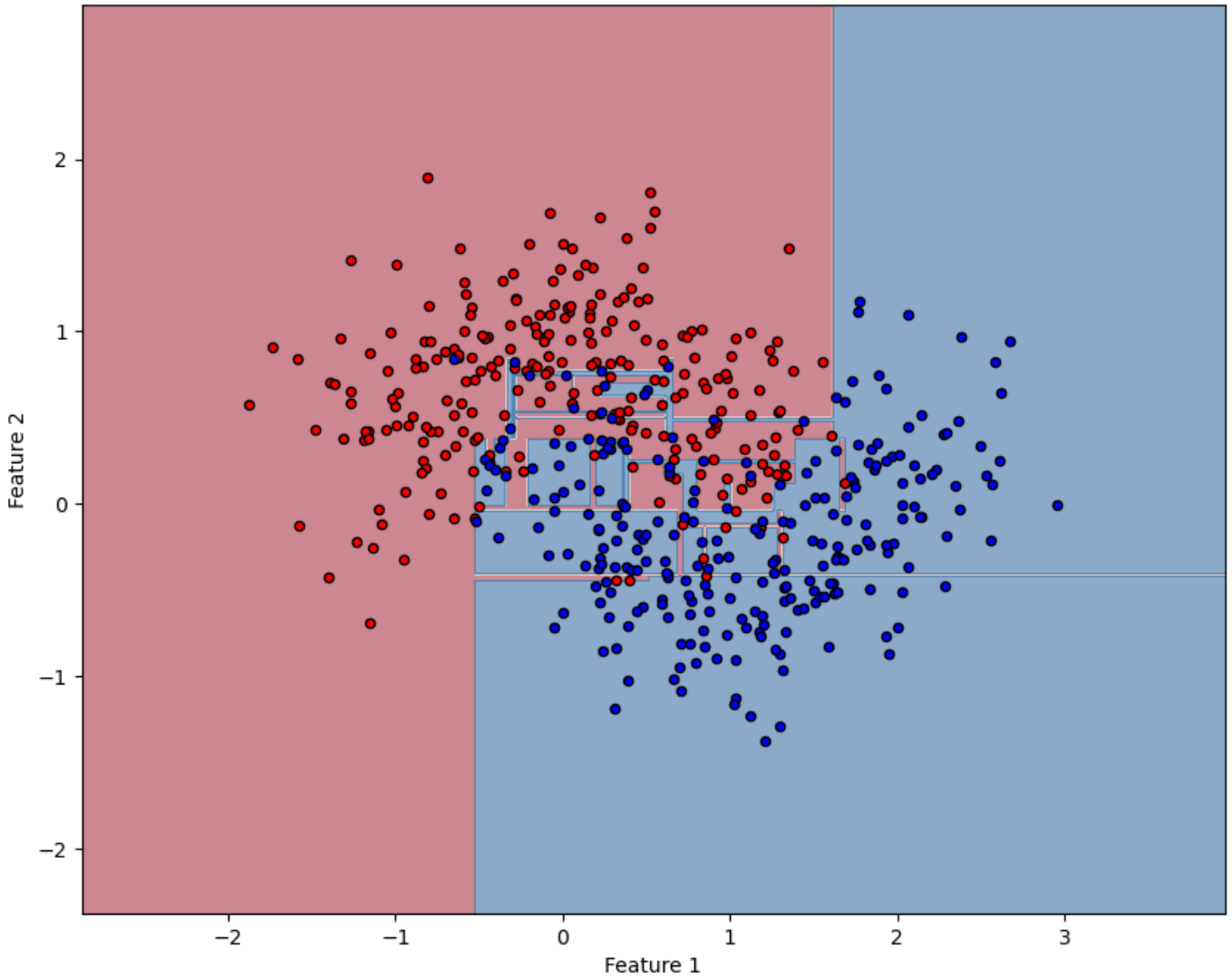
$$\mathbb{E}[y - \hat{y}]^2 = bias^2 + variance + noise$$

Разложение ошибки на смещение и разброс

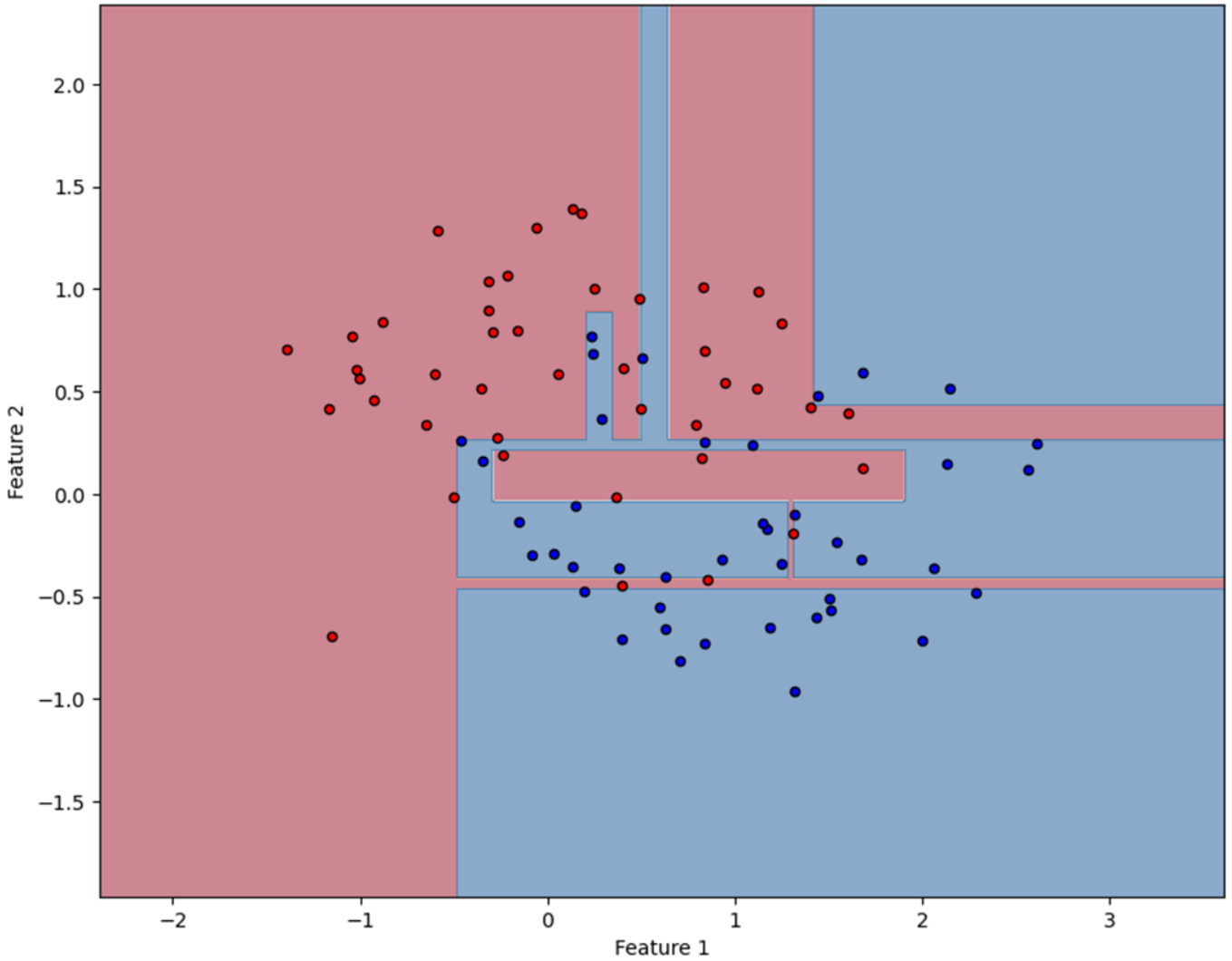


- Очень простая модель имеет большое смещение, но малую или нулевую дисперсию (модель недообучена)
- В меру сложная модель имеет небольшое смещение и небольшую дисперсию
- Очень сложная модель имеет небольшое смещение, но большую дисперсию (модель переобучена)

Неустойчивость деревьев



Неустойчивость деревьев

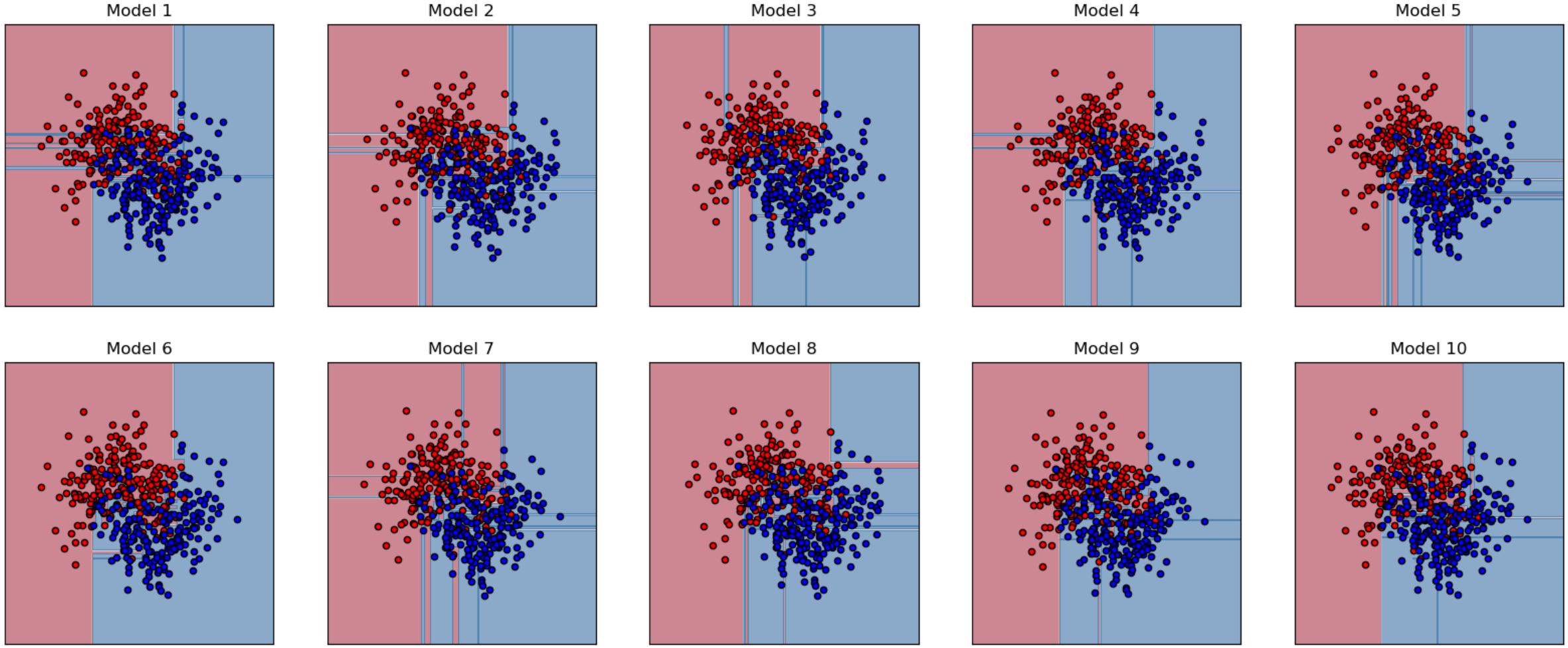


Композиция (Ensemble)

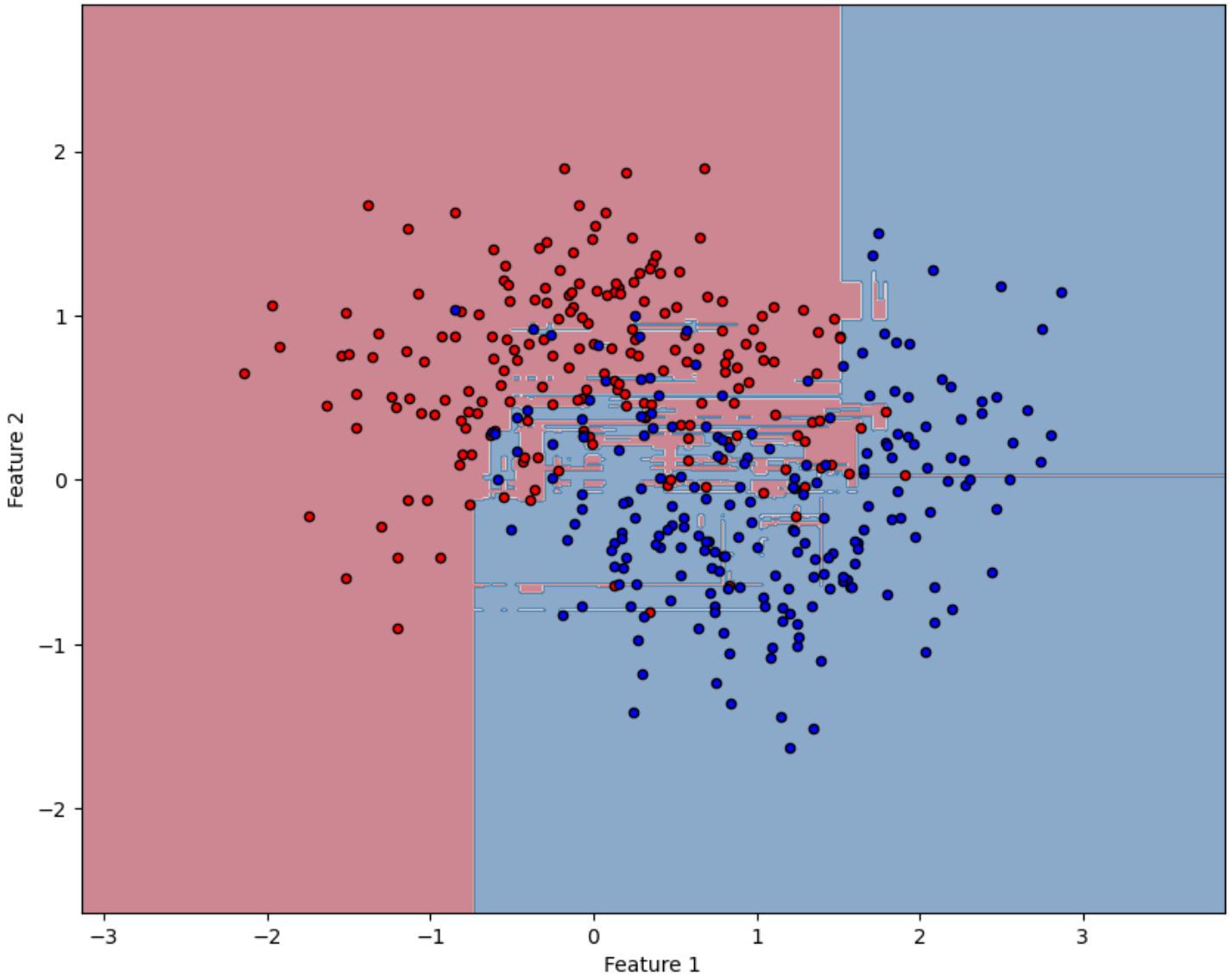
- $b_1(x), \ldots, b_n(x)$ - базовые модели (weak learners)
- Композиция: голосование по большинству (majority vote)

$$a(x) = \underset{y \in \mathbb{Y}}{\operatorname{argmax}} \sum_{n=1}^{n} [b_n(x) = y]$$

Неустойчивость деревьев



Композиция (Ensemble)

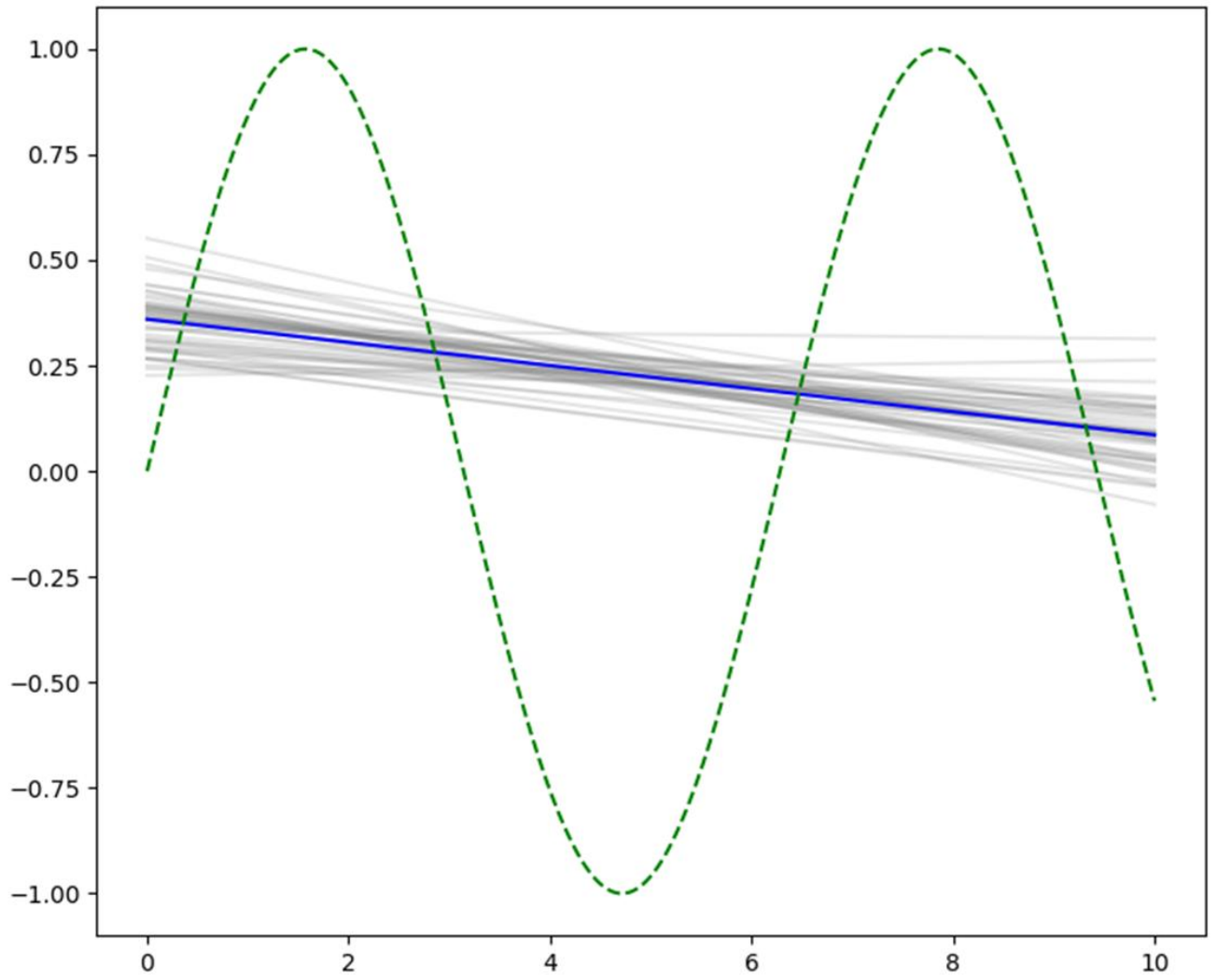


Регрессия

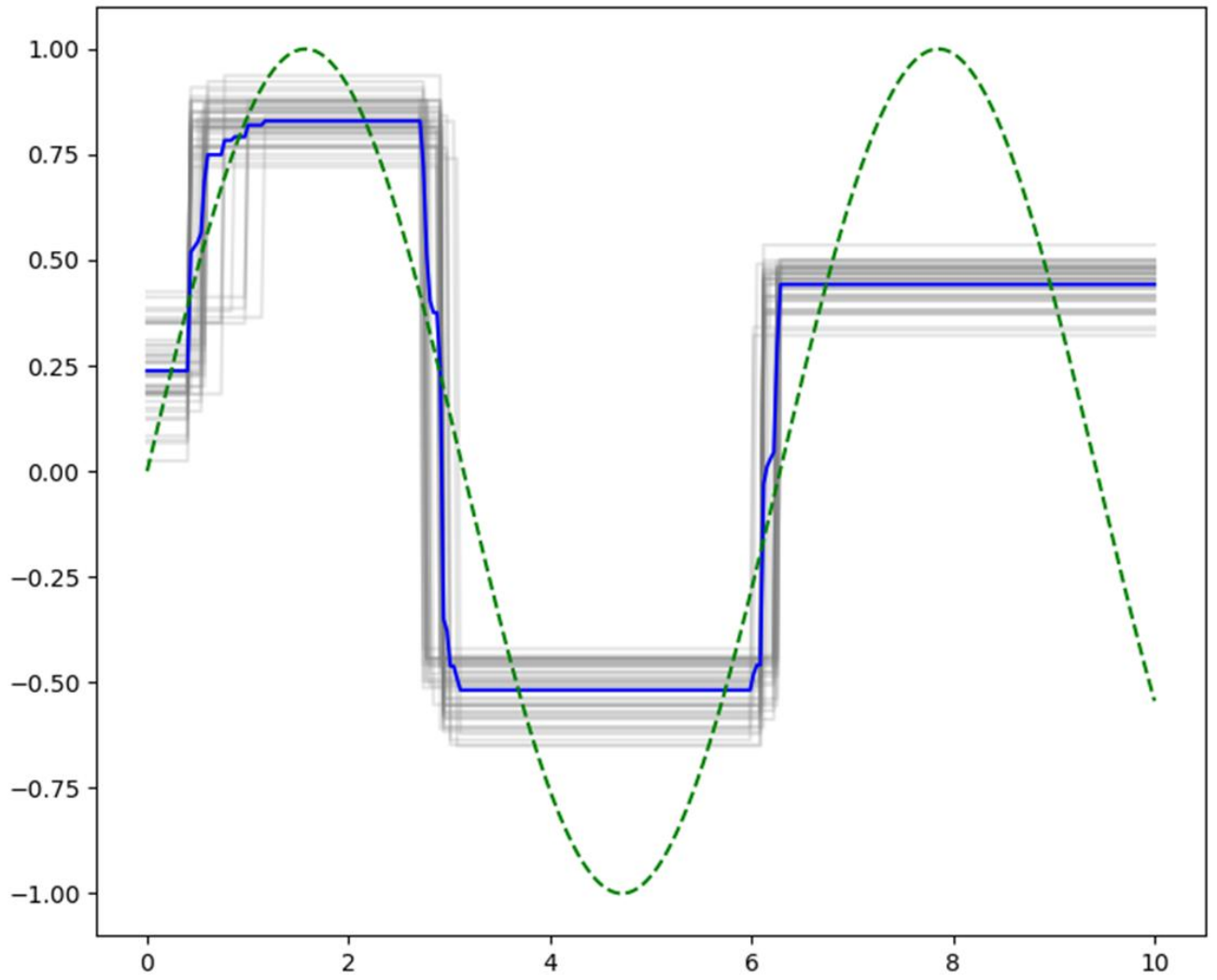
- $b_1(x), \ldots, b_n(x)$ - базовые модели (weak learners)
- Композиция: усреднение

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

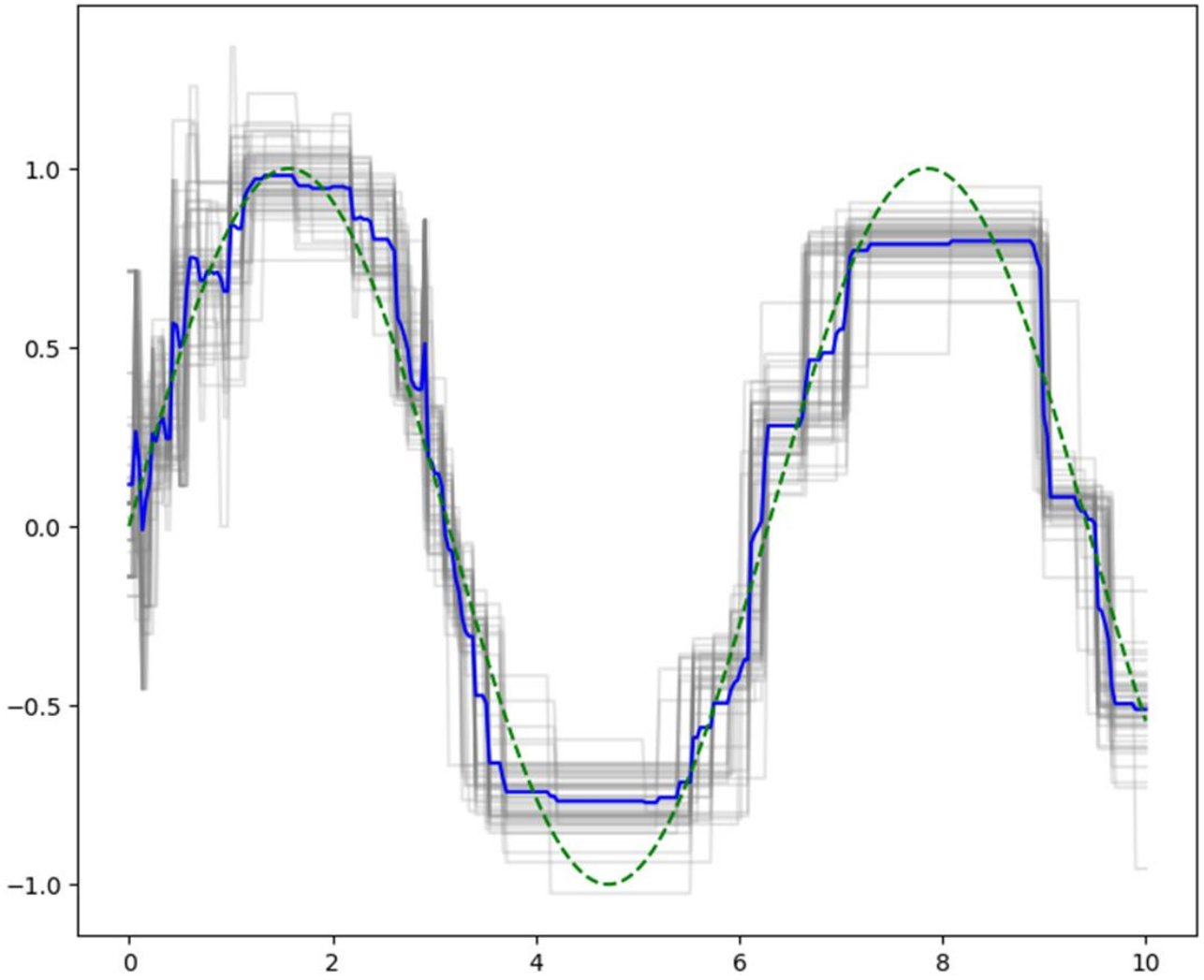
Разложение ошибки на смещение и разброс



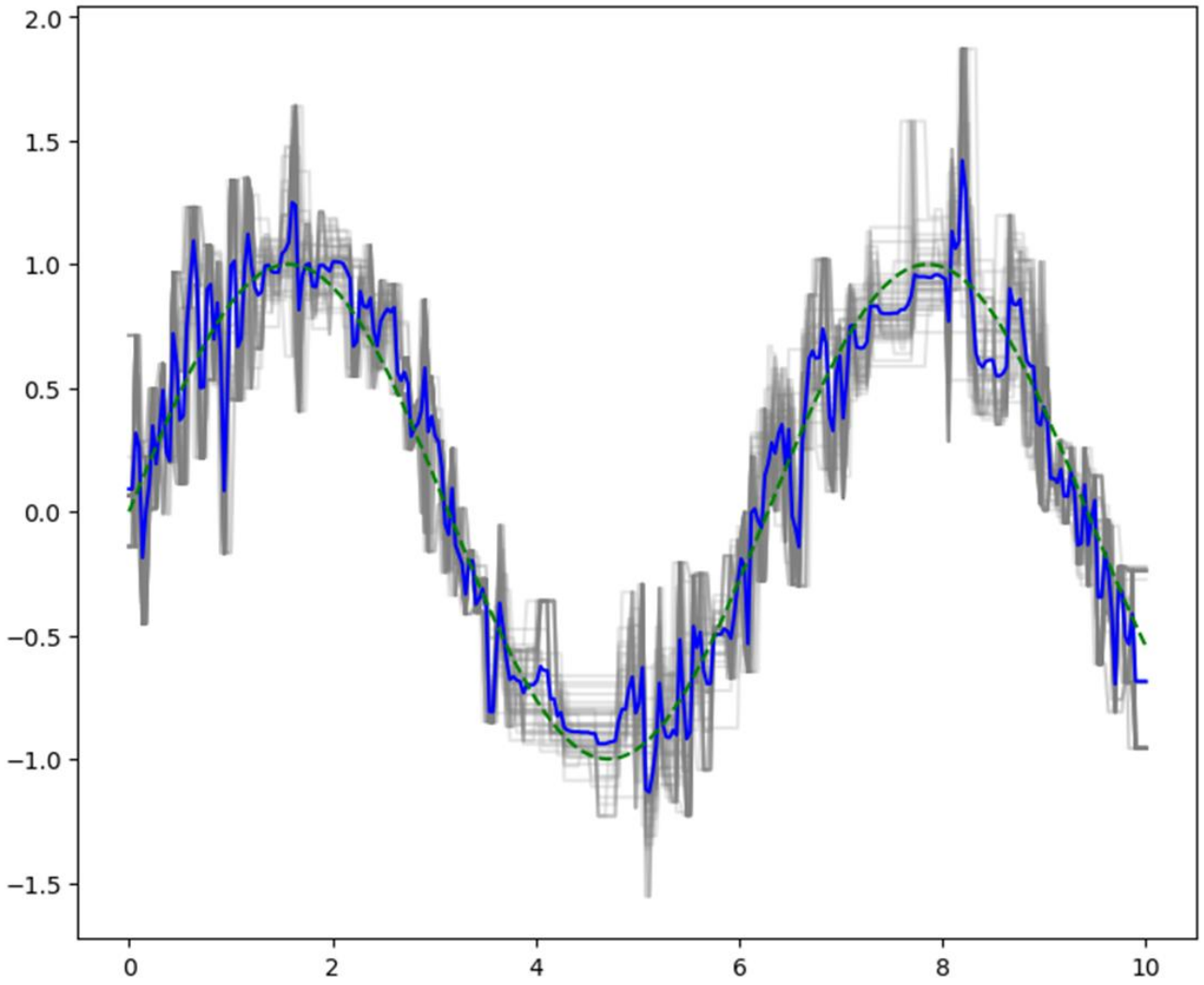
Разложение ошибки на смещение и разброс



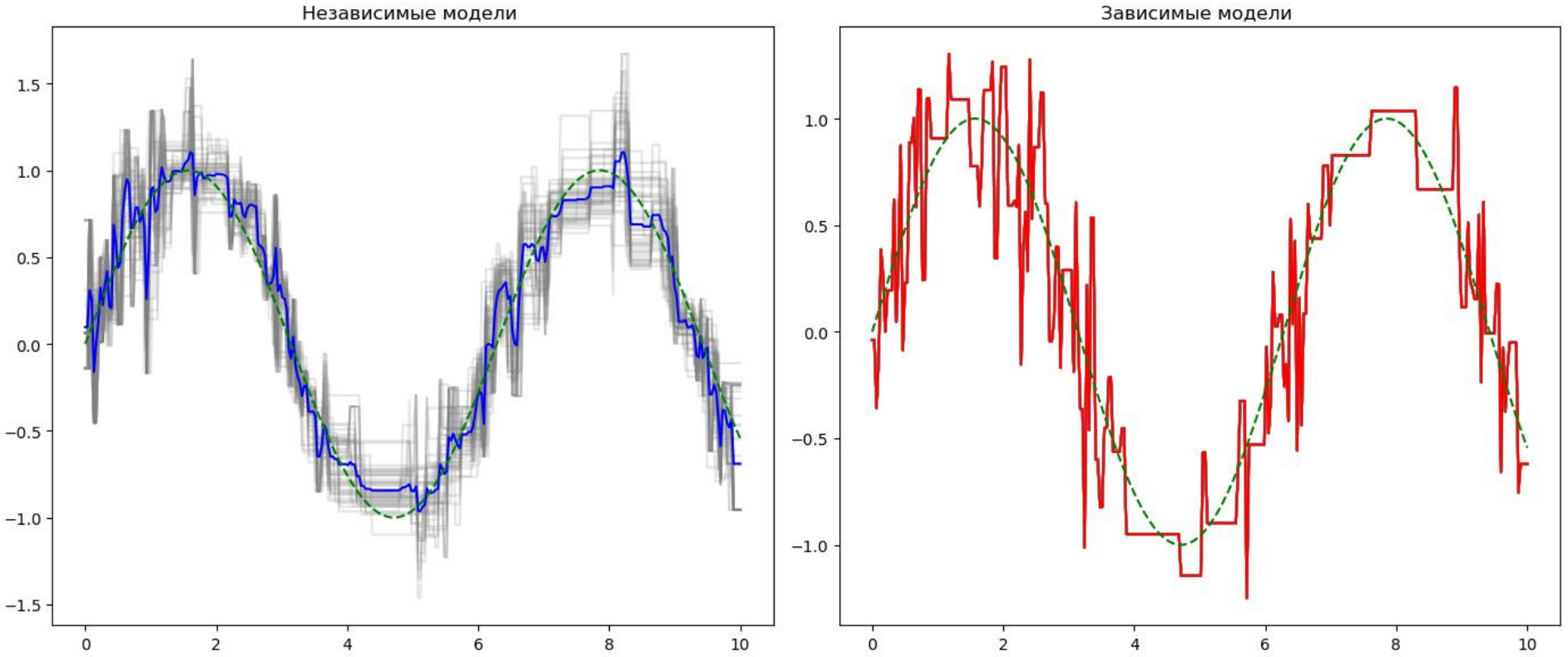
Разложение ошибки на смещение и разброс



Разложение ошибки на смещение и разброс



Разложение ошибки на смещение и разброс



Базовые модели

- $b_1(x), \ldots, b_n(x)$ - базовые модели

Как на одной выборке построить N различных моделей?

Bagging (bootstrap aggregation)

- Базовые модели обучаются независимо
- Каждый обучается на подмножестве обучающей выборки
- Подмножество выбирается с помощью бутстрапа

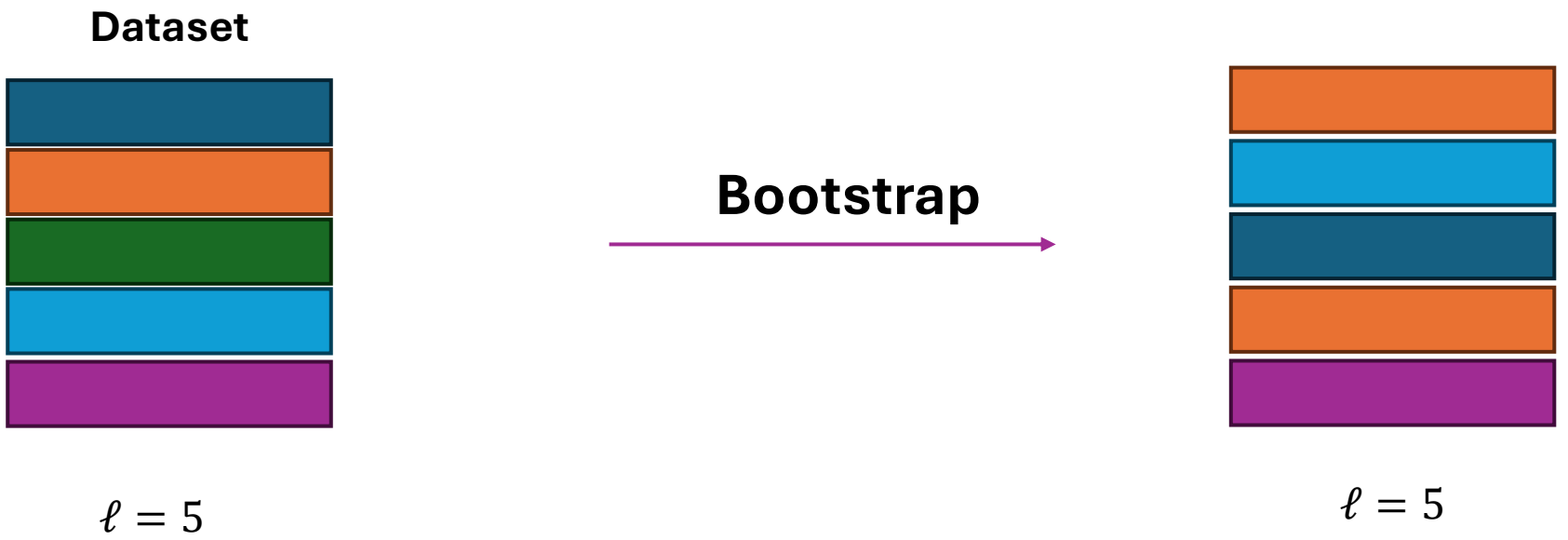
Bootstrap

- Выборка с возвращением
- Берём ℓ элементов из X (выборка с возвращением)
- Если объект входит в выборку несколько раз, то мы как бы повышаем его вес

$$\frac{1}{4}(L(y, a(\bigcirc)) + 2L(y, a(\bigcirc)) + l(y, a(\bigcirc))$$

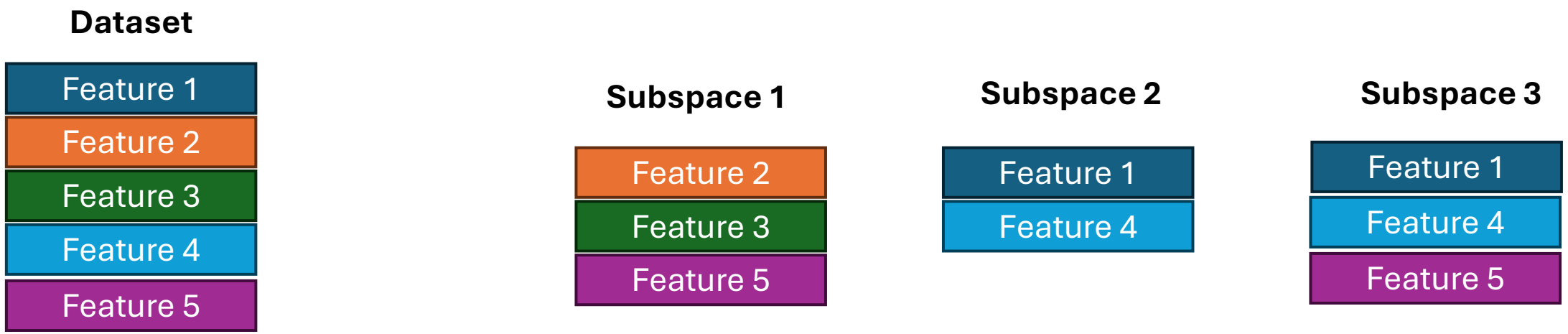
Bootstrap

- При bootstrap выборки размером ℓ из исходной выборки также размером ℓ
- Примерно 63.2% уникальных объектов из исходной выборки попадут хотя бы один раз в bootstrap выборку и 36.8% не попадут



Случайные подпространства (Random Subspaces)

- Выбираем случайное подмножество признаков
- Обучаем модель только на них



Жадный алгоритм

1. Поместить в корень всю выборку: $R_1 = X$
2. Запустить построение из корня: SplitNode(1, R_1)

SplitNode(m, R_m)

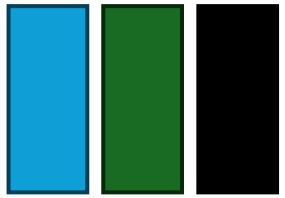
1. Если выполнен критерий остановка, то выход
2. Ищем лучший предикат: $j, t = \underset{j,t}{\operatorname{argmin}} Q(R_m, j, t)$
3. Разбиваем с его помощью объектов: $R_l = \{(x, y) \in R_m \mid [x_j < t]\}$, $R_r = \{(x, y) \in R_m \mid [x_j \geq t]\}$
4. Повторяем для дочерних вершин: SplitNode(l, R_l), : SplitNode(r, R_r)

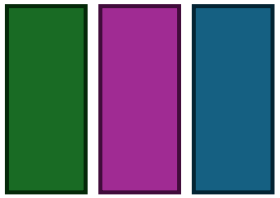
Выбор предиката

$$j, t = \underset{j,t}{\operatorname{argmin}}\, Q(R_m, j, t)$$

Будем искать лучший предикат среди случайного подмножества признаков размера q

q=3





Случайный лес (Random Forest)

Рекомендации для q:

Регрессия: $q = \frac{d}{3}$

Классификация : q $= \sqrt{d}$

Случайный лес

n=1,…,N

1. Для построения *i*-го дерева

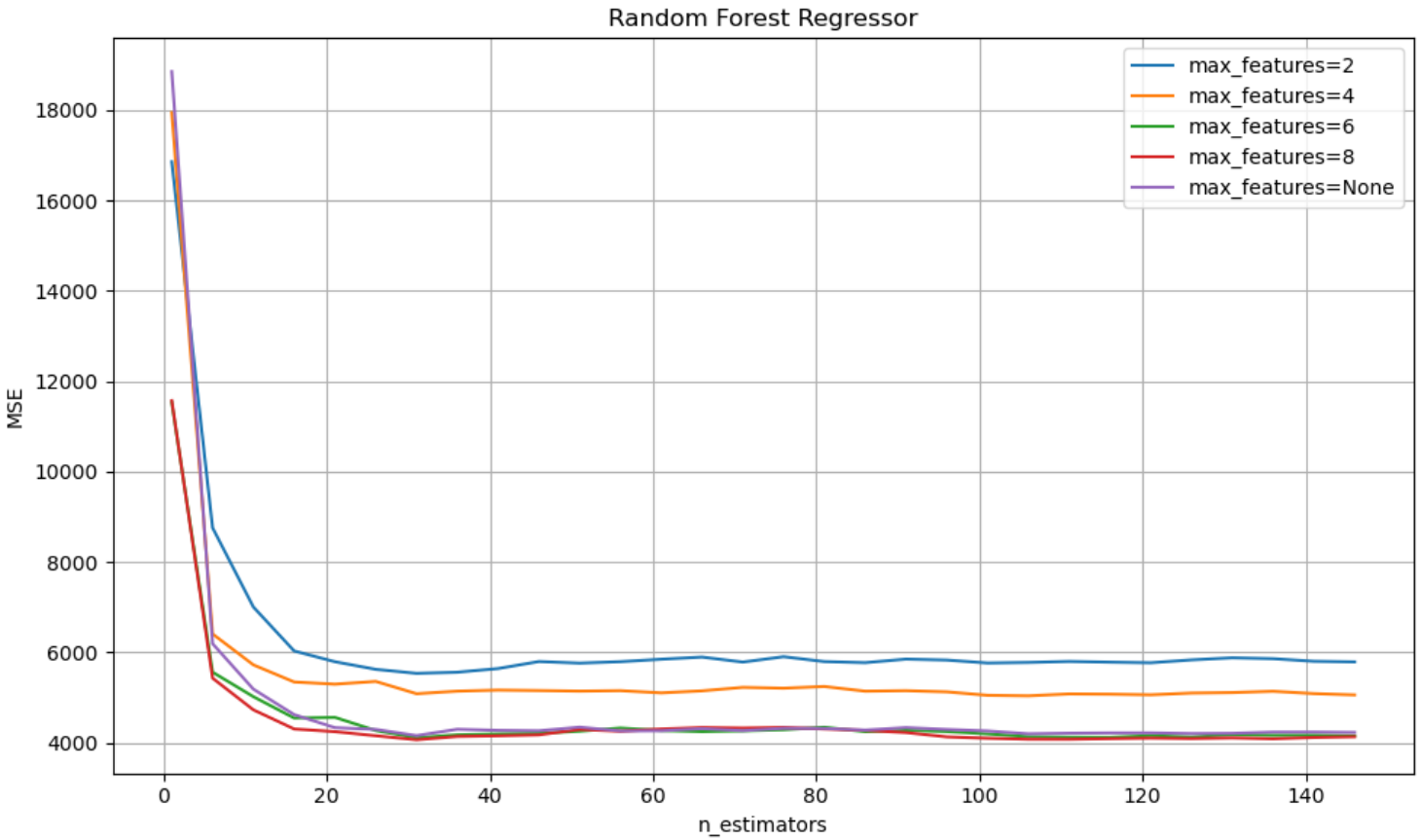
- Сначала, как в обычном бэггинге, из обучающей выборки X выбирается с возвращением случайная подвыборка X^i того же размера, что и X

В процессе обучения каждого дерева в каждой вершине случайно выбираются q признаков

2. Чтобы получить предсказание ансамбля на тестовом объекте, усредняем отдельные ответы деревьев (для регрессии) или берём самый популярный класс (для классификации)

Универсальный метод

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



Out-of-bag

- 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:

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

ℓ — Общее количество объектов

N — Количество деревьев в ансамбле

X_n Обучающая выборка для дерева n -

$[x_i \notin X_n]$ - Индикатор: объект не участвовал в обучении дерева n