

Ансамблирование

Сергей Скорик

ИСП РАН

10 апреля 2025

Ансамбли



Figure: Коллектив



Figure: Ансамбль

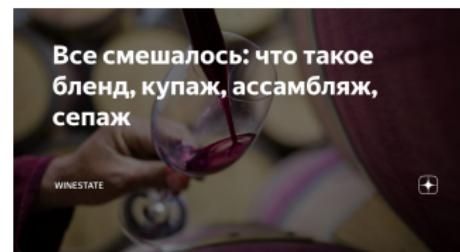
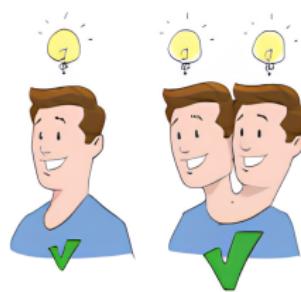


Figure: Ассамбль

Оценка моделей

Одна голова хорошо,
а две лучше!



© russkyl.info

Figure: Народная мудрость

- А что для моделей машинного обучения значит “лучше”?

Оценка моделей

	A, Precision	A, Recall	B, Precision	B, Recall
Model 1	0.986	0.845	0.894	0.791
Model 2	0.910	0.897	0.901	0.876

- Что можно сказать о моделях 1 и 2?

Оценка моделей

	A, Precision	A, Recall	B, Precision	B, Recall
Model 1	0.986	0.845	0.894	0.791
Model 2	0.910	0.897	0.901	0.876

- Что можно сказать о моделях 1 и 2?

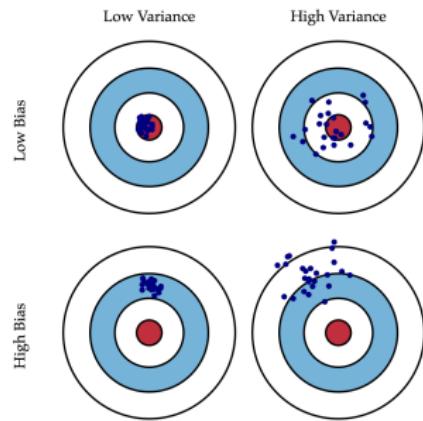


Figure: Иллюстрация сдвига и разброса для различных моделей.

Bias-Variance Decomposition I

- Рассмотрим задачу регрессии

$$y = f(x) + \varepsilon, \quad y, x, \varepsilon \in \mathbb{R}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \quad (1)$$

- Определим модель $a(x)$. Тогда, квадратичная функция потерь на объекте x : $\mathcal{L} = [y(x) - a(x)]^2$.
- Шум ε моделирует погрешности реальных измерений. Модель a обучается на конкретной реализации X распределения данных \mathcal{X} .
- В этом случае, целевая и моделируемая зависимости могут быть уточнены:

$$y(x) \rightarrow y(x, \varepsilon), \quad a(x) \rightarrow a(x, X),$$

где ε и X являются случайными величинами.

Bias-Variance Decomposition II

- Рассмотрим задачу регрессии

$$y(x, \varepsilon) = f(x) + \varepsilon.$$

- Определим модель $a(x, X)$. Как будет выглядеть функция потерь в этом случае?

Bias-Variance Decomposition II

- Рассмотрим задачу регрессии

$$y(x, \varepsilon) = f(x) + \varepsilon.$$

- Определим модель $a(x, X)$. Как будет выглядеть функция потерь в этом случае?
- Возьмём математическое ожидание по совместному распределению с.в. (X, ε) :

$$Q(a, x) = \mathbb{E}_{X, \varepsilon} [y(x, \varepsilon) - a(x, X)]^2. \quad (2)$$

- Случайные величины X, ε независимы, поэтому совместное матожидание можно заменить на последовательное $\mathbb{E}_{X, \varepsilon} \rightarrow \mathbb{E}_X \mathbb{E}_\varepsilon$.

Bias-Variance Decomposition III

Преобразуем $Q(a, x)$

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X \mathbb{E}_{\varepsilon} [y(x, \varepsilon) - a(x, X)]^2 = \\ &= \end{aligned}$$

Bias-Variance Decomposition III

Преобразуем $Q(a, x)$

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X \mathbb{E}_\varepsilon [y(x, \varepsilon) - a(x, X)]^2 = \\ &= \mathbb{E}_X \mathbb{E}_\varepsilon [\underbrace{(f(x) - a(x, X))^2}_{\text{не зависит от } \varepsilon} + \underbrace{2\varepsilon(f(x) - a(x, X))}_{\text{множители независимы}} + \varepsilon^2] = \end{aligned}$$

Bias-Variance Decomposition III

Преобразуем $Q(a, x)$

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X \mathbb{E}_\varepsilon [y(x, \varepsilon) - a(x, X)]^2 = \\ &= \mathbb{E}_X \mathbb{E}_\varepsilon [\underbrace{(f(x) - a(x, X))^2}_{\text{не зависит от } \varepsilon} + \underbrace{2\varepsilon(f(x) - a(x, X))}_{\text{множители независимы}} + \varepsilon^2] = \\ &= \mathbb{E}_X [f(x) - a(x, X)]^2 + 2 \underbrace{\mathbb{E}_\varepsilon[\varepsilon]}_{=0} \cdot \mathbb{E}_X [f(x) - a(x, X)] + \mathbb{E}_\varepsilon [\varepsilon^2] \end{aligned}$$

Bias-Variance Decomposition III

Преобразуем $Q(a, x)$

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X \mathbb{E}_\varepsilon [y(x, \varepsilon) - a(x, X)]^2 = \\ &= \mathbb{E}_X \mathbb{E}_\varepsilon [\underbrace{(f(x) - a(x, X))^2}_{\text{не зависит от } \varepsilon} + \underbrace{2\varepsilon(f(x) - a(x, X))}_{\text{множители независимы}} + \varepsilon^2] = \\ &= \mathbb{E}_X [f(x) - a(x, X)]^2 + 2 \underbrace{\mathbb{E}_\varepsilon[\varepsilon]}_{=0} \cdot \mathbb{E}_X [f(x) - a(x, X)] + \mathbb{E}_\varepsilon [\varepsilon^2] \\ &= \mathbb{E}_X [f(x) - a(x, X)]^2 + \sigma^2. \end{aligned}$$

Bias-Variance Decomposition IV

Продолжим преобразования

$$Q(a, x) = \mathbb{E}_X[f(x) - a(x, X)]^2 + \sigma^2 =$$

Bias-Variance Decomposition IV

Продолжим преобразования

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X [f(x) - a(x, X)]^2 + \sigma^2 = \\ &= \mathbb{E}_X [f(x) - \mathbb{E}_X a(x, X) + \mathbb{E}_X a(x, X) - a(x, X)]^2 + \sigma^2 = \end{aligned}$$

Bias-Variance Decomposition IV

Продолжим преобразования

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X[f(x) - a(x, X)]^2 + \sigma^2 = \\ &= \mathbb{E}_X[f(x) - \mathbb{E}_X a(x, X) + \mathbb{E}_X a(x, X) - a(x, X)]^2 + \sigma^2 = \\ &= \mathbb{E}_X \underbrace{[f(x) - \mathbb{E}_X a(x, X)]^2}_{\text{не зависит от } X} + \underbrace{\mathbb{E}_X[a(x, X) - \mathbb{E}_X a(x, X)]^2}_{\mathbb{V}_X[a(x, X)]} + \\ &\quad + 2\mathbb{E}_X \underbrace{[(f(x) - \mathbb{E}_X a(x, X)) \cdot (\mathbb{E}_X a(x, X) - a(x, X))]}_{\text{не зависит от } X} + \sigma^2 = \end{aligned}$$

Bias-Variance Decomposition IV

Продолжим преобразования

$$\begin{aligned} Q(a, x) &= \mathbb{E}_X[f(x) - a(x, X)]^2 + \sigma^2 = \\ &= \mathbb{E}_X[f(x) - \mathbb{E}_X a(x, X) + \mathbb{E}_X a(x, X) - a(x, X)]^2 + \sigma^2 = \\ &= \mathbb{E}_X \underbrace{[f(x) - \mathbb{E}_X a(x, X)]^2}_{\text{не зависит от } X} + \underbrace{\mathbb{E}_X [a(x, X) - \mathbb{E}_X a(x, X)]^2}_{\mathbb{V}_X[a(x, X)]} + \\ &\quad + 2\mathbb{E}_X \underbrace{[(f(x) - \mathbb{E}_X a(x, X)) \cdot (\mathbb{E}_X a(x, X) - a(x, X))]}_{\text{не зависит от } X} + \sigma^2 = \\ &= \underbrace{(f(x) - \mathbb{E}_X a(x, X))^2}_{bias_X^2[a(x, X)]} + \mathbb{V}_X[a(x, X)] + \sigma^2 + \\ &\quad + 2(f(x) - \mathbb{E}_X a(x, X)) \cdot \underbrace{(\mathbb{E}_X a(x, X) - \mathbb{E}_X a(x, X))}_{=0} \end{aligned}$$

Bias-Variance Decomposition V

Таким образом

$$Q(a, x) = \text{bias}_X^2[a(x, X)] + \mathbb{V}_X[a(x, X)] + \sigma^2. \quad (3)$$

- $\text{bias}_X[a(x, X)] = f(x) - \mathbb{E}_X a(x, X)$ – смещение “усреднённого” алгоритма $a(x)$ относительно целевой зависимости $f(x)$.
- $\mathbb{V}_X[a(x, X)]$ – разброс a в зависимости от X .
- Лосс по всем примерам $Q(a, x) \rightarrow Q(a) = \mathbb{E}_x[Q(a, x)]$. Выкладки проведены для квадратичной функции потерь.
- Для других видов лосса существуют общие формулы с похожим смыслом [2].

Bias-Variance trade-off I

- Чем больше параметров в модели, тем лучше она “запоминает” обучающую выборку: т.е. меньше смещение и больше разброс;
- Оптимальная сложность модели на пересечении смещения и разброса.

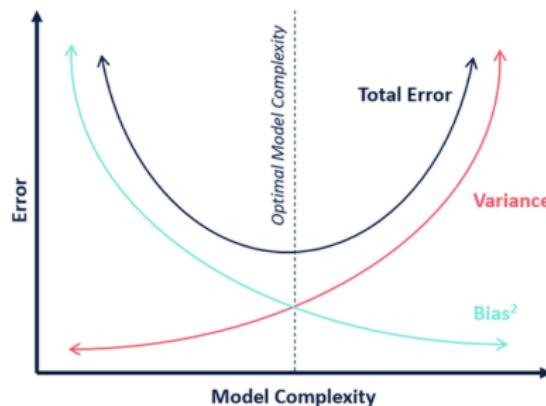


Figure: Компромисс смещения и разброса.

Bias-Variance trade-off II

- В современном обучении перепараметризованных моделей эта зависимость дополняется [1].
- Методы регуляризации (weight-decay) позволяют не проваливаться в локальные оптимумы функции потерь.
- На эту тему есть хороший [семинар](#) от Д.П. Ветрова.

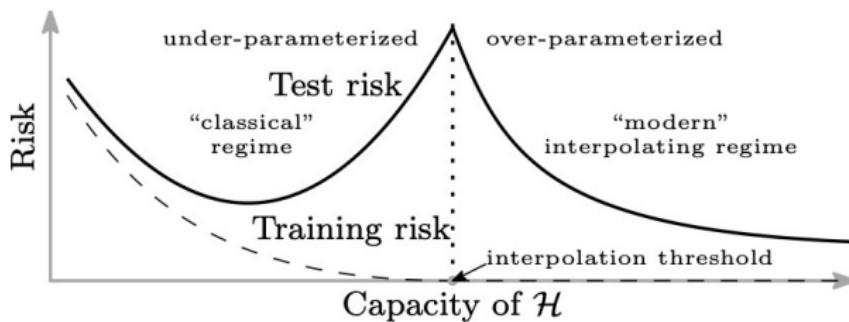


Figure: Двойной спуск

Бэггинг |

- Вернёмся к классическому bias-variance представлению (3)

$$Q(a, x) = \text{bias}_X^2[a(x, X)] + \mathbb{V}_X[a(x, X)] + \sigma^2.$$

- Пусть $|X| = n$. Получим X^1, X^2, \dots, X^k , выбирая n примеров из X равновероятно, с возвращением. Такая процедура называется **бутстрепом**.
- На соответствующих выборках получим k базовых моделей $b_i(x) = b(x, X^i)$. Ансамбль моделей $a(x)$ получается усреднением выходов $b_i(x)$.
- Процесс получения $a(x)$ называется **бэггингом (bootstrap aggregation)**.

Бэггинг ||

- Рассмотрим смещение и разброс ансамбля

$$\begin{aligned} bias_X[a(x, X)] &= f(x) - \mathbb{E}_X \left[\frac{1}{k} \sum_{i=1}^k b(x, X^i) \right] \\ &= f(x) - \frac{1}{k} \sum_{i=1}^k \mathbb{E}_X[b(x, X)] = f(x) - \mathbb{E}_X[b(x, X)] \end{aligned}$$

$$\begin{aligned} \mathbb{V}_X[a(x, X)] &= \mathbb{E}_X \left[\left(\frac{1}{k} \sum_{i=1}^k b(x, X^i) - \mathbb{E}_X \left[\frac{1}{k} \sum_{i=1}^k b(x, X^i) \right] \right)^2 \right] = \\ &= \frac{1}{k^2} \mathbb{E}_X \left[\sum_{i=1}^k (b(x, X^i) - \mathbb{E}_X b(x, X^i))^2 \right] \end{aligned}$$

Бэггинг III

- Продолжим преобразования

$$\begin{aligned}\mathbb{V}_X[a(x, X)] &= \frac{1}{k^2} \mathbb{E}_X \left[\sum_{i=1}^k (b(x, X^i) - \mathbb{E}_X b(x, X^i))^2 \right] = \\ &= \frac{1}{k^2} \sum_{i=1}^k \mathbb{V}_X b(x, X^i) + \frac{1}{k^2} \sum_{k_1 \neq k_2} \text{cov}(b(x, X_1^{k_1}), b(x, X_2^{k_2})).\end{aligned}$$

- Сделав предположение о нескоррелированности базовых моделей $b_i(x)$ мы получаем линейное убывание разброса при том же смещении.

Случайный Лес

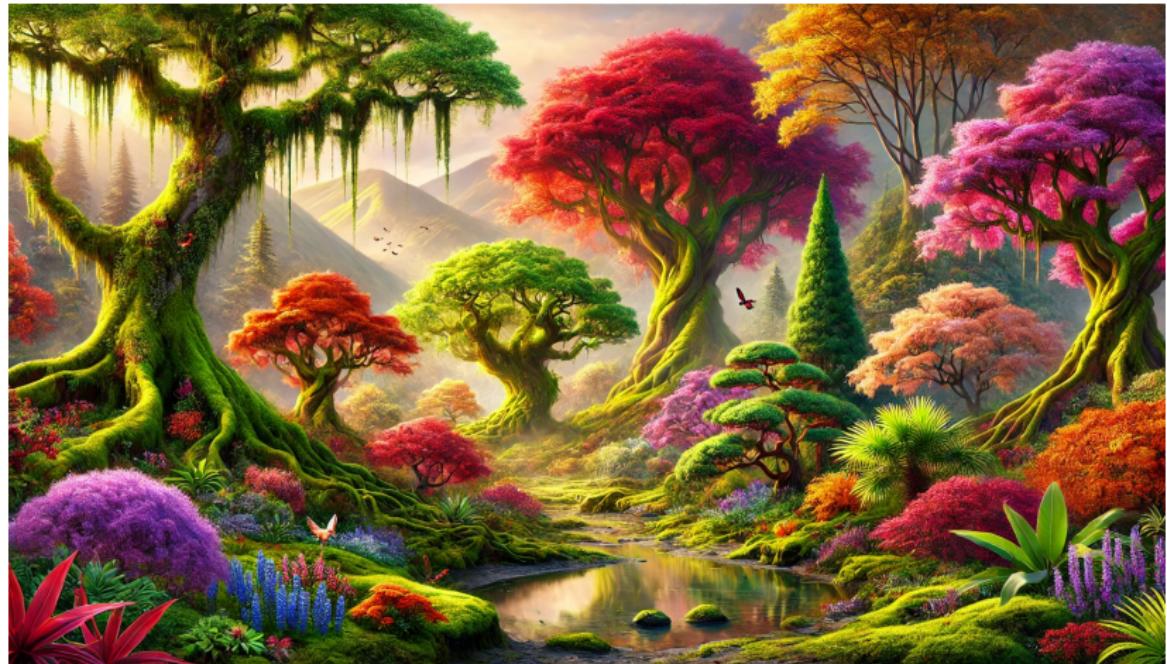


Figure: Случайный Лес

Напоминание: решающее дерево

- По обучающей выборке X ищем оптимальное условие $\mathbb{I}\{x_i < t\}$, $x_i \in \mathbb{R}^d$
- Правило разбивает X на X_ℓ и X_r .

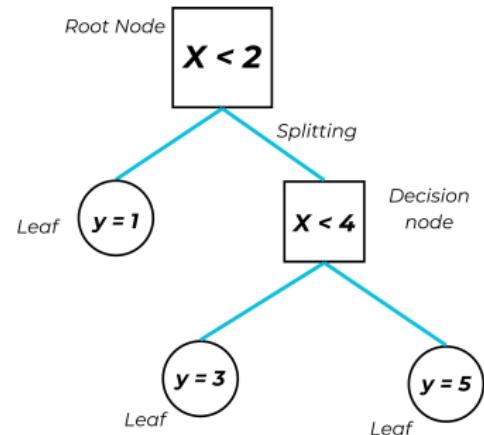


Figure: Решающее дерево

Случайный лес I

- Применим бэггинг над решающими деревьями:
 - Делаем бутстреп $X \rightarrow \{X^1, X^2, \dots, X^k\}$;
 - На каждой выборке строим решающее дерево b_i , ответы усредняем.
- Мы предполагали некоррелированность моделей b_i , в реальности это не так. *Как можно уменьшить корреляцию?*

Случайный лес I

- Применим бэггинг над решающими деревьями:
 - Делаем бутстреп $X \rightarrow \{X^1, X^2, \dots, X^k\}$;
 - На каждой выборке строим решающее дерево b_i , ответы усредняем.
- Мы предполагали некоррелированность моделей b_i , в реальности это не так. Как можно уменьшить корреляцию?
- Метод Случайных Подпространств (МСП) – строим правило $\mathbb{I}\{x'_i < t\}$ на основе подмножества признаков x_i : $x'_i \in \mathbb{R}^p$, $p < d$.
- Случайный лес – комбинация бэггинга и МСП над решающими деревьями.

Случайный лес II

- *Как выбрать глубину дерева?*

Случайный лес II

- Как выбрать глубину дерева?
- Чем больше глубина, тем меньше смещение и больше разброс.
Разброс компенсируем ансамблированием, поэтому выгодно брать глубокие решающие деревья.
- Как выбрать число r в МСП?

Случайный лес II

- Как выбрать глубину дерева?
- Чем больше глубина, тем меньше смещение и больше разброс.
Разброс компенсируем ансамблированием, поэтому выгодно брать глубокие решающие деревья.
- Как выбрать число p в МСП?
- p отражает trade-off между смещением базового алгоритма и вкладом от ансамблирования.
- Для регрессии рекомендуют $p = \frac{d}{3}$, для классификации $p = \sqrt{d}$.
- Как выбрать число деревьев?

Случайный лес II

- Как выбрать глубину дерева?
- Чем больше глубина, тем меньше смещение и больше разброс.
Разброс компенсируем ансамблированием, поэтому выгодно брать глубокие решающие деревья.
- Как выбрать число p в МСП?
- p отражает trade-off между смещением базового алгоритма и вкладом от ансамблирования.
- Для регрессии рекомендуют $p = \frac{d}{3}$, для классификации $p = \sqrt{d}$.
- Как выбрать число деревьев?
- В какой-то момент, корреляция алгоритмов нивелирует эффект уменьшения разброса. Можно строить график, можно ориентироваться от общей сложности ансамбля.

Стекинг



Figure: Риторический вопрос

Стекинг



Figure: Риторический вопрос

- Почему используем усреднение или голосование базовых алгоритмов?

Стекинг



Figure: Риторический вопрос

- Почему используем усреднение или голосование базовых алгоритмов?
- Давайте доверим ансамблирование очередному алгоритму.

Блендинг

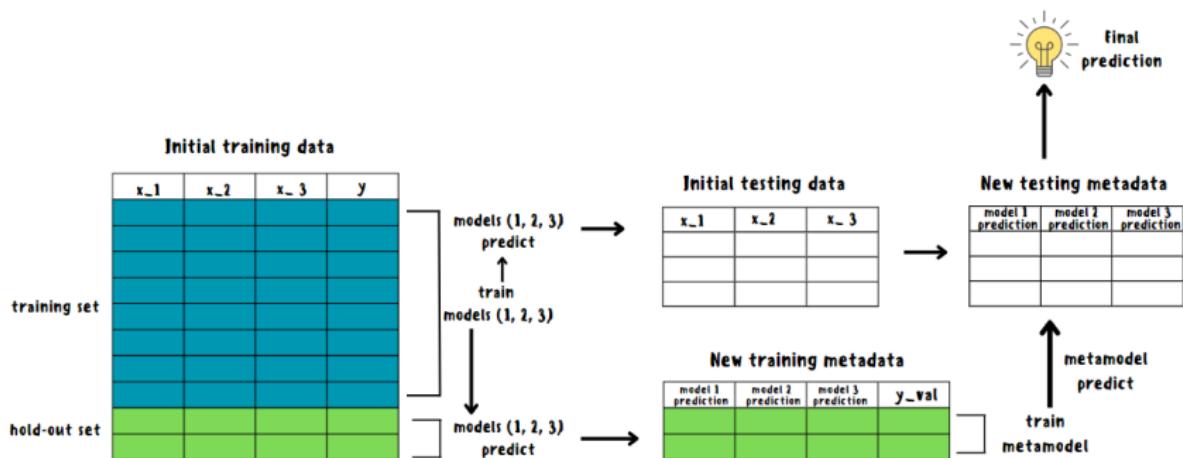


Figure: схема блендинга

Стекинг II

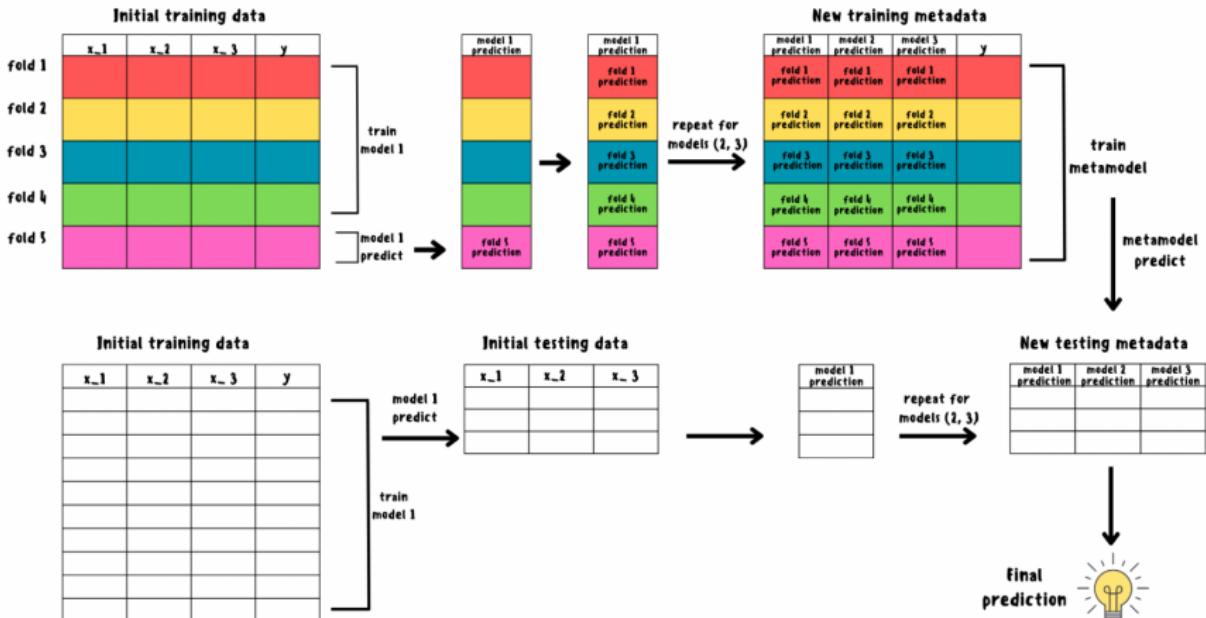


Figure: схема стекинга

References I

-  [Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandala.
Reconciling modern machine learning practice and the bias-variance
trade-off, 2019.](#)
[arXiv preprint, version 2, 10 Sep 2019.](#)
-  [Pedro Domingos.
A unified bias-variance decomposition.
In Proceedings of 17th international conference on machine learning,
pages 231–238. Morgan Kaufmann Stanford, 2000.](#)