
Занятие № 9

Ансамблирование моделей

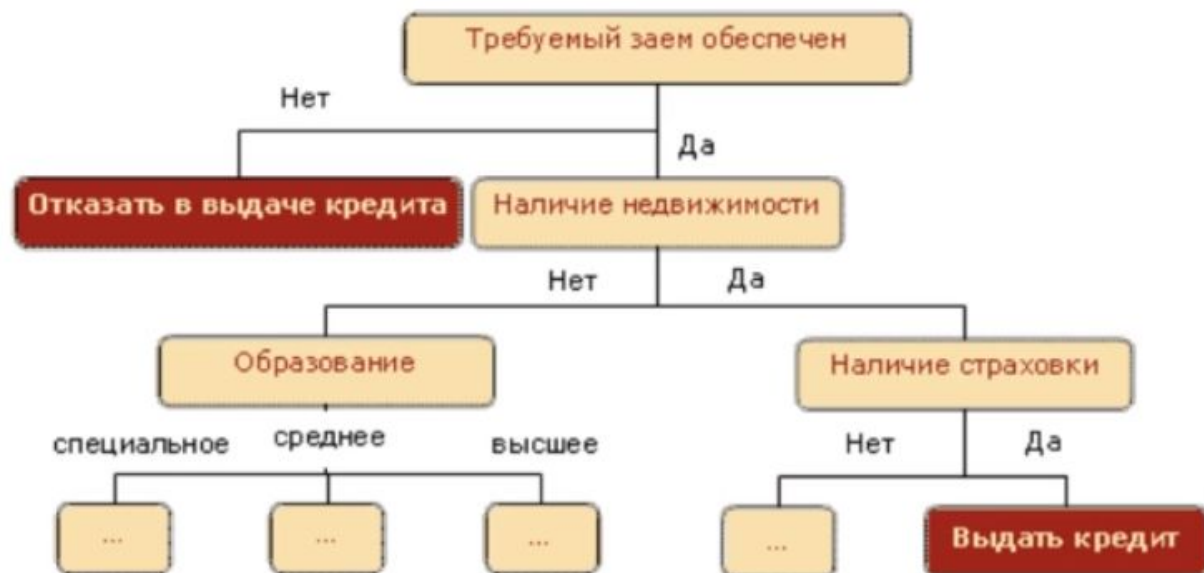


Содержание

- 1 Введение
- 2 Что не так с деревьями?
- 3 Ансамбли
- 4 Бутстреп. Бэггинг. Случайный лес. Блэндинг. Стекинг. Бустинг
- 5 Практика.



Введение



Что не так с деревьями?

1. Острая проблема переобучения
2. Неустойчивость
3. Не учитывает нелинейные зависимости или даже простые линейные, которые идут не по осям координат (f.e., представьте дерево для классификатора вида $y > x$)
4. Чувствителен к несбалансированным классам
5. Хорошо интерполирует, плохо экстраполирует



«Мудрость толпы»



Фрэнсис Гальтон «Мудрость толпы»

Фрэнсис Гальтон в 1906 году посетил рынок, где проводилась некая лотерея для крестьян.

Их собралось около 800 человек, и они пытались угадать вес быка, который стоял перед ними. Бык весил 1198 фунтов. Ни один крестьянин не угадал точный вес быка, но если посчитать среднее от их предсказаний, то получим 1197 фунтов.

Эту идею уменьшения ошибки применили и в машинном обучении.



Теорема Кондорсе «о жюри присяжных» (1784).

Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице. Если же вероятность быть правым у каждого из членов жюри меньше 0.5, то вероятность принятия правильного решения присяжными в целом монотонно уменьшается и стремится к нулю с увеличением количества присяжных.

N — количество присяжных

p — вероятность правильного решения присяжного

μ — вероятность правильного решения всего жюри

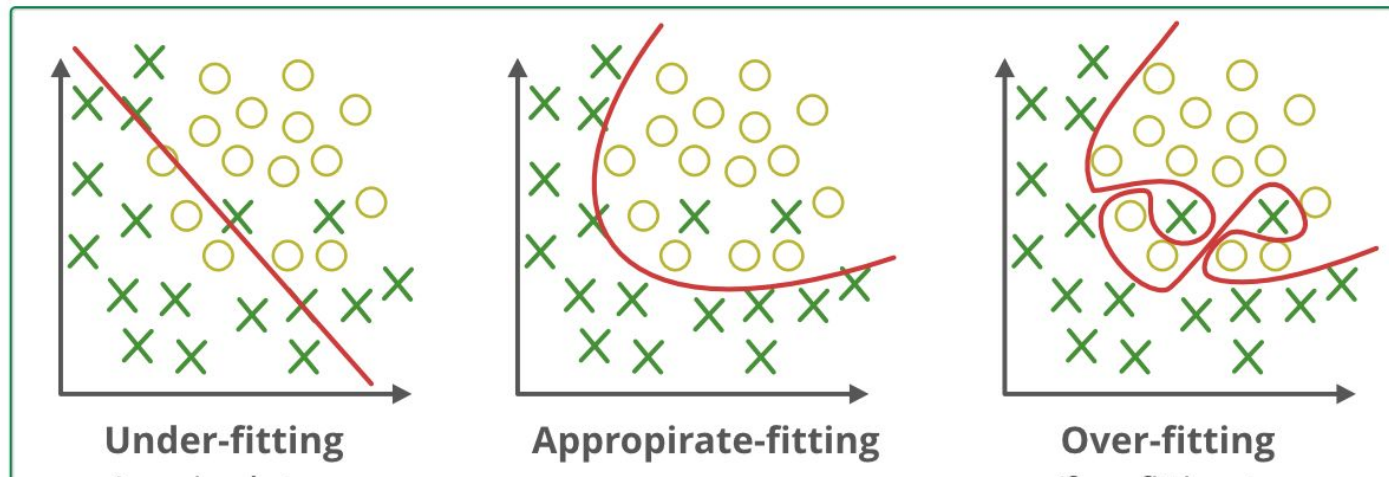
m — минимальное большинство членов жюри, $m = \text{floor}(N/2) + 1$

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 $N \rightarrow \infty$, то $\mu \rightarrow 1$



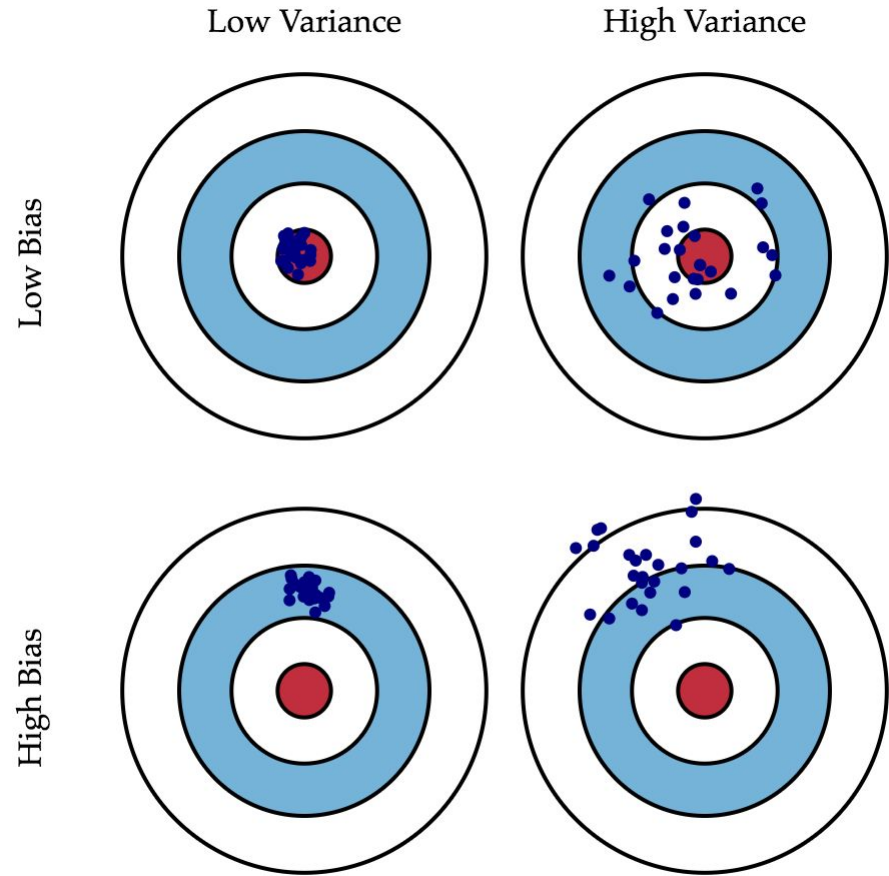
переобучение (overfitting) – явление, когда ошибка на тестовой выборке заметно больше ошибки на обучающей.

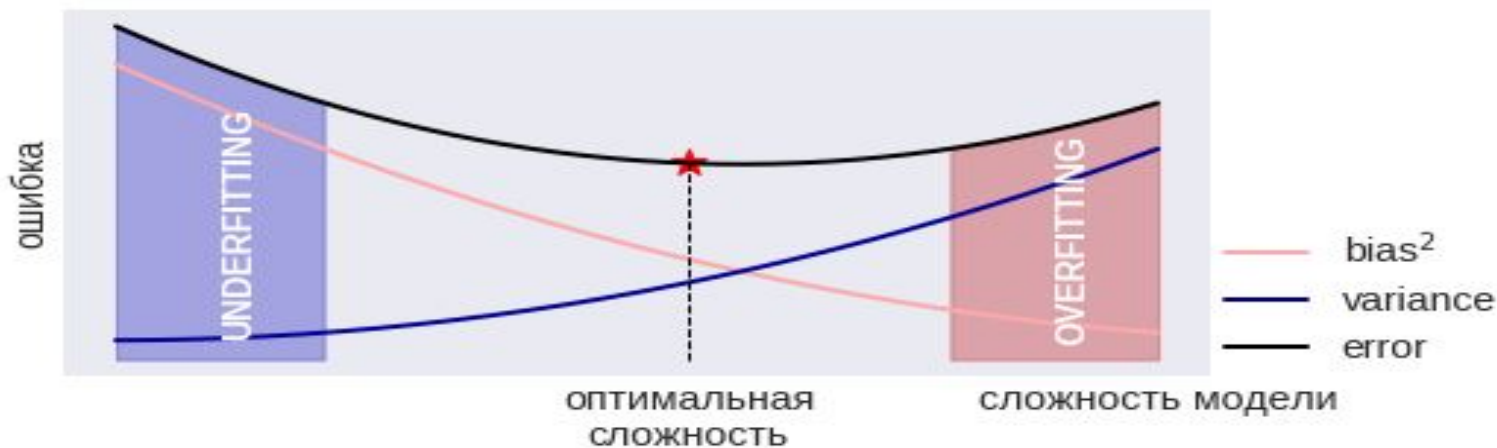
недообучение (underfitting) – явление, когда ошибка на обучающей выборке достаточно большая, часто говорят «не удаётся настроиться на выборку».

сложность (complexity) модели алгоритмов – оценивает, насколько разнообразно семейство алгоритмов в модели с точки зрения их функциональных свойств (например, способности настраиваться на выборки). Повышение сложности (т.е. использование более сложных моделей) решает проблему недообучения и вызывает переобучение.

разброс (variance) - дисперсию ответов алгоритмов характеризует разнообразие алгоритмов (из-за случайности обучающей выборки, в том числе шума, и стохастической природы настройки)

смещением (bias) – матожидание разности между истинным ответом и выданным алгоритмом, характеризует способность модели алгоритмов настраиваться на целевую зависимость.



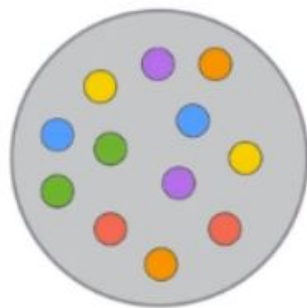


Для простых моделей характерно недообучение (они слишком простые, не могут описать целевую зависимость и имеют большое смещение), для сложных – переобучение (алгоритмов в модели слишком много, при настройке мы выбираем ту, которая хорошо описывает обучающую выборку, но из-за сильного разброса она может допускать большую ошибку на тесте).



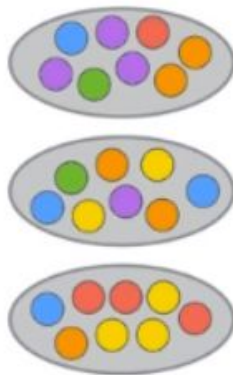
Бутстреп

Исходная выборка



Статистика по
выборке

Бутстрэп выборки



Статистики по
бутстрэп выборкам

Статистика 1

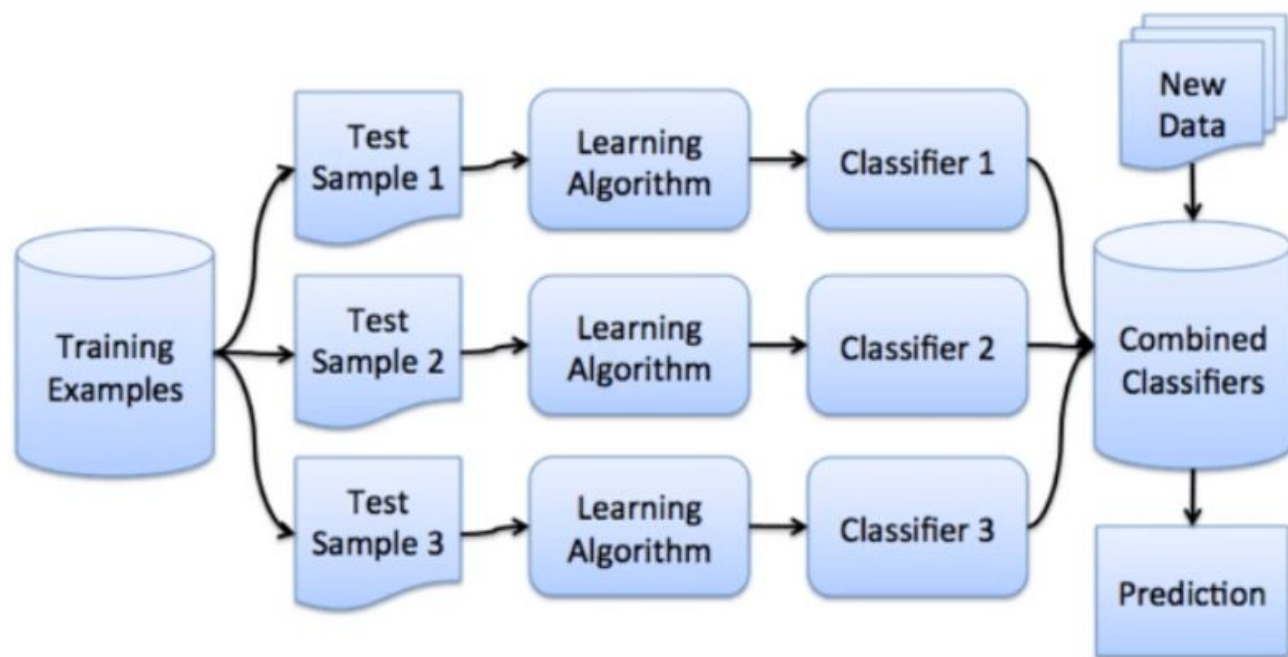
Статистика 2

Статистика 3

Бутстрэп
распределение



БЭГИНГ



Бэгинг

```
from sklearn.ensemble import BaggingRegressor
```

base_estimator object or None – модель регрессии из sklearn (по умолчанию деревья решений)

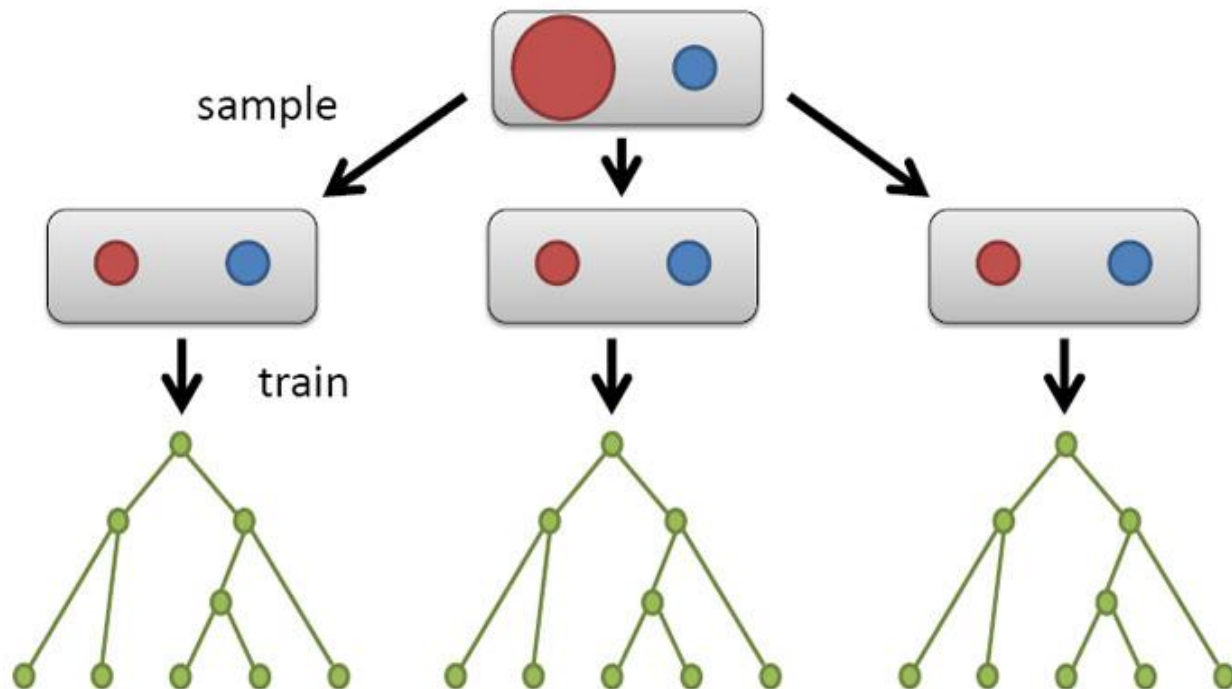
n_estimators – количество моделей

max_samples int or float, optional (default=1.0) Количество сэмплов для обучения

max_features int or float, optional (default=1.0) Количество признаков для обучения



Случайный лес



Случайный лес

Бэггинг + случайные подпространства = случайный лес

Случайный лес - вариация беггинга над деревьями, дающая даже лучшие результаты: Как и в беггинге, мы создаем ансамбль деревьев решений, используя выборки из обучающего набора. Однако при построении каждого дерева каждый раз, когда производится расщепление, признак выбирается из случайной выборки размера m из всех признаков. Новая случайная выборка признаков формируется для каждого отдельного дерева в каждом отдельном расщеплении. Для классификации m обычно выбирается как квадратный корень из p . Для регрессии m обычно выбирается где-то между $p/3$ и p .

Случайный лес

Случайные подпространства

Предположим, что в наборе данных есть один очень сильный признак. При использовании беггинга большая часть деревьев будет использовать этот признак в качестве первого, по которому производится деление, в результате чего образуется ансамбль похожих деревьев, которые сильно коррелированы. Усреднение высокоррелированных величин не приводит к значительному уменьшению дисперсии (что является целью беггинга). Случайно исключая признаки из каждого расщепления, Random Forest «декоррелирует» деревья, так что процесс усреднения может уменьшить дисперсию результирующей модели.

Случайный лес

```
from sklearn.ensemble import DecisionTreeRegressor
```

criterion – метод оценки ошибки (MSE по умолчанию)

splitter str – метод разделения "better" или "random"

n_estimators – количество моделей

max_depth – максимальная глубина деревьев

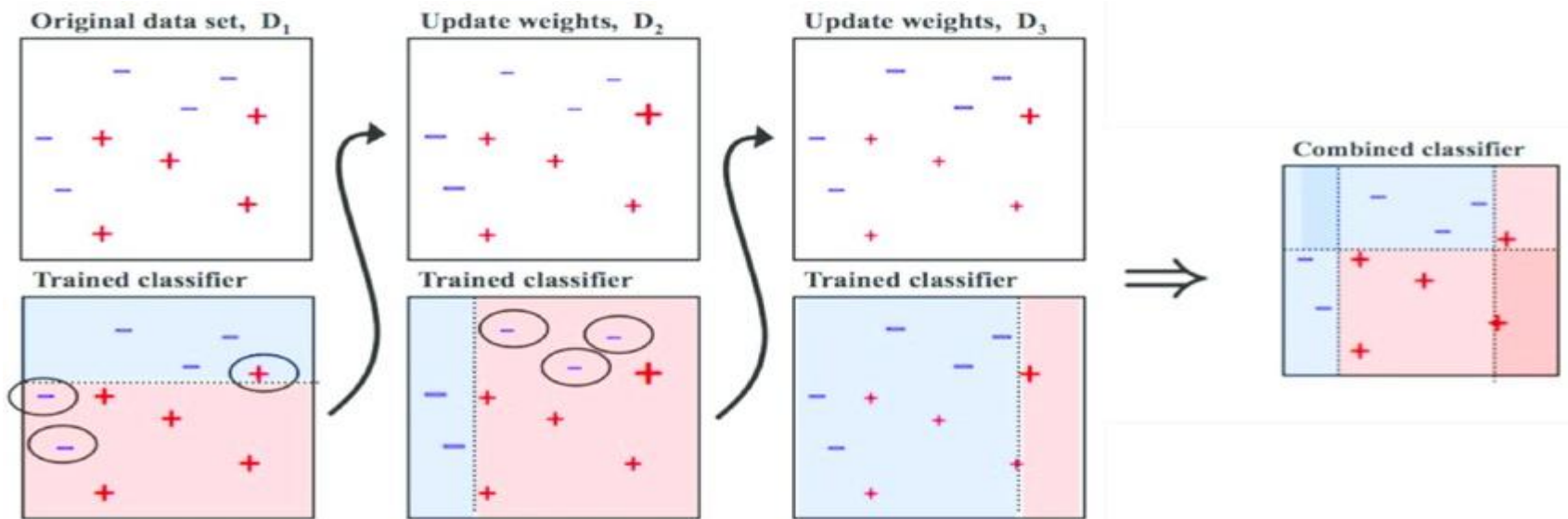
max_samples int or float, optional (default=1.0) Количество сэмплов для обучения

max_features int or float, optional (default=1.0) Количество признаков для обучения

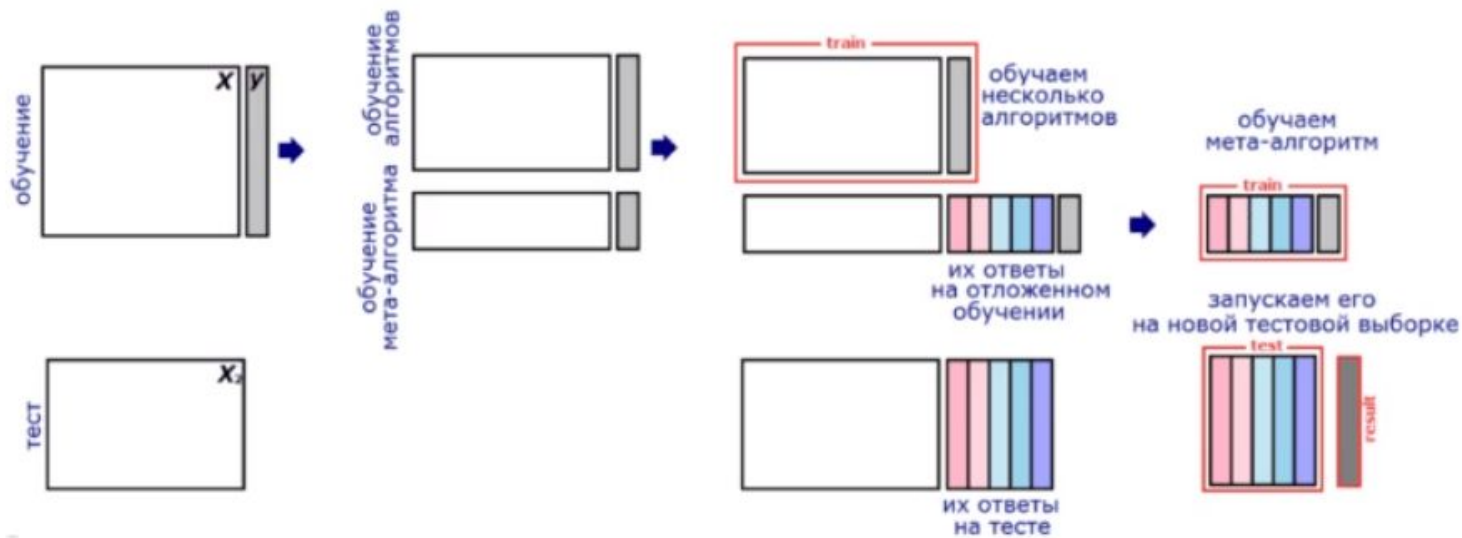


Бустинг

Бустинг (англ. boosting — улучшение) — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов.



Блендинг



Стекинг

