Разложение ошибки и бэггинг

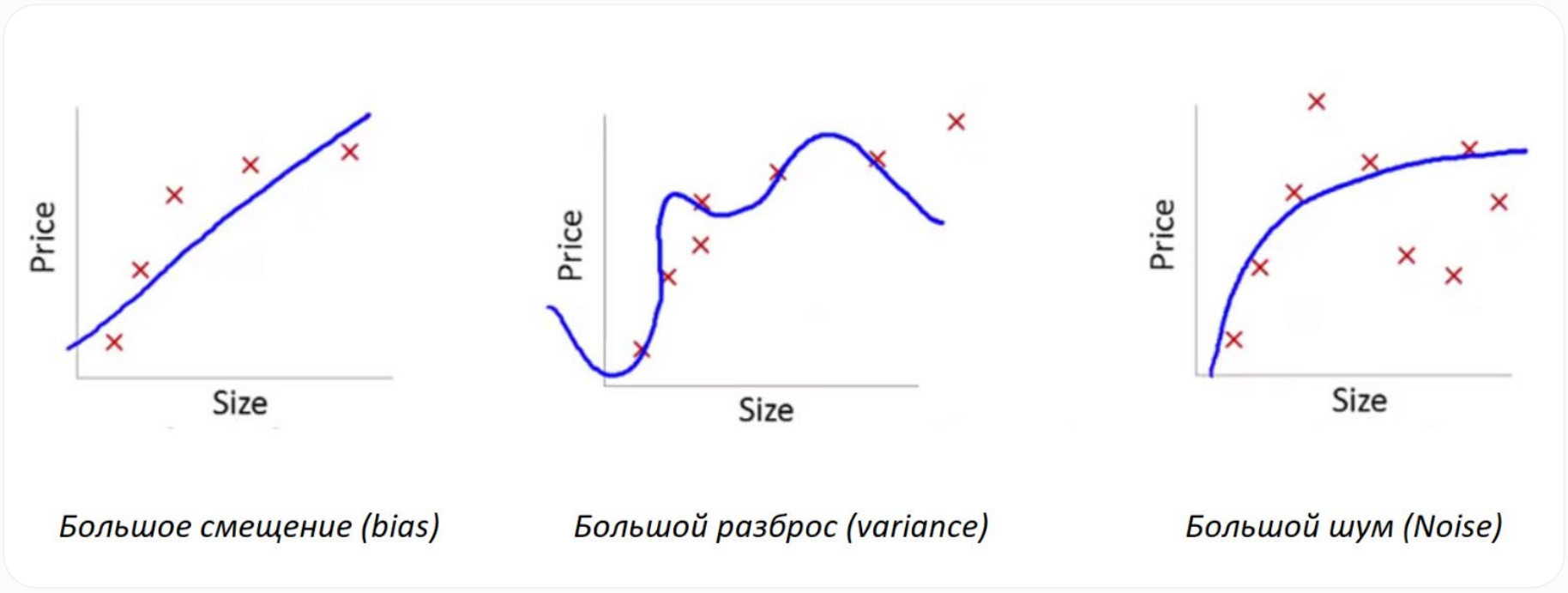
Skillbox

Елена Кантонистова

Почему модель плохо работает?

- Модель переобучена, т. е. имеет большой разброс
- Модель плохо предсказывает целевую переменную, то есть плохо приближает зависимость целевой переменной от признаков — имеет большое смещение
- В данных может быть много неточностей (или шумов)

Почему модель плохо работает?



Большое смещение (bias) *Большой разброс (variance)* *Большой шум (Noise)*

Изображение: Пример

Ошибка модели

Ошибка модели на тестовой выборке:

$$Error = Bias + Variance + Noise$$

- Bias — смещение
- Variance — разброс
- Noise — шум

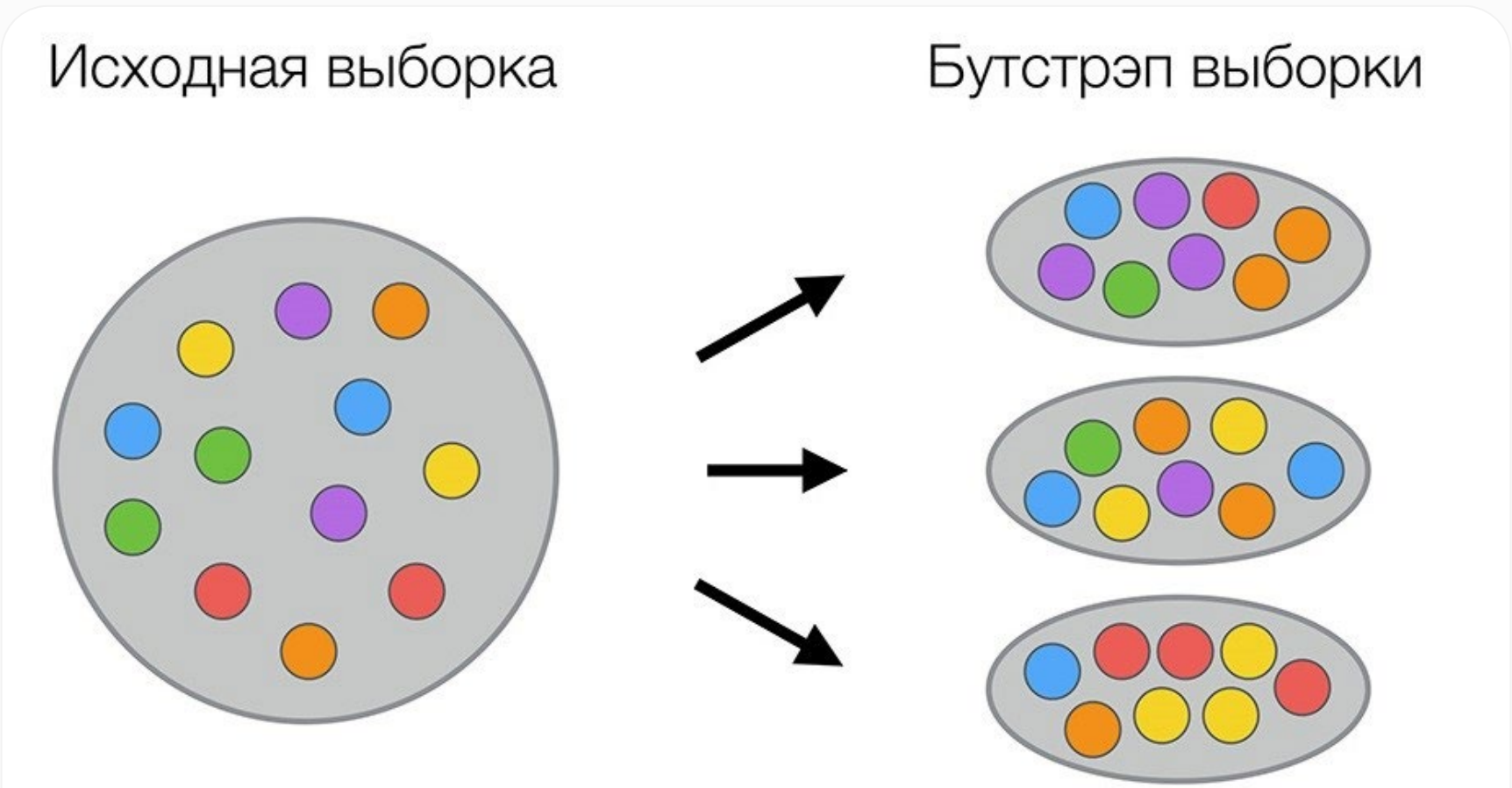
Все слагаемые неотрицательные.

Решающее дерево

- Bias (смещение) — низкое
- Variance (разброс) — высокий

Бутстреп (bootstrap)

Возьмите из выборки несколько объектов с возвращением
(т. е. в новой выборке будут повторяющиеся объекты).



Изображение: Бутстреп

Бэггинг (bagging)

Рассмотрите задачу регрессии:

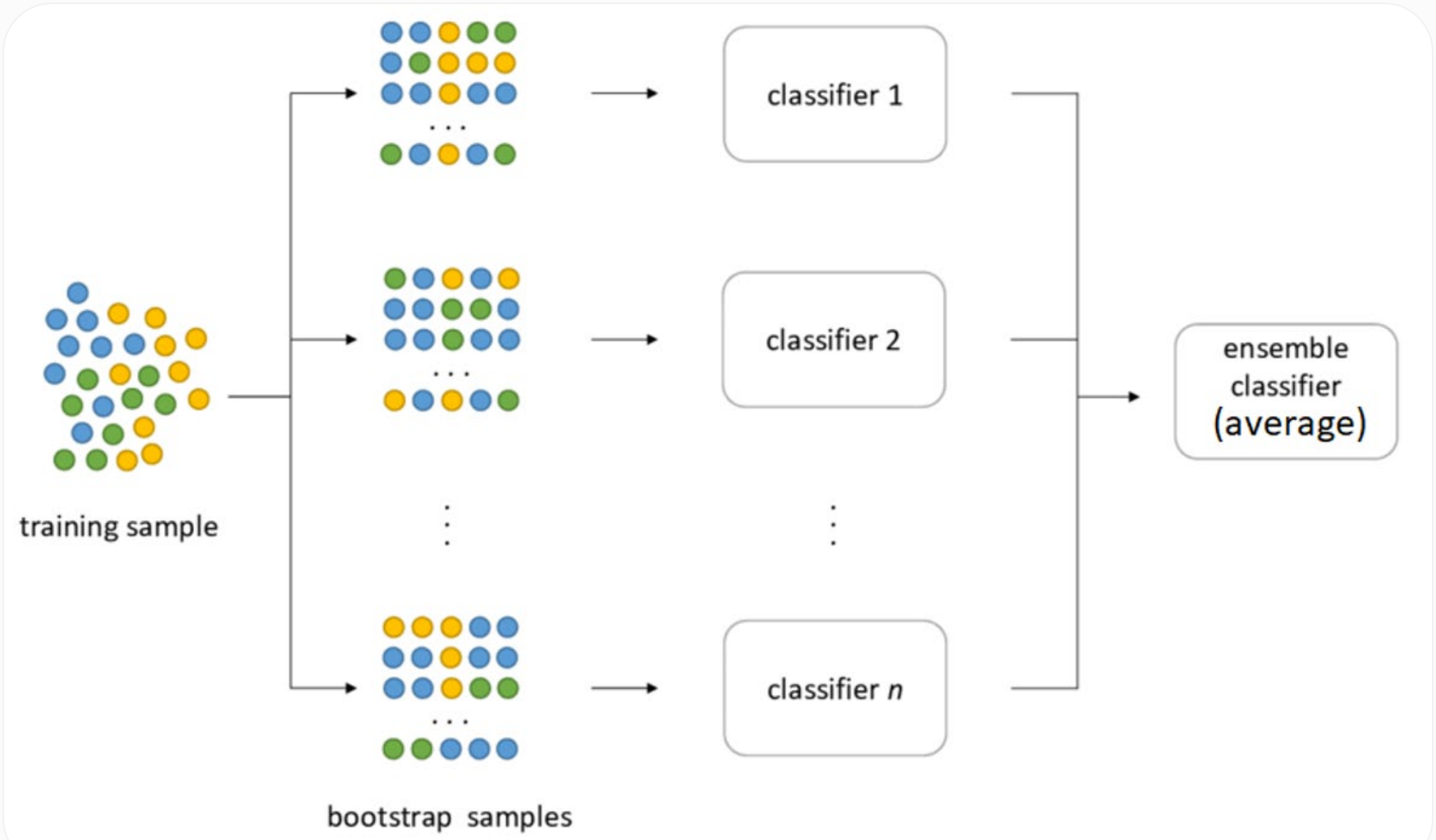
- с помощью бутстрэпа вы получили выборки $X_1, \ldots, X_N$
- обучите по каждой из них дерево — получите базовые алгоритмы $b_1(x), \ldots, b_N(x)$

Постройте новую функцию регрессии:

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

Бэггинг (bagging)

Bagging (bootstrap aggregation) в задаче регрессии.



Изображение: Бэггинг

Вы узнали

- Из чего складывается ошибка модели
- Об алгоритме под названием «бэггинг» над решающими деревьями, который в общем случае имеет меньшую ошибку, чем решающее дерево