



Методы ансамблирования

Ансамбль

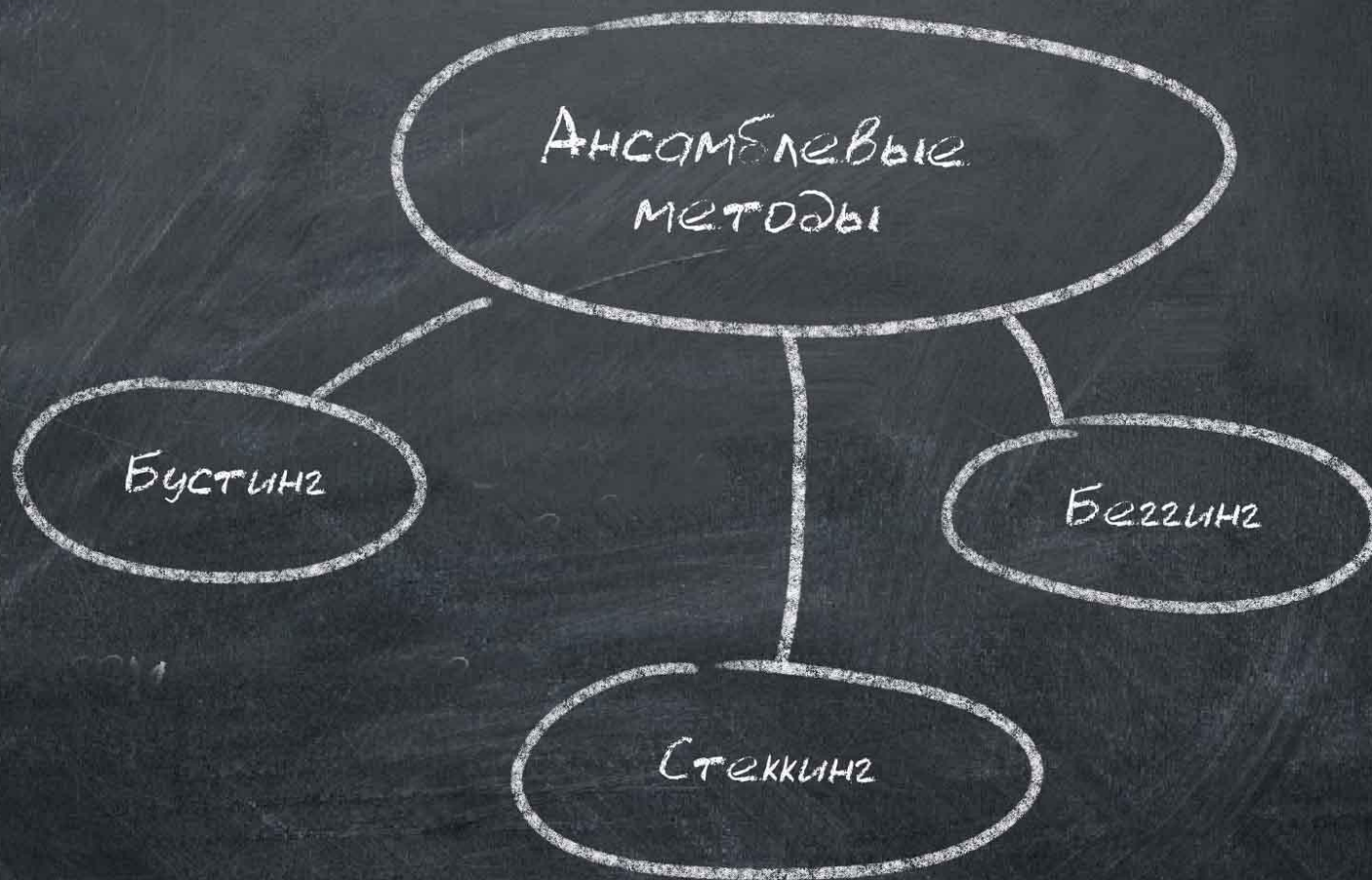
- **Ансамбль** — алгоритм машинного обучения состоящий из простых базовых алгоритмов.

$$a(x) = b(b_1(x), \dots, b_n(x))$$

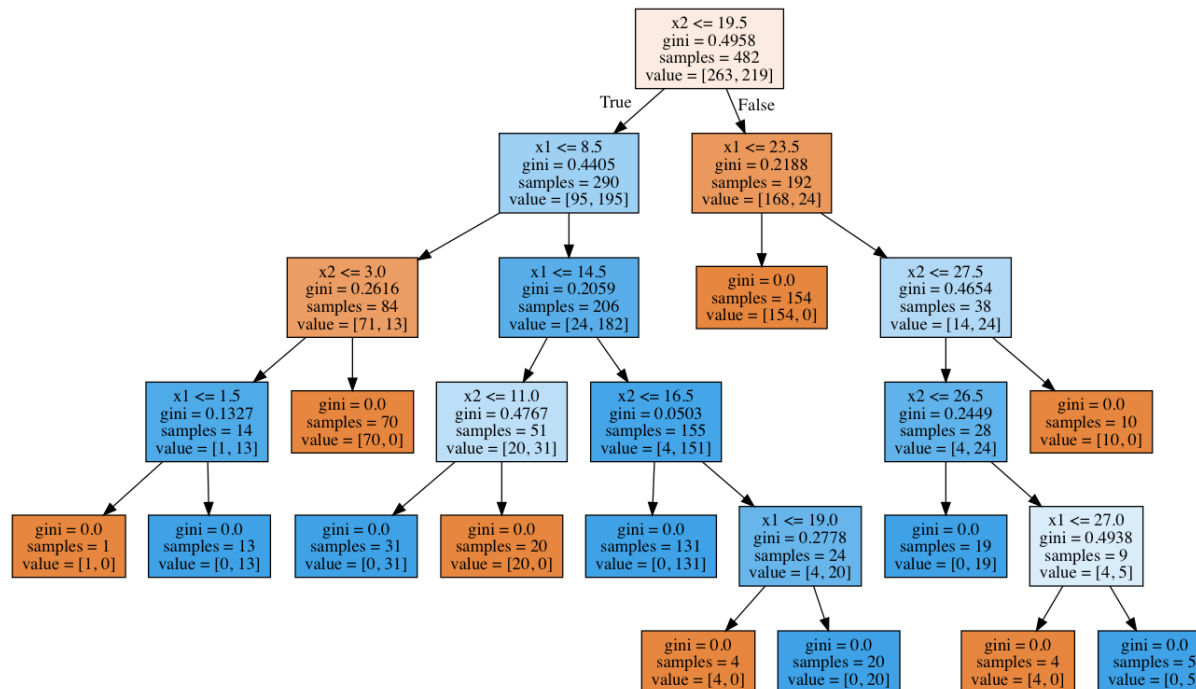
Мета-алгоритм b



Ансамбль



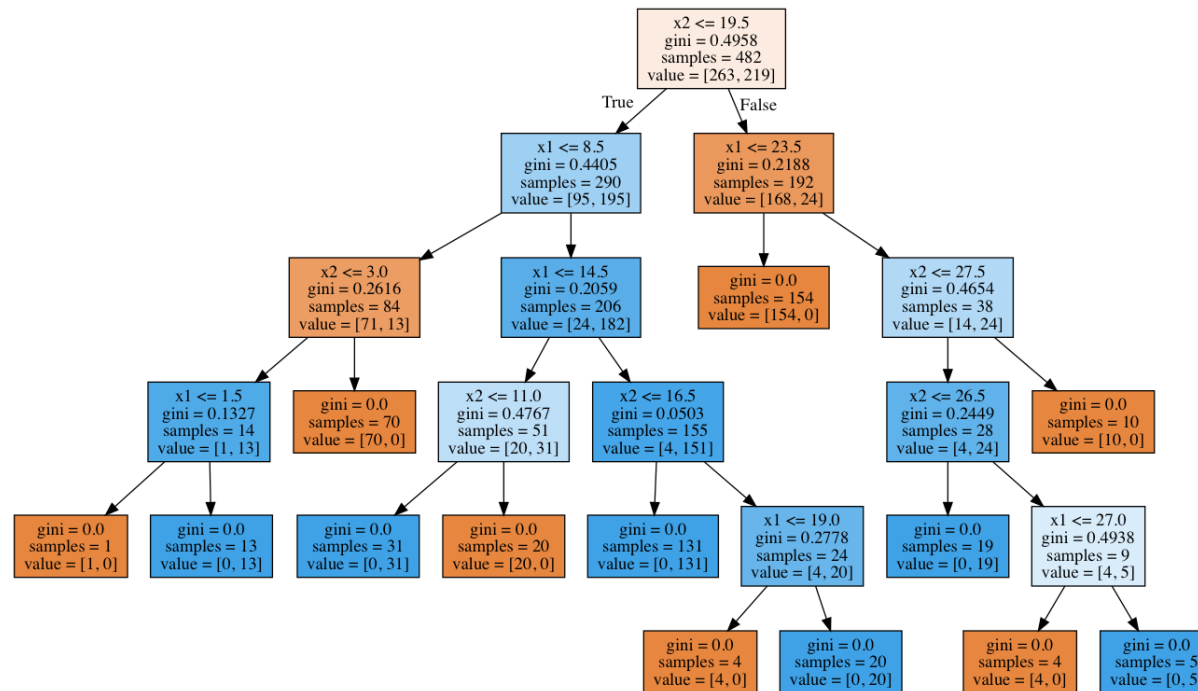
Дерево решений



Разбиение происходит путем сравнения значения одного из признаков с некоторым порогом. Параметры в условии $x^j \leq t$ подбираются так, чтобы минимизировать критерий ошибки:

$$Q(X_m, j, t) \rightarrow \min_{j, t}$$

Дерево решений



В задаче регрессии прогноз в листах вычисляется как среднее: $a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$

В задаче классификации наиболее популярный класс: $a_m = \operatorname{argmax}_{y \in Y} \sum_{i \in X_m} [y_i = y]$

Критерий ошибки



$$Q(X_m, j, t) = \frac{X_l}{X_m} H(X_l) + \frac{X_r}{X_m} H(X_r),$$

$H(X)$ - критерий информативности

$H(X)$ для регрессии:

$$H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \bar{y})^2$$

$H(X)$ для классификации:

Джини:
$$H(X) = \sum_{k=1}^K p_k (1 - p_k),$$

Энтропия:
$$H(X) = - \sum_{k=1}^K p_k \ln(p_k),$$

где
$$p_k = \frac{1}{X} \sum_{i \in X} [y_i = k]$$

Критерий останова



Разбиение вершин останавливается если выполняется какое либо из условий:

- 1) все критерии в вершине относятся к одному классу;
- 2) количество объектов в вершине $\leq n$;
- 3) дерево достигло определенной глубины.

Недостатки решающего дерева



- Сильно переобучается;
- Сильно меняется при небольшом изменении выборки.

Бэггинг

Бэггинг позволяет получить ансамблевую модель с меньшим разбросом прогнозов, чем у ее базовых алгоритмов при этом сохраняя ту же точность.

Идея бэггинга:

- 1) в качестве базовой модели выбирается несмещенная модель с высоким разбросом (например глубокое решающее дерево);
- 2) с помощью бутстрепа формируется обучающая подвыборка для i -го базового алгоритма;
- 3) базовые алгоритмы параллельно обучаются;
- 4) полученные прогнозы агрегируются согласно функции некоего мета-алгоритма.

Бэггинг

Почему ансамбль на основе бэггинга имеет ту же точность, но меньший разброс прогнозов чем ее базовые алгоритмы:

$$\Theta = \frac{1}{n} (\theta_1 + \dots + \theta_n)$$

$$E \Theta = \frac{1}{n} (E \theta_1 + \dots + E \theta_n) = E \theta_i$$

$$D \Theta = \frac{1}{n^2} (D \theta_1 + \dots + D \theta_n) = \frac{D \theta_i}{n}$$

Бэггинг

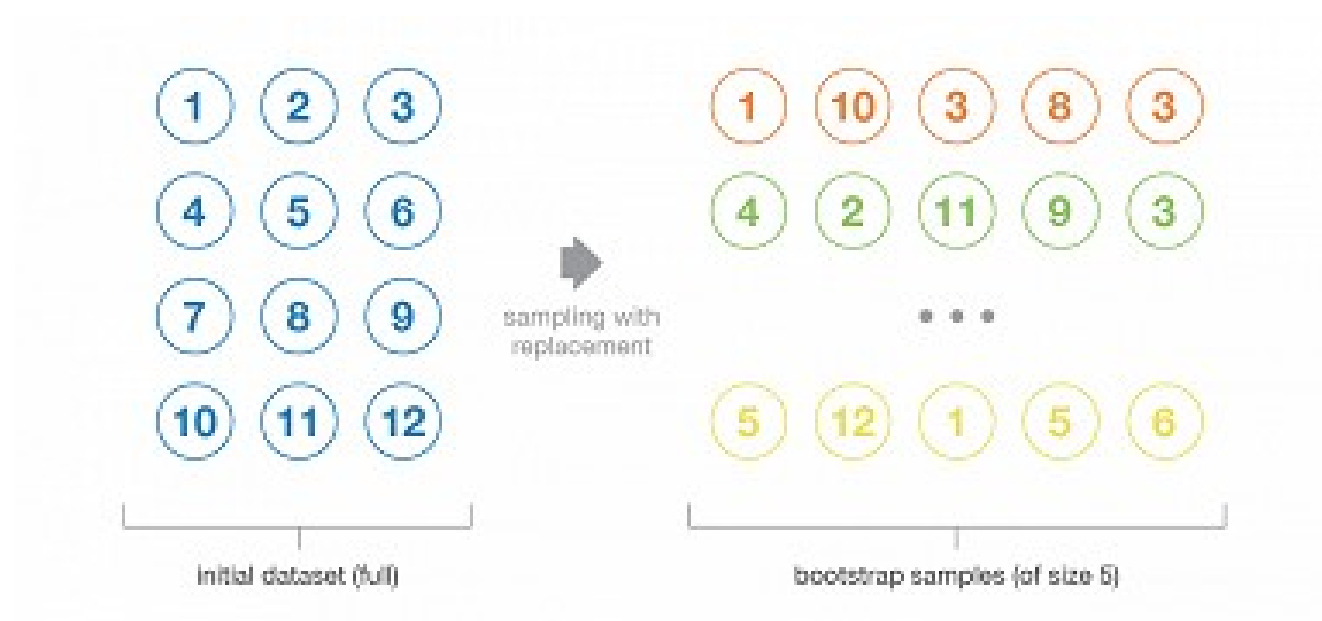
Для задач регрессии и классификации агрегирование происходит путем усреднения полученных прогнозов базовых моделей и путем голосования большинства соответственно:

$$a(x) = \frac{1}{n} (b_1(x), \dots, b_n(x))$$

$$a(x) = \text{mode}(b_1(x), \dots, b_n(x))$$

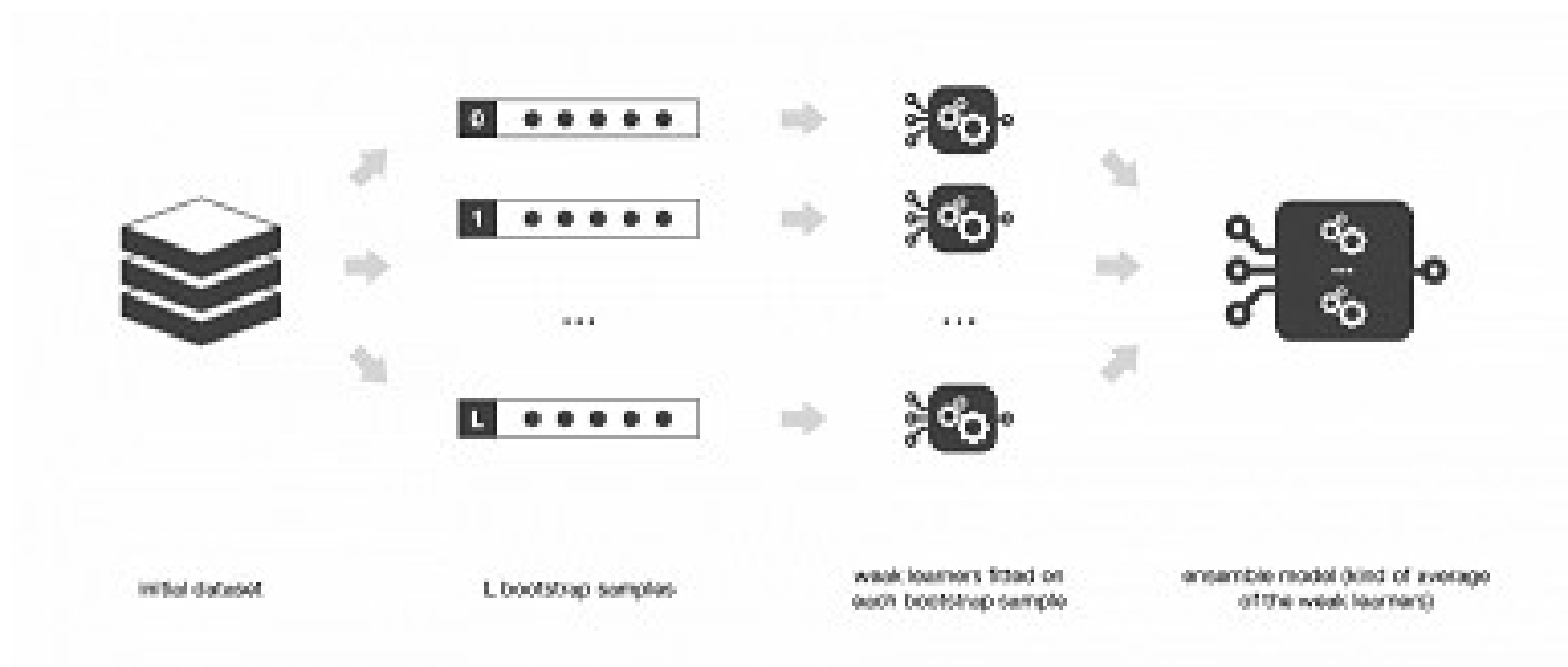
Бэггинг

Бутстрэп:



Бэггинг

Параллелизм:



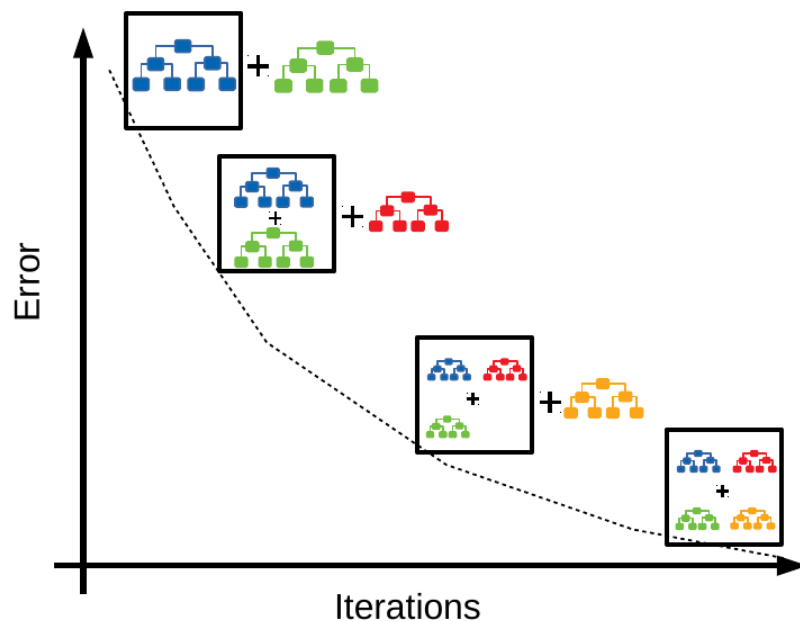
Бэггинг

Бэггинг + деревья решений

RANDOM, FORREST, RANDOM!



Градиентный бустинг



Идея бустинга:

- 1) базовые алгоритмы строятся последовательно;
- 2) каждый последующий базовый алгоритм исправляет ошибки построенной композиции.

Градиентный бустинг

Идея бустинга:

- 1) базовые алгоритмы строятся последовательно;
- 2) каждый последующий базовый алгоритм исправляет ошибки построенной композиции.

Функции потерь при обучении базовых алгоритмов (для одного объекта):

- 1) в задаче регрессии: $L(y, z) = (y - z)^2$
- 2) в задаче классификации: $L(y, z) = \log(1 + \exp(-yz))$,

где y — истинный ответ, z — прогноз алгоритма на некотором объекте.

Градиентный бустинг

Обучение базовых алгоритмов

Каждый последующий базовый алгоритм обучается на векторе антиградиента s :

$$s = -\nabla F = \begin{pmatrix} -L'_z(y_1, a_{N-1}(x_1)) \\ \dots \\ -L'_z(y_l, a_{N-1}(x_l)) \end{pmatrix}$$

Подходы к борьбе с переобучением:

- 1) Стохастический градиентный бустинг;
- 2) Сокращение размера шага: $a_N(x) = a_{N-1}(x) + \eta b_N(x)$, где $\eta \in (0, 1]$