Лекция 9
Ансамбли моделей

Макаренко В.А., Габдуллин Р.А.

МГУ им. М.В. Ломоносова

10 марта 2023

Ансамбли моделей

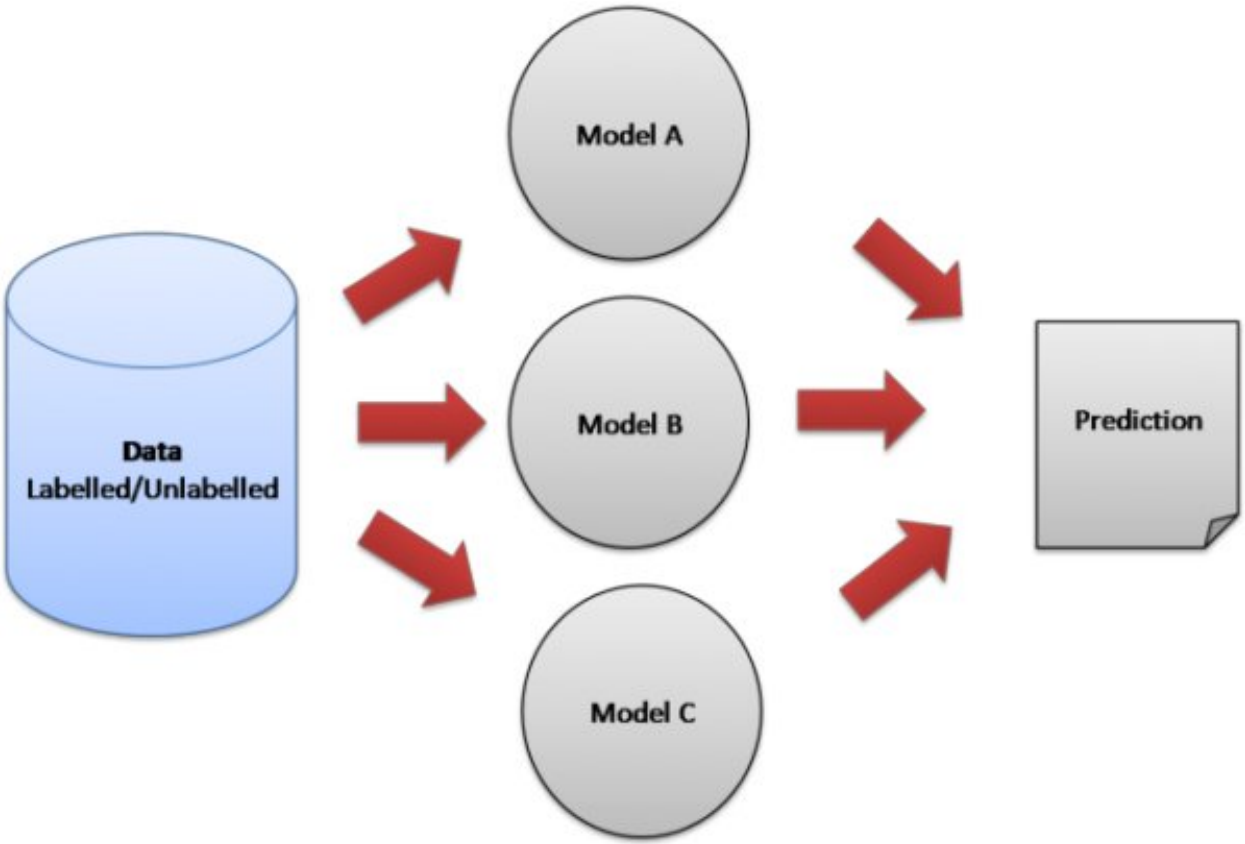


Рис.: Источник: kdnuggets.com

Для построения прогноза можно использовать сразу несколько моделей.

Бэггинг (Bootstrap AGGregatING)

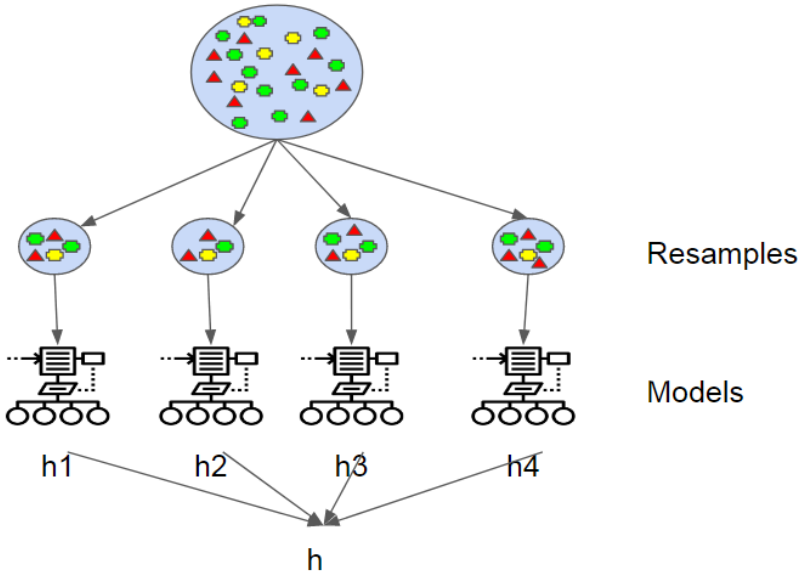


Рис.: Источник: hackernoon.com

Пусть $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ – обучающая выборка.

- Строим N выборок длины ℓ, сэмплируя с возвращением данные из тренировочной выборки (по факту генерируем независимые выборки из эмпирического распределения).
- Обучаем модель на каждой из выборок.
- Усредняем прогнозы (для регрессии) / берем моду (для классификации).

Бэггинг. Мотивация

Обозначения:

- Тренировочная выборка: $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ – н.о.р.с.в. с плотностью $p_{X,Y}$.
- $\hat{f}$ – функция, полученная в результате обучения. Является случайной функцией, так как зависит от тренировочной выборки.
- Ответ для нового объекта с признаковым описанием x: $y_x \sim p_{Y|X=x}$.

Дилемма смещения–дисперсии (bias-variance tradeoff):

$$\mathbb{E}(\hat{f}(x) - y_x)^2 = \mathbb{E}(\hat{f}(x) - \mathbb{E}\hat{f}(x) + \mathbb{E}\hat{f}(x) - \mathbb{E}y_x + \mathbb{E}y_x - y_x)^2 =$$

$$= \underbrace{\mathbb{D}\hat{f}(x)}_{\text{variance}} + \underbrace{(\mathbb{E}\hat{f}(x) - \mathbb{E}y_x)^2}_{\text{bias}^2} + \underbrace{\mathbb{D}y_x}_{\sigma^2}.$$

Бэггинг. Мотивация

Пусть $X_1, \ldots, X_n$ – случайные величины. Имеем

$$\mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathbb{E}X_1 + \ldots + \mathbb{E}X_n}{n},$$

$$\mathbb{D}\left(\frac{X_1 + \ldots X_n}{n}\right) = \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathbb{D}X_k + 2\sum_{1\leqslant k<j<n}\operatorname{cov}(X_k, X_j)\right).$$

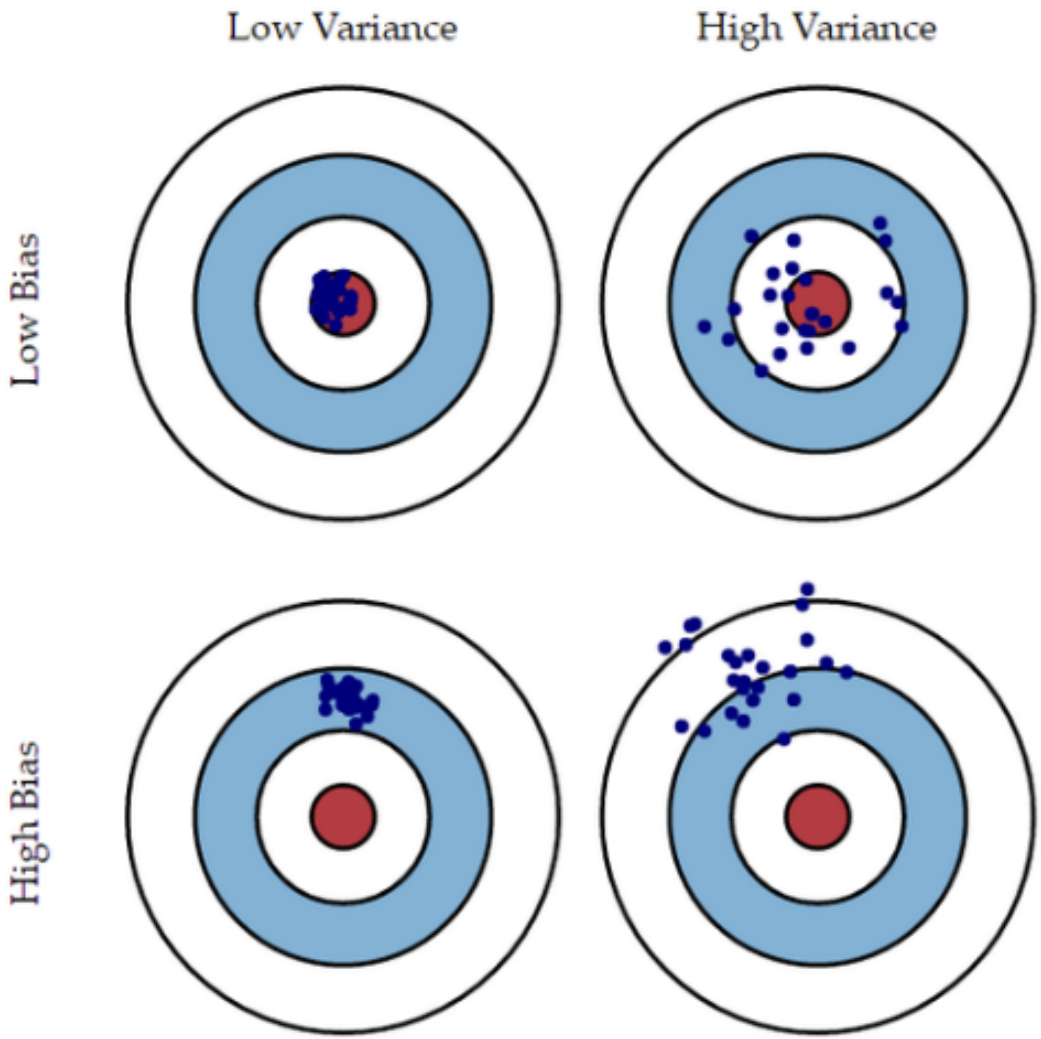
Если $X_1, \ldots, X_n$ – некоррелированные и одинаково распределенные, то

$$\mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \mathbb{E}X_1,$$

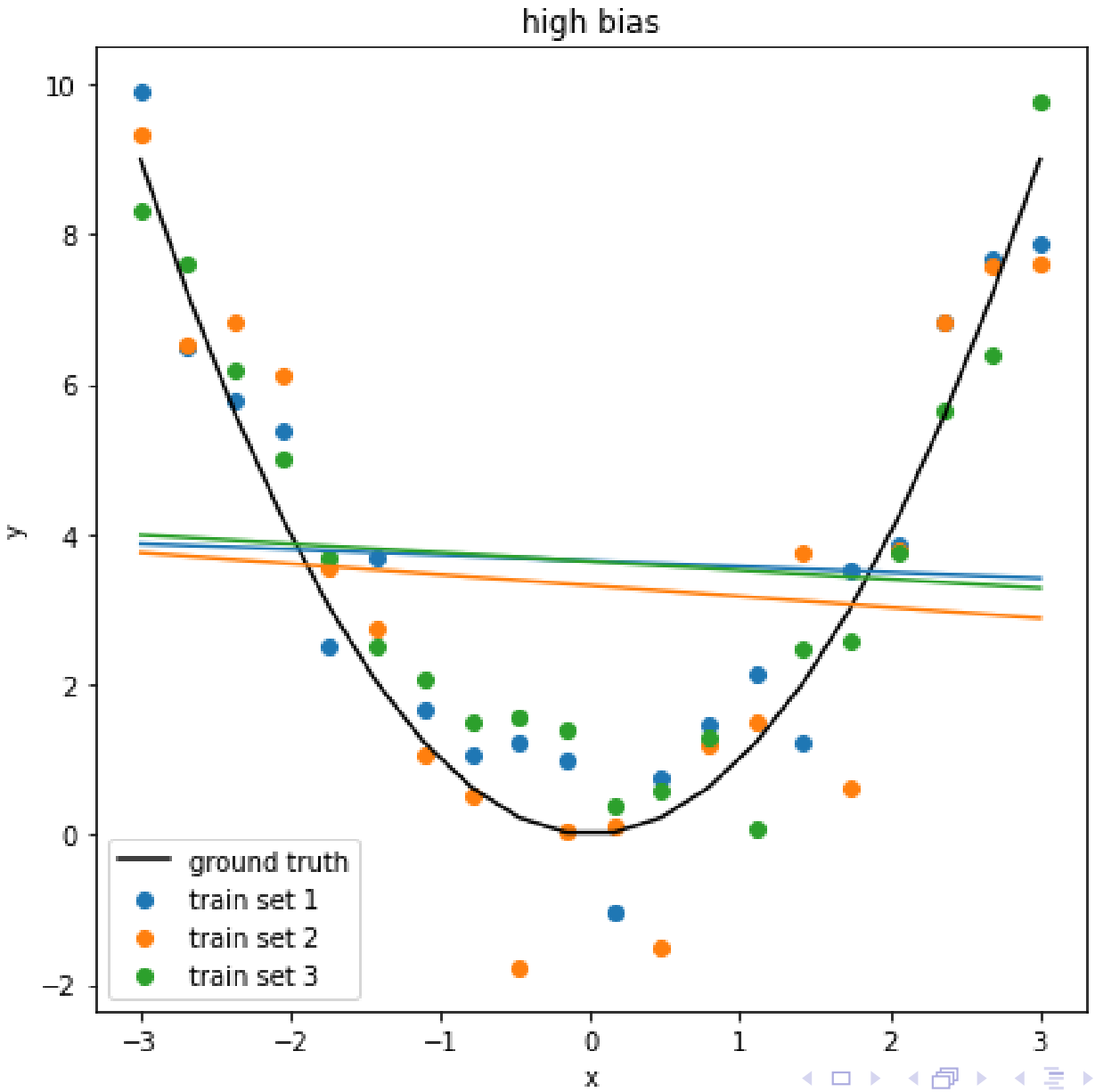
$$\mathbb{D}\left(\frac{X_1 + \ldots X_n}{n}\right) = \frac{\mathbb{D}X_1}{n}.$$

Усредняя, уменьшаем дисперсию!

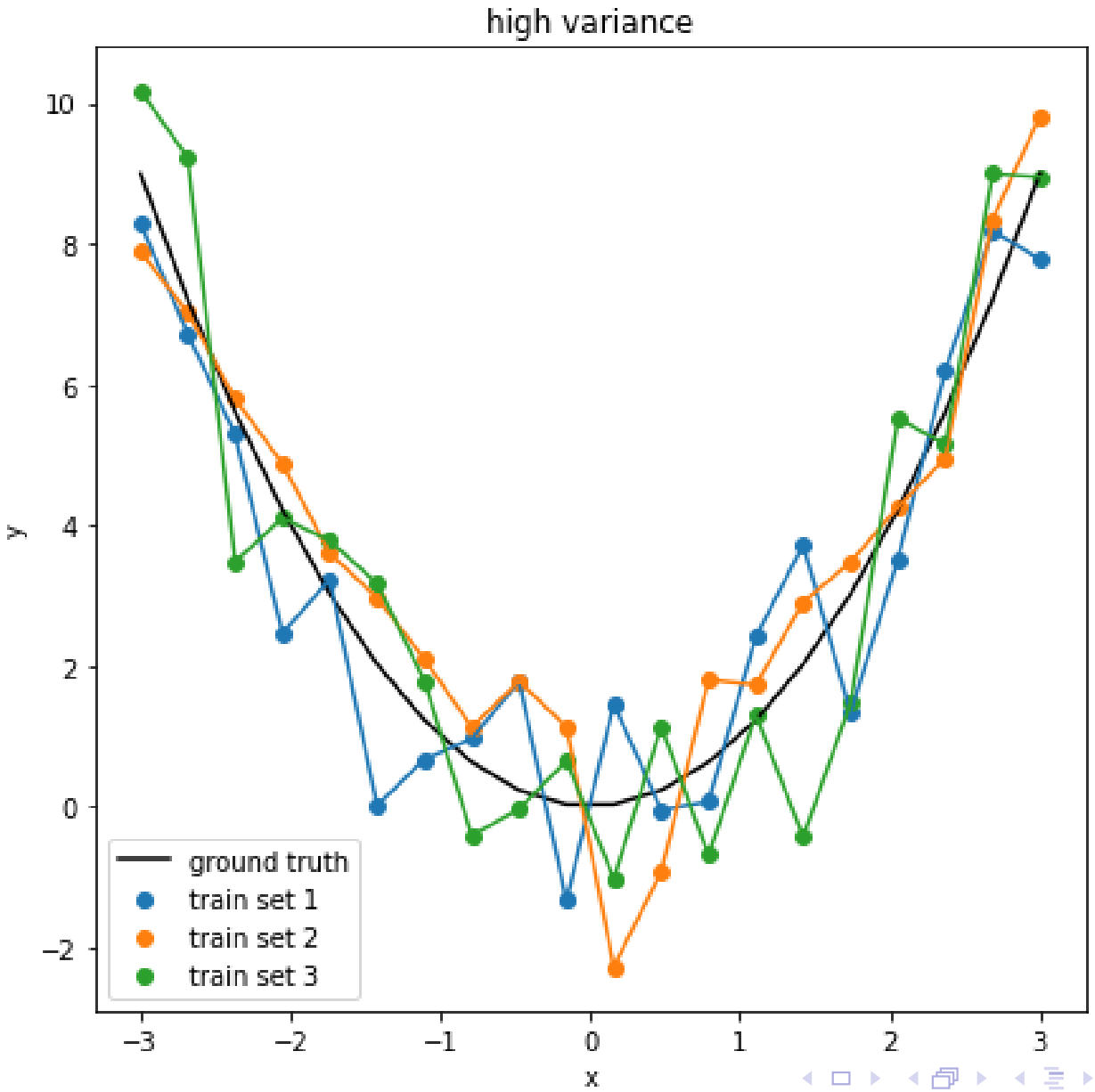
Bias vs Variance



Большое смещение



Большая дисперсия



Случайный лес (Random forest)

Random forest – ансамбль над деревьями, при построении которого:

- Каждое дерево строится на bootstrap-выборке (Bagging).
- При разбиении (split) очередной вершины дерева перебираем не все признаки, а случайные d штук (Feature sampling), и из них выбираем наилучший по критерию информативности.

Можно оценивать обобщающую способность через out-of-bag: для каждого наблюдения получаем прогноз, используя только те деревья, которые "не видели"это наблюдение.

Extremely randomized trees (Extra trees)

Extra trees – ансамбль над деревьями, при построении которого:

- Каждое дерево строится на всей выборке.
- При разбиении очередной вершины перебираем не все признаки, а а случайные d штук (Feature sampling). Для каждого рассматриваемого признака случайно выбираем порог разбиения. Выбираем наилучший из признаков по критерию информативности.

Бустинг (Boosting)

Идея:

- Будем использовать N «слабых» моделей $b_1(x), b_2(x), \ldots, b_n(x)$ и с помощью них строить одну «сильную» модель в виде:

$$b(x) = \alpha_1(x)b_1(x) + \alpha_2(x)b_2(x) + \ldots + \alpha_N(x)b_N(x).$$

- Процесс построения итеративный: каждая новая модель учится исправлять ошибки итогового алгоритма с прошлой итерации.
- Последовательно уменьшаем смещение (bias).

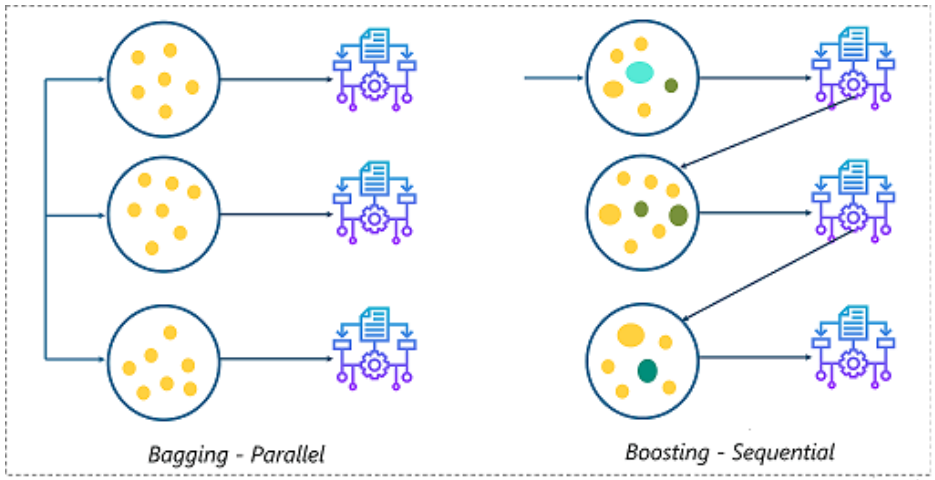


Рис.: Источник: edureka.co

Бустинг (Boosting)

Цель – построить алгоритм, минимизирующий эмпирический риск:

$$b(x) = \sum_{j=1}^{N} \alpha_j b_j(x),$$

$$\alpha_k, b_k = \underset{\alpha_k, b_k}{\operatorname{argmin}} = \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(y_i, \sum_{j=1}^{N} \alpha_j b_j(x)\right).$$

Итеративный процесс:

$$\alpha_m, b_m = \underset{\alpha_m, b_m}{\operatorname{argmin}} \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(y_i, \sum_{j=1}^{m-1} \alpha_j b_j(x) + \alpha_m b_m(x)\right).$$

Градиентный бустинг (Gradient boosting)

Итеративный процесс:

$$\alpha_m, b_m = \underset{\alpha_m, b_m}{\operatorname{argmin}} \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(y_i, \hat{b}(x) + \alpha_m b_m(x)\right).$$

Приблизим с помощью $b_m(x)$ антиградиент функции потерь

$$b_m(x_i) \approx -\frac{\partial L(y_i, z)}{\partial z}\Big|_{z=\hat{b}(x_i)},$$

после чего решим задачу оптимизации

$$\alpha_m = \underset{\alpha_m}{\operatorname{argmin}} \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(y_i, \hat{b}(x) + \alpha_m b_m(x_i)\right).$$

Таким образом, эмпирический риск будет убывать с каждой итерацией.

Обучение мета-алгоритмов

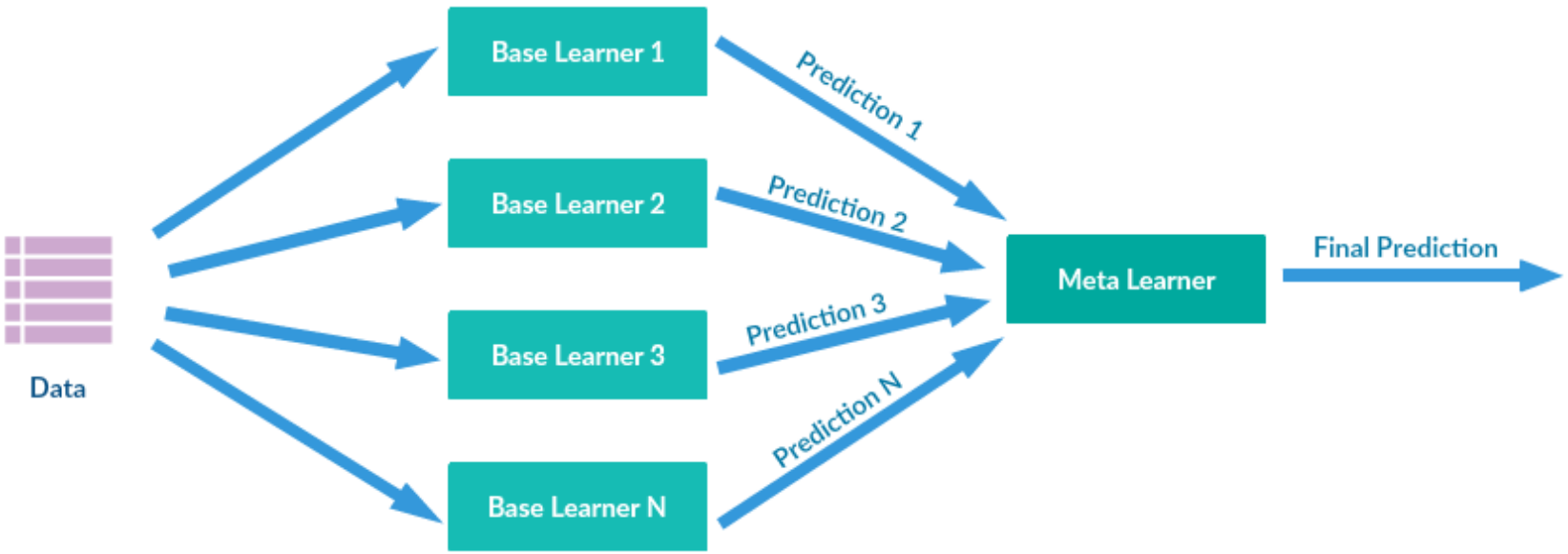


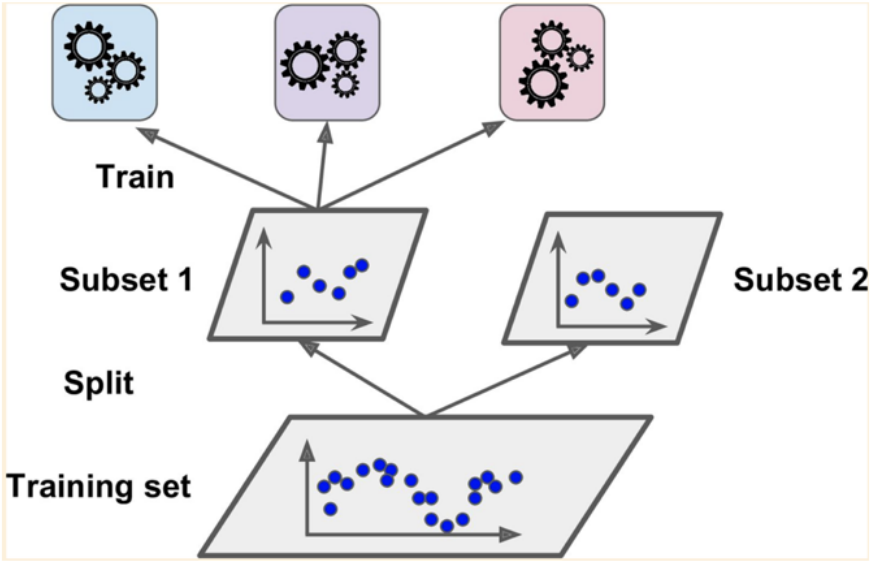
Рис.: Источник: towardsdatascience.com

Блендинг (Blending)

Пусть (X, y) – тренировочная выборка, $(X_{\text{test}}, y_{\text{test}})$ – тестовая.

- Разобьем тренировочную выборку на две части $(X_{\text{train}}, y_{\text{train}})$ и $(X_{\text{meta}}, y_{\text{meta}})$.
- Обозначим $(X_{\text{train}}, y_{\text{train}})$, $(X_{\text{meta}}, y_{\text{meta}})$ и $(X_{\text{test}}, y_{\text{test}})$ через A, B и C соответственно.
- Обучим модели $b_1(x), \ldots, b_N(x)$ на A.
- Обучим модель $a(x)$ на выборке B, используя в качестве признаков предсказания моделей $b_1(x), \ldots, b_n(x)$.
- Оцениваем итоговое качество модели $a(x)$ на выборке C.

Блендинг (Blending)



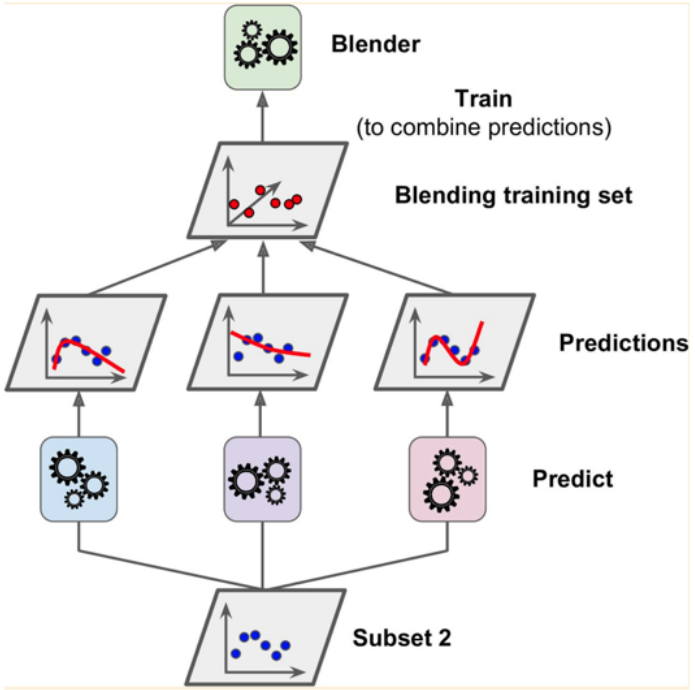


Рис.: Источник: stats.stackexchange.com

Стекинг (Stacking)

Пусть (X, y) – тренировочная выборка, $(X_{\text{test}}, y_{\text{test}})$ – тестовая.

- Разобьем тренировочную выборку на K фолдов $(X_{\text{train,i}}, y_{\text{train,i}})$.
- Обозначим $(X_{\text{train,i}}, y_{\text{train,i}})$ и $(X_{\text{test}}, y_{\text{test}})$ через A_i и C соответственно.
- Для каждого фолда A_i обучаем модели $b_1(x), \ldots, b_N(x)$ на остальных фолдах, получаем предсказание на данном фолде.
- Обучаем модель $a(x)$ на всей тренировочной выборке, используя в качестве признаков прогнозы, данные моделями $b_1(x), \ldots, b_N(x)$.
- Обучаем модели $b_1(x), \ldots, b_N(x)$ на всей тренировочной выборке.
- Тестируем итоговый алгоритм на выборке C.

Стекинг (Stacking)

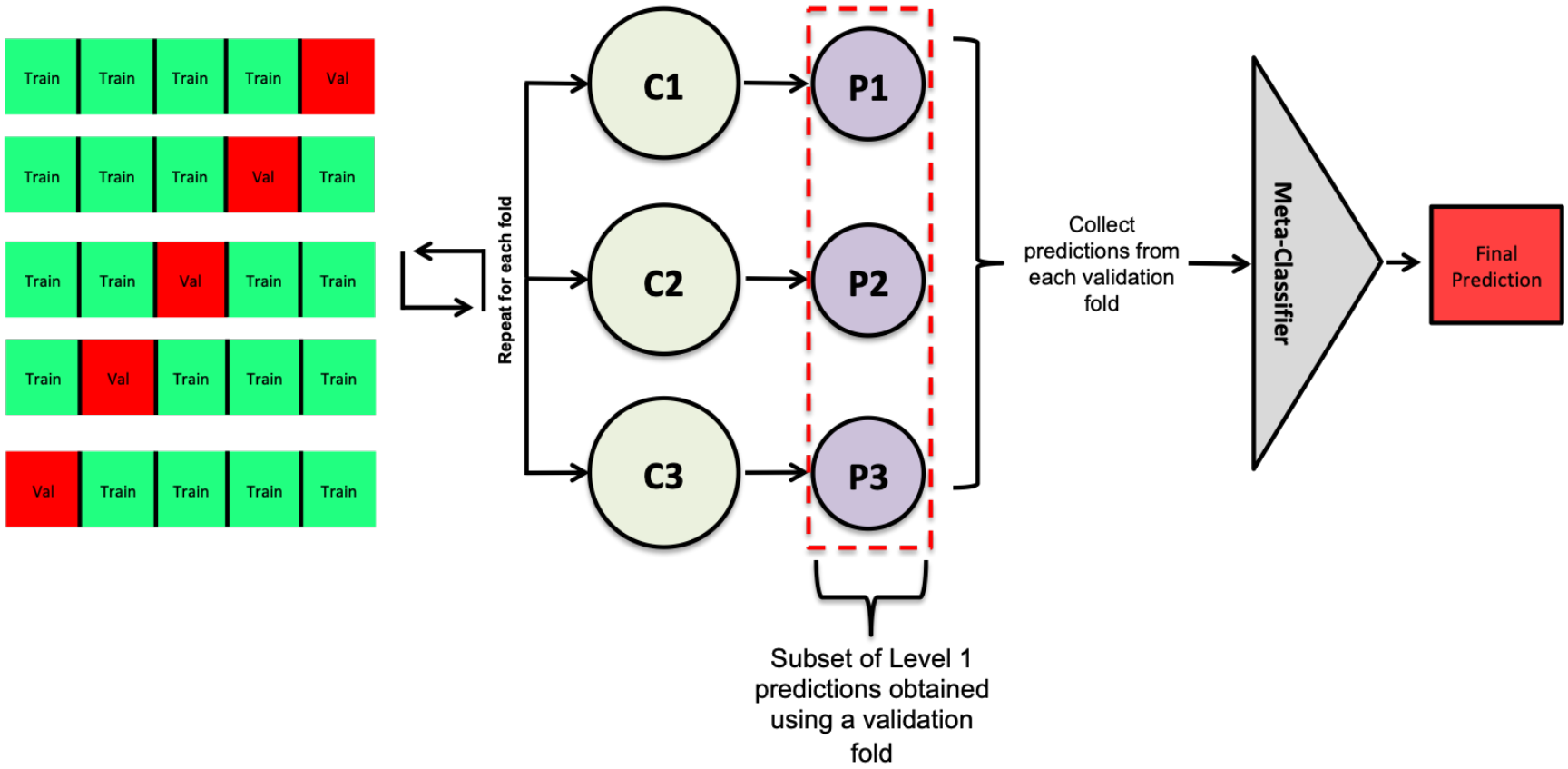


Рис.: Источник: towardsdatascience.com