Введение в АД

Лекция 6
Кросс-валидация, градиентный бустинг

ФЭФМ МФТИ

Весенний семестр 2023

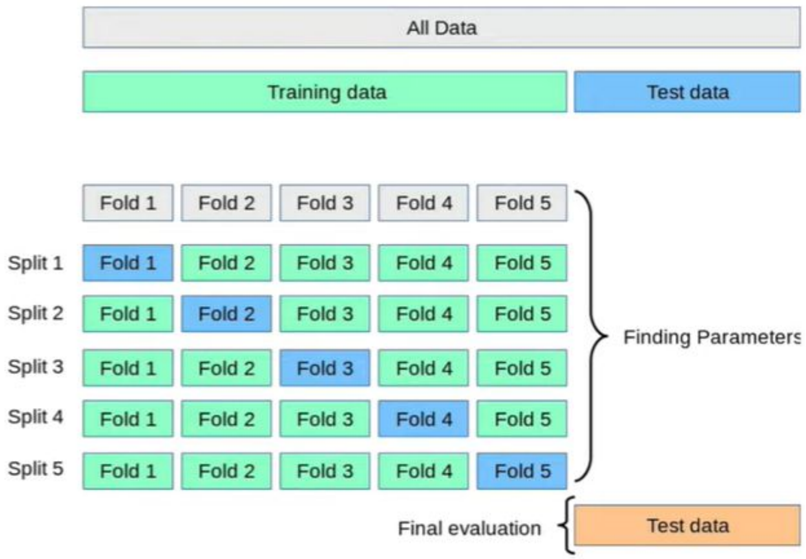
Cross-validation

По какому принципу нужно делить выборку на train/test?

Что делать, если данных мало, а хочется обучиться хорошо?

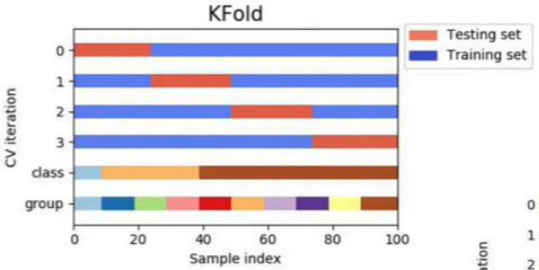
Стратегии валидации:

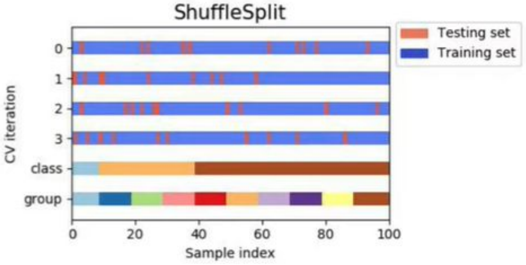
- hold-out
- k-fold
- leave-one-out (LOO)
- etc.

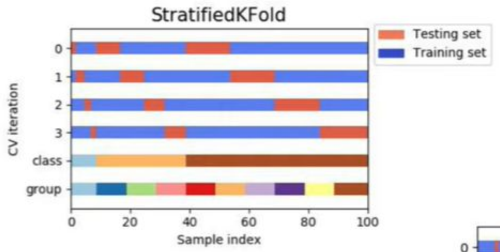


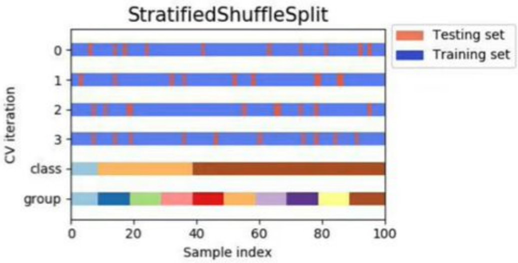
Cross-validation

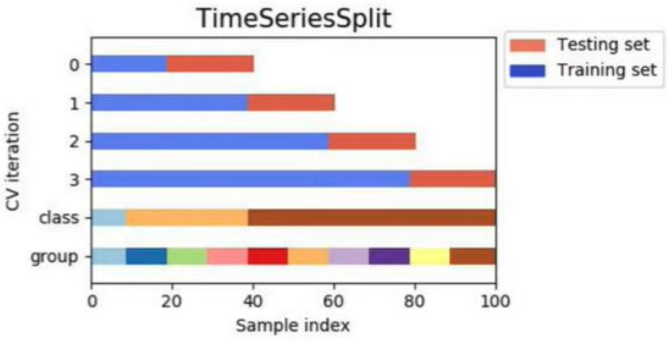
Важно: кросс-валидация не дает оценку качества обучения, она дает оценку качества подобранной **модели**







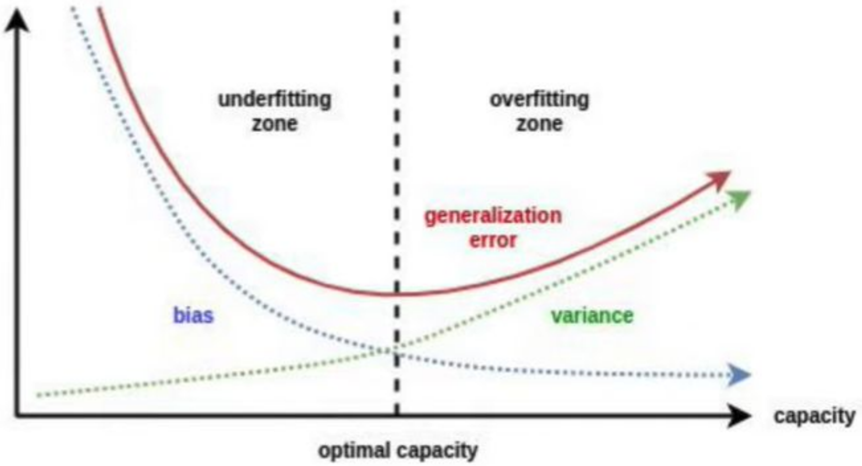




Bias-variance tradeoff

При обобщении результатов работы модели на генеральную совокупность, ошибку модели можно разложить на три части:

- **bias**, неправильные или недостаточные предположения модели о данных
- **variance**, излишняя чувствительность к флуктуациям в данных, “обучение на шум”
- **шум** в исходных данных, также называемый **irreducible error**



Bias-variance decomposition

The dataset $X = (x_i, y_i)_{i=1}^{\ell}$ with $y_i \in \mathbb{R}$ for regression problem.

Denote loss function $L(y,a) = \big(y - a(x)\big)^2$.

The corresponding risk estimation is

$$R(a) = \mathbb{E}_{x,y}\left[\big(y - a(x)\big)^2\right] = \int_{\mathbb{X}}\int_{\mathbb{Y}} p(x,y)\big(y-a(x)\big)^2 dxdy.$$

Let's show that $a_*(x) = \mathbb{E}[y \,|\, x] = \int_{\mathbb{Y}} yp(y \,|\, x)dy = \arg\min_a R(a).$

$$\begin{aligned} L(y,a(x)) &= (y-a(x))^2 = (y - \mathbb{E}(y \,|\, x) + \mathbb{E}(y \,|\, x) - a(x))^2 = \\ &= (y - \mathbb{E}(y \,|\, x))^2 + 2\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) + (\mathbb{E}(y \,|\, x) - a(x))^2. \end{aligned}$$

Let's return to the risk estimation:

$$\begin{aligned} R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\ &= \mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2 + \\ &+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big). \end{aligned}$$

22

$$\begin{aligned} &\mathbb{E}_x\mathbb{E}_y\Big[\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) \,|\, x\Big] = \\ &= \mathbb{E}_x\Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\mathbb{E}_y\Big[\big(y - \mathbb{E}(y \,|\, x)\big) \,|\, x\Big]\Big) = \\ &= \mathbb{E}_x\Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\big(\mathbb{E}(y \,|\, x) - \mathbb{E}(y \,|\, x)\big)\Big) = \\ &= 0 \end{aligned}$$

So the risk takes form:

Does not depend on a(x)

$$R(a) = \boxed{\mathbb{E}_{x,y}(y - \mathbb{E}(y\,|\,x))^2} + \mathbb{E}_{x,y}(\mathbb{E}(y\,|\,x) - a(x))^2.$$

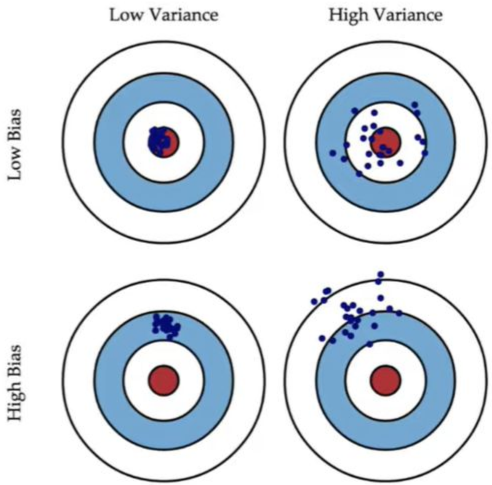
The minimum is reached when $a(x) = \mathbb{E}(y\,|\,x).$

So the optimal regression model with square loss is

$$a_*(x) = \mathbb{E}(y\,|\,x) = \int_{\mathbb{Y}} yp(y\,|\,x)dy.$$

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y\,|\,x]\right)^2\right]}_{\text{noise}} +$$

$$+ \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mathbb{E}[y\,|\,x]\right)^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X) - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right]}_{\text{variance}}.$$



Gradient boosting

Бустинг – подход, при которой для обучения следующих моделей мы используем данные о предыдущих. Реализован в CatBoost, XGBoost

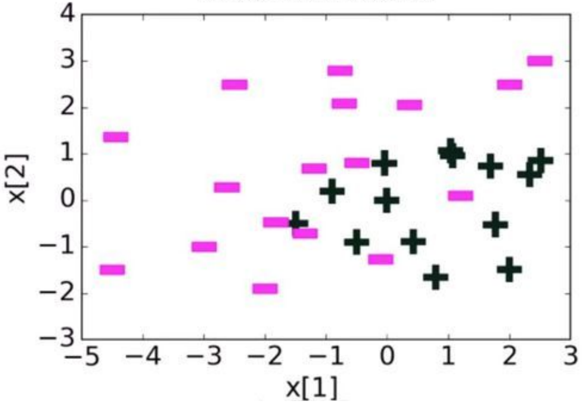
Плюсы:

- легкое построение модели
- быстрое обучение, если за базовую модель берем быстрые алгоритмы
- можно хорошо идентифицировать выбросы

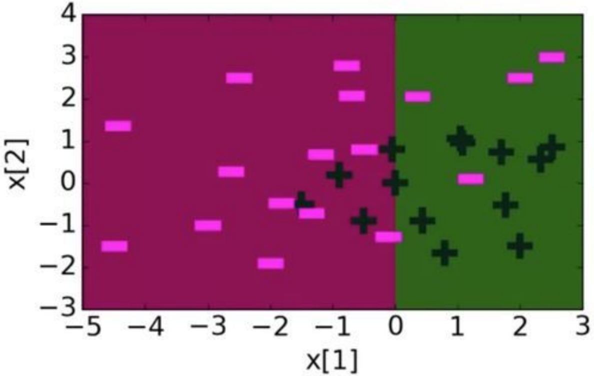
Минусы:

- теряется интерпретируемость
- слабая параллелизация

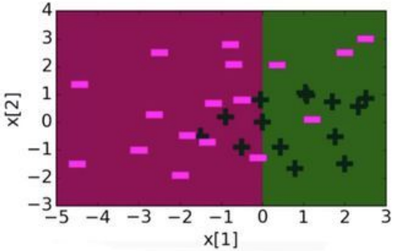
НАЧАЛЬНЫЕ ДАННЫЕ



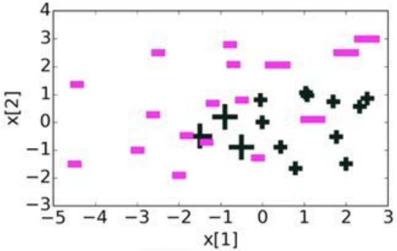
РЕШЕНИЕ F(X)



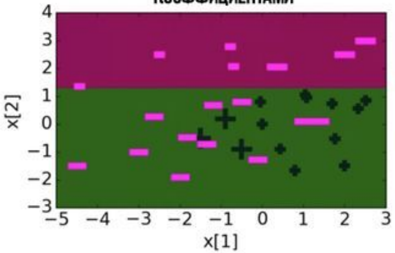
РЕШЕНИЕ F(X)

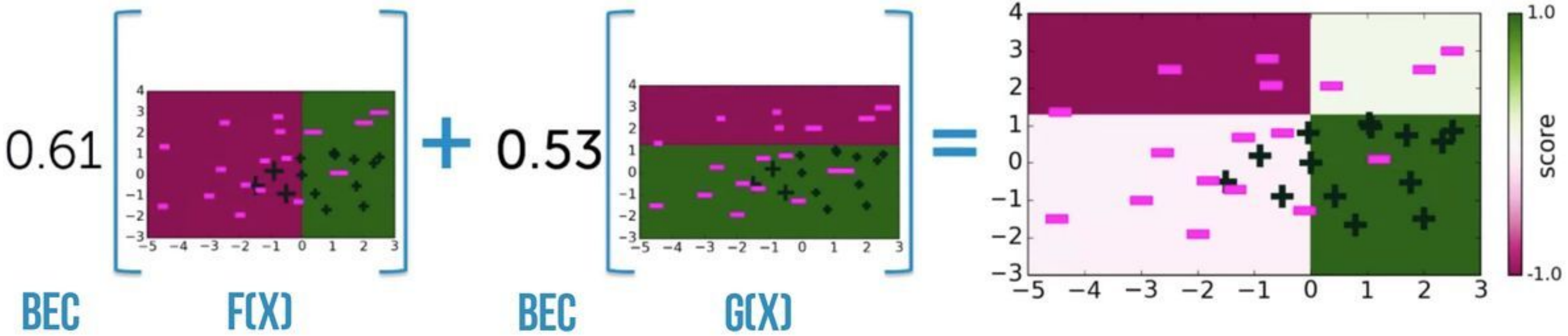


ПЕРЕСЧЕТ ВЕСОВЫХ КОЭФФИЦИЕНТОВ

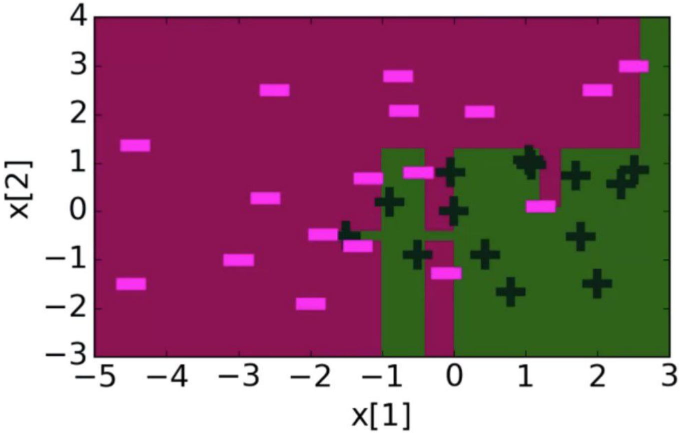


РЕШЕНИЕ G(X) С НОВЫМИ ВЕСОВЫМИ КОЭФФИЦИЕНТАМИ





после 30 итераций:



Откуда берутся коэффициенты?

Optimal model:

$$\hat{f}(x) = \underset{f(x)}{\arg\min}\ L(y, f(x)) = \underset{f(x)}{\arg\min}\ \mathbb{E}_{x,y}[L(y, f(x))]$$

Let it be from parametric family:

$$\hat{f}(x) = f(x, \hat{\theta}),$$

$$\hat{\theta} = \underset{\theta}{\arg\min}\ \mathbb{E}_{x,y}[L(y, f(x, \theta))]$$

$$\hat{f}(x) = \sum_{i=0}^{t-1} \hat{f}_i(x),$$

$$(\rho_t, \theta_t) = \underset{\rho,\theta}{\arg\min}\ \mathbb{E}_{x,y}[L(y, \hat{f}(x) + \rho \cdot h(x, \theta))],$$

$$\hat{f}_t(x) = \rho_t \cdot h(x, \theta_t)$$

Попробуем использовать градиентный спуск *в пространстве функций*

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)}, \quad \text{for } i = 1, \dots, n,$$

$$\theta_t = \arg\min_{\theta} \sum_{i=1}^{n} (r_{it} - h(x_i, \theta))^2, \qquad \rho_t = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i) + \rho \cdot h(x_i, \theta_t))$$

In linear regression case with MSE loss:

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)} = -2(\hat{y}_i - y_i) \propto \hat{y}_i - y_i$$