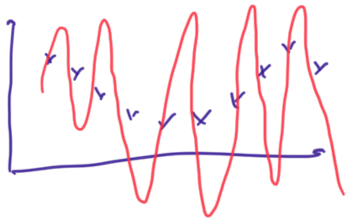
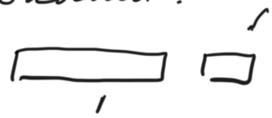
# Линейная регрессия



1

Оценивание обобщ. способности:

1) отлож. выборка



2) кросс-валидация

$K$ - число блоков (fold)

$X^k$ - $k$-й блок

$X^{\setminus k}$ - все блоки, кроме $k$-го

Leave-one-out
$K = l$

$$CV = \frac{1}{K} \sum_{k=1}^{K} Q(a(x; X^{\setminus k}), X^k)$$

обучена на $X^{\setminus k}$

Итоговая модель: 1) обучить на всех данных
2) усреднить $K$ моделей

$$c(x) = \frac{1}{K} \sum_{k=1}^{K} a(x, X^{\setminus k})$$

## Борьба с переобучением

В 1 $w_j$ очень большие
$10^6, 10^7, \dots$

Регуляризация - запрет на большие веса

$Q(w, X) \to \min$ - наш функционал

$$\boxed{Q(w,X) + \underbrace{\lambda}_{\text{коэф. регуляриз.}} \underbrace{\|w\|_2^2}_{\text{регуляризатор}} \to \min_w}$$

$$\left\{ \begin{array}{l} w = (X^T X + \underbrace{\lambda I}_{})^{-1} X^T y \\ \text{(для MSE)} \end{array} \right.$$

$\boxed{\lambda \geq 0}$

**Важно**: в $\|w\|_2^2$ не входит $w_0$ !!!

(иначе $a(x)$ не сможет быть одного порядка с $y$)

Почему большие веса — плохо?

$a(x) = 10^6 + \underbrace{3 \cdot 10^7 \cdot (\text{площадь})}_{} - 5 \cdot 10^6 \cdot (\text{расст. до метро})$

площадь $+ 0.001$

$a(x) + \underbrace{3 \cdot 10^4}_{30.000}$

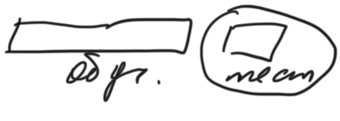
Как выбирать $\lambda$?

- по обуч. выборке: выбираем $\lambda$, при которой $Q(a, X)$ минимальна

плохо — оптимально $\lambda = 0$

Гиперпараметр — нельзя подбирать по обуч. выб.
вводятся для улучш. качества на новых данных

нужно подбирать по CV или отлож. выб.




Регуляриз. не обязательно через $L_2$-норму!

$$Q(w,X) + \lambda \|w\|_1 \to \min, \quad \|w\|_1 = \sum_{j=1}^{d} |w_j|$$

$\lambda$ $\pi$

$\|w\|_1$

Допустим, в $X$ есть лин. зав. признаки

[illegible]

$\exists v \ \forall x \in X \quad \langle v, x \rangle = 0$

$w_*$ - лучшие по MSE веса

$$\langle \underbrace{w_* + \alpha v}, x \rangle = \langle w_*, x \rangle + \alpha \underbrace{\langle v, x \rangle}_{=0} = \langle \underbrace{w_*}, x \rangle$$

т.е. решений много

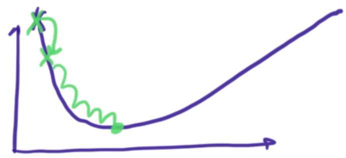
$w_* + \alpha v \xrightarrow[\alpha \to \infty]{}$

---

Обучение лин. регр.

для MSE: $w = \underbrace{(X^T X)^{-1}}_{O(d^3)} X^T y$

## Градиентное обучение моделей



1) $w^{(0)}$ - начальное приближение

$\nabla_w Q(w)$ - градиент $Q$ по $w$

$-\nabla Q(w)$ - в сторону наискорейш. убывания

2) $w^{(k)} = w^{(k-1)} - \eta_k \nabla Q(w^{(k-1)})$ - шаг град. спуска

$\eta_k$ - длина шага (learning rate)

3) когда останавливаться?

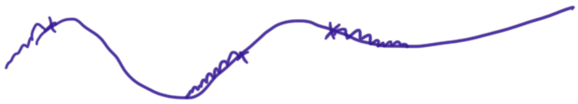
- когда ошибка на тесте перестает уменьшаться
- $\| w^{(k)} - w^{(k-1)} \| < \varepsilon$
- $| Q(w^{(k)}, X) - Q(w^{(k-1)}, X) | < \varepsilon$

— . . .

Сходимость: 1) к $\nabla Q(w) \approx 0$

для лин. моделей с матрицей полного ранга такая точка одна

2) если решений несколько, то можно делать мульти-старт



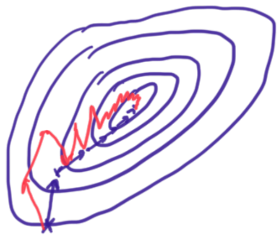
Улучшения GD

оценки градиента

градиентный шаг

## Оценивание градиента

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$



$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} q_i(w)$$

$$\nabla_w Q(w, X) = \underbrace{\frac{1}{\ell} \sum_{i=1}^{\ell} \nabla_w q_i(w)}$$

полный градиент

Стохастический GD (SGD): $\nabla Q(w) \approx \nabla q_i(w)$

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla q_{i_k}(w^{(k-1)})$$

$q_{i_k} = (\langle w, x_i \rangle - y_i$

$i_k$ - случ. номер объекта

→ $\sum_{k=1}^{\infty} \eta_k = \infty$

→ $\sum_{k=1}^{\infty} \eta_k^2 = const$

$\eta_k = \frac{1}{k}$

$\eta_k = \lambda \left( \frac{s_0}{s_0 + k} \right)^p$

$\lambda,\ s_0, p$ - параметры

Full GD:

$Q(w^{(k)}) - Q(w_*) =$

$= O\left(\frac{1}{k}\right)$

SGD: $= O\left(\frac{1}{\sqrt{k}}\right)$

---

Mini-batch GD:

$$\nabla Q(w) \approx \frac{1}{n} \sum_{j=1}^{n} \nabla q_{i_{k_j}}(w)$$

---

SAG: $z_i^{(0)} = \nabla q_i(w^{(0)})$

$k$-я итерация:

$$z_i^{(k)} = \begin{cases} \nabla q_i(w^{(k-1)}), & i = i_k \\ z_i^{(k-1)}, & \text{иначе} \end{cases}$$

$$\nabla Q(w^{(k-1)}) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} z_i^{(k)}$$

$O\left(\frac{1}{k}\right)$