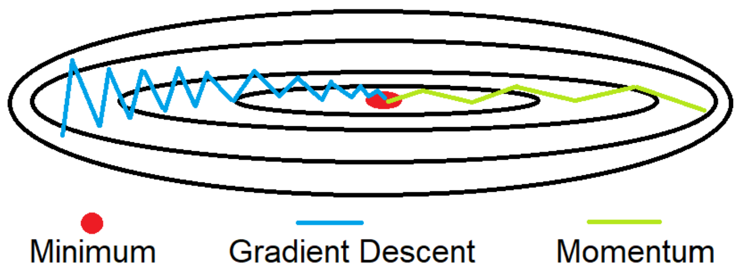
Лекция 4. Модификации градиентного спуска, регуляризация.

Модификации градиентного спуска.

Метод инерции (*momentum*).



$w^{(0)}$ — инициализация весов, h_0 — вектор инерции

Шаг: $\boxed{h_k = \alpha \cdot h_{k-1} + \underbrace{\eta_k \nabla_w Q\left(w^{(k-1)}\right)}_{\substack{\text{можно}\\ \text{заменять}\\ \text{на оценку град.}}}}$ $\Rightarrow$ $w^{(k)} = w^{k-1} - h_k$

Адаптивный шаг (*AdaGrad*). в основном для разреженных данных.

G_{0j}, j-номер признака.

$G_{kj} = G_{k-1,j} + \left(\nabla Q\left(w^{(k-1)}\right)\right)_j^2$ — насколько сильно уже обучена w_j

Шаг: $\boxed{w_j^{(k)} = w_j^{(k-1)} - \frac{\eta_k}{\sqrt{G_{kj} + \varepsilon}} \cdot \left(\nabla Q\left(w^{(k-1)}\right)\right)_j}$

Проблема: G_{kj} только растёт и может быстро остановить оптимизацию.

Модификация: *RMSProp* $G_{kj} = \alpha G_{k-1,j} + (1-\alpha) \cdot \left(\nabla Q\left(w^{(k-1)}\right)\right)_j^2$, $\alpha \in (0,1)$

Adam (*momentum* + *AdaGrad*) одновременно и инерция, и адаптивность.

Регуляризация.

Известный экспериментальный факт:

Линейная модель переобучена $\underbrace{\Leftrightarrow}_{\text{по слухам}}$ большие веса.

Почему?

Объяснение №1. Пусть есть линейно зависимые признаки $\exists v \in \mathbb{R}^d$: $\forall x \in \mathbb{X}$ $< v, x > = 0$

$w_\star$ — решение $\frac{1}{\ell}\sum_{i=1}^{\ell}(< w, x_i > - y_i)^2 \to \min\limits_w$

$\alpha > 0$: $< w_\star + \alpha v, x > = < w_\star, x > + \alpha \underbrace{< v, x >}_{0} = < w_\star, x >$, т. е. $w_\star + \alpha v$ — решение (еще одно)

Объяснение №2. $a(x) = 10^8 \cdot \text{площадь} - 10^9 \cdot \text{этаж} + 10^{11} \cdot \text{район}$. Добавим совсем чуть чуть к признаку

$10^8(\text{площадь} + 0{,}001) = 10^8 \cdot \text{площадь} + \underbrace{10^8 \cdot 0{,}001}_{10^5}$

Гиперчувствительность к изменениям в признаках — не соответствует тому, как работает мир.

❗ Запретить большие веса ❗

$Q(w) + \underbrace{\alpha}_{\substack{\text{коэф.}\\ \text{регуляр.}}} \cdot \underbrace{R(w)}_{\text{регуляризатор}} \to \min$

$$R(w) = \left[\begin{array}{l} \|w\|_1 = \sum_{j=1}^{d}|w_j| \text{ — это } L_1 \text{ регуляризатор.} \\ \|w\|_2^2 = \sum_{j=1}^{d} w_j^2 \text{ — это } L_2 \text{ регуляризатор.} \end{array}\right. , \quad \left[\begin{array}{lll} \alpha \gg 0 & \Rightarrow & w_\star = 0 \\ \alpha = 0 & \Rightarrow & \text{нет рег.} \end{array}\right.$$

α нельзя подбирать по обучающей выборке — гиперпараметр. Подбираем по новым данным (отложенной выборке, …)

Стратегии подбора α:

- *Grid Search*
- *Random Search*
- *AutoML*

Ridge-регрессия $\frac{1}{\ell}\sum_{i=1}^{\ell}(< w, x_i > - y_i)^2 + \alpha\|w\|_2^2 \to \min\limits_w$, невырожденная матрица

LASSO $\frac{1}{\ell}\sum(< w, x_i > - y_i)^2 + \alpha\|w\|_1 \to \min$, зануляет часть весов