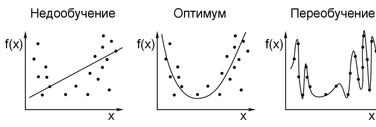
# Лекция 3. Оценка обобщающей способности, градиентные методы обучения.

## Переобучение.




Переобучение (*overfitting*). Ошибка на новых данных > ошибка на обучающей выборке.

Обобщающая способность (*generalization*). Качество модели на новых данных не сильно хуже, чем на обучении.

## Оценка обобщающей способности.

1. Отложенная выборка (*hold-out set*).

| Обучающая выборка | Валид. выборка | $\xrightarrow{\text{Лучшая модель}}$ | Тестовая выборка |
|---|---|---|---|

2. Кросс-валидация ($CV$).

$K$-число блоков ($folds$).

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|

Обучение на $X_1 \cup X_2$, тест на $X_3$, получаем $Q_1$.
↓
Обучение на $X_1 \cup X_3$, тест на $X_2$, получаем $Q_2$. $\qquad Q = \frac{1}{3}(Q_1 + Q_2 + Q_3)$
↓
Обучение на $X_2 \cup X_3$, тест на $X_1$, получаем $Q_3$.

При $k = \ell \quad \Rightarrow \quad LOO(leave\text{-}one\text{-}out)$

Что делать дальше?

⋆ Обучить модель на всей выборке.
⋆ $x \to \frac{1}{3}\big(a_1(x) + a_2(x) + a_3(x)\big)$

Замечания:

⋆ Если $\ell \gg 0$, то $CV$ вряд ли оправдано.
⋆ $CV$ требует обучения $k$ моделей.

## Обучение линейных моделей.

$\frac{1}{\ell}\sum_{i=1}^{\ell}(< w, x_i > - y_i)^2 \;\to\; \min_{w \in \mathbb{R}^d} \square, \qquad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ x_{21} & \dots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{\ell 1} & \dots & x_{\ell d} \end{pmatrix}, \qquad w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}, \qquad X \cdot w = \begin{pmatrix} < w, x_1 > \\ \vdots \\ < w, x_\ell > \end{pmatrix}$

Тогда векторный вид: $\quad Q(w) = \frac{1}{\ell}\|Xw - y\|_2^2 \to \min_w \square$

$\nabla Q(w) = 0 \quad \Rightarrow \quad \boxed{w_* = (X^T X)^{-1} X^T y}$

Проблемы:

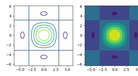
⋆ Матрица может быть вырожденной.
⋆ Обращение матрицы за $O(d^3)$
⋆ Если функция потерь $L(y, z)$ более хитрая, то решить $\nabla Q(w) = 0$ не выйдет.

## Градиентные методы обучения.

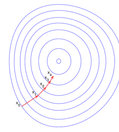
$Q(w_1, \dots, w_d) \to \min_w \square, \qquad Q$ – дифференцируемый.

Важные свойства градиента:

⋆ $\nabla Q(w)$ показывает направление наискорейшего роста в этой точке $w$, $\qquad (-\nabla Q(w)$ наискорейшее убывание)
⋆ $\nabla Q(w)$ ортогонален линии уровня



Градиентный спуск.



$w^{(0)}$ – инициализация (в линейной моделе можно брать случайно)

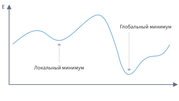
Шаг спуска: $\boxed{w^{(k)} = w^{(k-1)} - \eta \nabla Q\big(w^{(k-1)}\big)}$, $\qquad \eta$ – длина шага ($learning\ rate$)

Останавливаемся, если:

⋆ $\big\|w^{(k)} - w^{(k-1)}\big\| < \varepsilon$
⋆ $\big|Q\big(w^{(k)}\big) - Q\big(w^{(k-1)}\big)\big| < \varepsilon$
⋆ $\big\|\nabla Q\big(w^{(k)}\big)\big\| < \varepsilon$
⋆ $k > N$
⋆ ошибка на отложенной выборке перестала убывать.

## Некоторые наблюдения:

⋆ Локальные минимумы:




Решение: мультистарт (сомнительно для большого количества весов)

⋆ Много условий сходимости:

$$\begin{cases} Q(w) - \text{выпуклая} + \text{дифференцируемая} \\ \nabla Q(w) - \text{липшицева}\big(\|\nabla Q(w_1) - \nabla Q(w_2)\| \le L \cdot \|w_1 - w_2\|\big) \\ \eta \text{ не очень большая } \left(\eta \le \frac{1}{L}\right) \end{cases} \Rightarrow \quad \text{градиентный спуск сходится к минимуму.}$$

⋆ С линейными моделями и адекватными функциями потерь $Q(w)$ всегда (почти всегда)выпуклые.

⋆ $Q\big(w^{(k)}\big) - Q(w_*) = \underline{O}\left(\frac{1}{k}\right)$

## Оценивание градиента.

$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a(x_i, w)\big), \qquad Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} q_i(w), \qquad \nabla Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} \nabla q_i(w), \qquad \ell = 10^6 \Rightarrow 10^6$ градиентов

Стохастический градиентный спуск ($SGD$).

$\nabla Q \approx \nabla q_i(w) \quad \Rightarrow \quad$ шаг: $i_k$ – индекс случайного объекта.

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

Проблема: если $\big\|w^{(k)} - w_*\big\| \gg 0, \quad$ то $\nabla q_{i_k}(w) \approx \nabla Q(w)$
если $\big\|w^{(k)} - w_*\big\| \approx 0, \quad$ то $\nabla q_{i_k}(w) \neq \nabla Q(w)$

Идея: $w^{(k)} = w^{(k-1)} - \eta \nabla q_{i_k}\big(w^{(k-1)}\big)$, где $\underbrace{\sum_{k=1}^{\infty} \eta_k = \infty \quad \text{и} \quad \sum_{k=1}^{\infty} \eta_k^2 < \infty}_{\text{условия Роббинса-Монро}} \quad \Rightarrow \quad SGD$ сойдется к минимуму (если угадаем $w^{(0)}$)

$\eta_k = \lambda\left(\frac{s_0}{s_0 + k}\right)^p, \qquad \lambda, s_0, p$ – надо подбирать.

Скорость сходимости: $\mathbb{E}\big[Q\big(w^{(k)}\big) - Q(w_*)\big] = \underline{O}\left(\frac{1}{\sqrt{k}}\right)$

Замечания:

⋆ $mini\text{-}batch\ GD$: $\quad \nabla Q(w) \approx \frac{1}{t}\sum_{j=1}^{t} \nabla q_{i_j}(w)$

⋆ $SGD$ хорош для онлайн-обучения (можно обучаться на огромных выборках).

## *Stochastic average gradient (SAG)*

$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} q_i(w), \qquad z_i^{(0)} = \nabla q_i\big(w^{(0)}\big)$

Итерация $SAG$: $\quad i_k \sim \{1, \dots, \ell\}, \qquad z_i^{(k)} = \begin{cases} \nabla q_i\big(w^{(k-1)}\big), & i = i_k \\ z_i^{(k-1)}, & i \neq i_k \end{cases}$

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

$\mathbb{E}\big[Q\big(w^{(k)}\big) - Q(w_*)\big] = \underline{O}\left(\frac{1}{k}\right)$