# Машинное обучение
# Семинар 3
# Матрично-векторное дифференцирование

Как правило, дифференцируемые модели обучаются с помощью градиентного спуска, а для него важно уметь считать градиент функционала ошибки по параметрам модели. Можно считать градиент покоординатно, а потом пристально смотреть на формулы и пытаться понять, как это может выглядеть в векторной форме. Гораздо проще считать градиент напрямую — а для этого поможет знание градиентов для основных функций и основных правил матрично-векторого дифференцирования.

## 1 Вывод основных формул

Введём следующие определения:

- При отображении вектора в число $f(x) : \mathbb{R}^n \to \mathbb{R}$

$$\nabla_x f(x) = \left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right]^T$$

- При отображении матрицы в число $f(A) : \mathbb{R}^{n \times m} \to \mathbb{R}$

$$\nabla_A f(A) = \left(\frac{\partial f}{\partial A_{ij}}\right)_{i,j=1}^{n,m}$$

Мы хотим оценить, как функция изменяется по каждому из аргументов по отдельности. Поэтому производной функции по вектору будет вектор, по матрице — матрица. Теперь поупражняемся в дифференцировании:

**Задача 1.1.** *Пусть $a \in \mathbb{R}^n$ — вектор параметров, а $x \in \mathbb{R}^n$ — вектор переменных. Необходимо найти производную их скалярного произведения по вектору переменных $\nabla_x a^T x$.*

**Решение.**

$$\frac{\partial}{\partial x_i} a^T x = \frac{\partial}{\partial x_i} \sum_j a_j x_j = a_i,$$

поэтому $\nabla_x a^T x = a$.

Заметим, что $a^T x$ — это число, поэтому $a^T x = x^T a$, следовательно,

$\nabla_x x^T a = a$.

■

**Задача 1.2.** *Пусть теперь $A \in \mathbb{R}^{n\times n}$. Необходимо найти $\nabla_x x^T Ax$.*

**Решение.**

$$\frac{\partial}{\partial x_i} x^T Ax = \frac{\partial}{\partial x_i} \sum_j x_j (Ax)_j = \frac{\partial}{\partial x_i} \sum_j x_j \left( \sum_k a_{jk} x_k \right) = \frac{\partial}{\partial x_i} \sum_{j,k} a_{jk} x_j x_k =$$
$$= \sum_{j\neq i} a_{ji} x_j + \sum_{k \neq i} a_{ik} x_k + 2a_{ii} x_i = \sum_j a_{ji} x_j + \sum_k a_{ik} x_k = \sum_j (a_{ji} + a_{ij}) x_j.$$

Поэтому $\nabla_x x^T Ax = (A + A^T)x$.

■

**Задача 1.3.** *Пусть $A \in \mathbb{R}^{n\times n}$. Необходимо найти $\nabla_A \det A$.*

**Решение.** Воспользуемся теоремой Лапласа о разложении определителя по строке:

$$\frac{\partial}{\partial A_{ij}} \det A = \frac{\partial}{\partial A_{ij}} \left[ \sum_k (-1)^{i+k} A_{ik} M_{ik} \right] = (-1)^{i+j} M_{ij},$$

где $M_{ik}$ — дополнительный минор матрицы $A$. Также вспомним формулу для элементов обратной матрицы

$$(A^{-1})_{ij} = \frac{1}{\det A} (-1)^{i+j} M_{ji}.$$

Подставляя выражение для дополнительного минора, получаем ответ $\nabla_A \det A = (\det A) A^{-T}$.

■

**Задача 1.4.** *Пусть* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$. *Необходимо найти* $\nabla_A tr(AB)$.

**Решение.**

$$\frac{\partial}{\partial A_{ij}}\mathrm{tr}(AB) = \frac{\partial}{\partial A_{ij}}\sum_k (AB)_{kk} = \frac{\partial}{\partial A_{ij}}\sum_{k,l} A_{kl}B_{lk} = B_{ji}.$$

То есть, $\nabla_A \mathrm{tr}(AB) = B^T$.

■

**Задача 1.5.** *Пусть* $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, $y \in \mathbb{R}^m$. *Необходимо найти* $\nabla_A x^T Ay$.

**Решение.** Воспользовавшись циклическим свойством следа матрицы (для матриц подходящего размера):

$$\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$$

и результатом предыдущей задачи, получаем

$$\nabla_A x^T Ay = \nabla_A \mathrm{tr}(x^T Ay) = \nabla_A \mathrm{tr}(Ayx^T) = xy^T.$$

■

Наконец, научимся считать градиенты для сложных функций. Допустим, даны функции $f : \mathbb{R}^n \to \mathbb{R}^m$ и $g : \mathbb{R}^m \to \mathbb{R}$. Тогда градиент их композиции можно вычислить как

$$\nabla_x g\left(f(x)\right) = J_f^T(x) \nabla_z \, g(z)\big|_{z=f(x)},$$

где $J_f(x) = \left(\frac{\partial f_i(x)}{\partial x_j}\right)_{i,j=1}^{m,n}$ — матрица Якоби для функции $f$. Если $m = 1$ и функция $g(z)$ имеет всего один аргумент, то формула упрощается:

$$\nabla_x g\left(f(x)\right) = g'(f(x))\nabla_x f(x).$$

**Задача 1.6.** *Вычислите градиент логистической функции потерь для линейной модели по параметрам этой модели:*

$$\nabla_w \log\left(1 + \exp(-y\langle w, x\rangle)\right).$$

**Решение.**

$$\begin{aligned}
&\nabla_w \log\left(1 + \exp(-y\langle w, x\rangle)\right) = \\
&\qquad = \frac{1}{1+\exp(-y\langle w, x\rangle)}\nabla_w\left(1 + \exp(-y\langle w, x\rangle)\right) = \\
&\qquad = \frac{1}{1+\exp(-y\langle w, x\rangle)}\exp(-y\langle w, x\rangle)\nabla_w\left(-y\langle w, x\rangle\right) = \\
&\qquad = -\frac{1}{1+\exp(-y\langle w, x\rangle)}\exp(-y\langle w, x\rangle)yx = \\
&\qquad = \left\{\sigma(z) = \frac{1}{1+\exp(-z)}\right\} = \\
&\qquad = -\sigma(-y\langle w, x\rangle)yx
\end{aligned}$$

■

## §1.1 Решение задачи регрессии для многомерного случая

Вспомним, зачем мы хотели научиться дифференцировать. В общем случае мы имеем выборку $\{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ $i = \overline{1, \ell}$, и хотим найти наилучшие параметры модели $a(x) = \langle w, x \rangle$ с точки зрения минимизации функции ошибки

$$Q(w) = (y - Xw)^T (y - Xw).$$

Здесь $X \in \mathbb{R}^{\ell \times d}$ — матрица «объекты-признаки» для обучающей выборки, $y \in \mathbb{R}^{\ell}$ — вектор значений целевой переменной на обучающей выборке, $w \in \mathbb{R}^d$ — вектор параметров. Выпишем градиент функции ошибки по $w$:

$$\begin{aligned}\nabla_w Q(w) &= \nabla_w [y^T y - y^T X w - w^T X^T y + w^T X^T X w] = \\ &= 0 - X^T y - X^T y + (X^T X + X^T X) w = 0.\end{aligned}$$

Таким образом, искомый вектор параметров выражается как

$$w = (X^T X)^{-1} X^T y.$$

Заметим, что это общая формула, и нет необходимости выводить формулу для регрессии вида $a(x) = Xw + w_0$, т.к. мы всегда можем добавить признак (столбец матрицы $X$), который всегда будет равен 1, и по уже выведенной формуле найдём параметр $w_0$.

Покажем, почему найденная точка — точка минимума, если матрица $X^T X$ обратима. Из курса математического анализа мы знаем, что если матрица Гессе функции положительно определёна в точке, градиент которой равен нулю, то эта точка является локальным минимумом.

$$\nabla^2 Q(w) = 2 X^T X.$$

Необходимо понять, является ли матрица $X^T X$ положительно определённой. Запишем определение положительной определённости матрицы $X^T X$:

$$z^T X^T X z > 0, \ \forall z \in \mathbb{R}^d, z \neq 0.$$

Видим, что тут записан квадрат нормы вектора $Xz$, то есть это выражение будет не меньше нуля. В случае, если матрица $X$ имеет «книжную» ориентацию (строк не меньше, чем столбцов) и имеет полный ранг (нет линейно зависимых столбцов), то вектор $Xz$ не может быть нулевым, а значит выполняется

$$z^T X^T X z = ||Xz||^2 > 0, \ \forall z \in \mathbb{R}^d, z \neq 0.$$

То есть $X^T X$ является положительно определённой матрицей. Также, по критерию Сильвестра, все главные миноры (в том числе и определитель) положительно определённой матрицы положительны, а, следовательно, матрица $X^T X$ обратима, и решение существует. Если же строк оказывается меньше, чем столбцов, или $X$ не является полноранговой, то $X^T X$ необратима и решение $w$ определено неоднозначно.