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

Когда мы работаем с одномерными функциями, для поиска любых производных нам хватает небольшой таблицы со стандартными случаями и пары правил. Для случая матриц все эти правила можно обобщить, а таблицы дополнить специфическими функциями вроде определителя. Удобнее всего оказывается работать в терминах «дифференциала» — с ним можно не задумываться о промежуточных размерностях, а просто применять стандартные правила.

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 $f(A) : \mathbb{R}^{n \times m} \to \mathbb{R}$

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

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

$$\mathfrak{J}_x = \left(\frac{\partial f_j}{\partial x_i}\right)_{i,j=1}^{n,m}$$

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

Нарисуем таблицу с тем, как выглядят дифференциалы для разных случаев. По строчкам будем откладывать то, откуда бьёт функция, то есть входы. По столбцам будем откладывать то, куда бьёт функция, то есть выходы. На пересечении будут расположены дифференциалы. Для ситуаций обозначенных прочерками обобщения получить не выйдет.

	скаляр	вектор	матрица
скаляр	$f'(x)\,\mathrm{d}x$	$\mathfrak{J} * \mathrm{d}x$	—
вектор	$\nabla_x f(x)^T\,\mathrm{d}x$	$\mathfrak{J}\,\mathrm{d}x$	—
матрица	$\mathrm{tr}(\nabla_A f(A)\,\mathrm{d}A)$	—	—

Всегда, когда мы будем сталкиваться с дифференцированием на практике, мы будем выяснять, к какой из ситуации относится задача, а дальше сводит дифференциал к виду из таблицы выше и вытаскивать из него производную.

Дифференциал — это линейная часть приращения функции. Если мы находимся в какой-то точке x_0 и делаем из неё небольшое приращение $\mathrm{d}x$, то наша функция изменится примерно на $\mathrm{d}f(x)$.

1. Для функции, которая бьёт **из скаляров в скаляры** $f(x) : \mathbb{R} \to \mathbb{R}$ дифференциал выглядит как $\mathrm{d}f(x) = f'(x)\,\mathrm{d}x$.

2. Когда функция бьёт **из векторов в скаляры** $f(x) : \mathbb{R}^n \to \mathbb{R}$., мы имеем дело с функцией нескольких переменных. Нам нужно взять производную по каждой из них и получить вектор производных, градиент

$$\nabla_x f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

   Если умножить транспонированный градиент на вектор приращений, у нас получится дифференциал

$$\mathrm{d}f(x) = \nabla_x f^T\,\mathrm{d}x = \begin{pmatrix} \frac{\partial f}{\partial x_1}, & \dots, & \frac{\partial f}{\partial x_n} \end{pmatrix} \begin{pmatrix} \mathrm{d}x_1 \\ \dots \\ \mathrm{d}x_n \end{pmatrix} = \frac{\partial f}{\partial x_1} \cdot \mathrm{d}x_1 + \ldots + \frac{\partial f}{\partial x_n} \cdot \mathrm{d}x_n.$$

   При изменении x_i на $\mathrm{d}x_i$ функция будет при прочих равных меняться пропорционально соответствующей частной производной.

3. Когда функция бьёт **из векторов в векторы** $f(x) : \mathbb{R}^n \to \mathbb{R}^m$, мы взаимодействуем с семейством функций. Например, если $n = 1$ то у нас есть m функций, каждая из которых применяется к x. На выходе получается вектор

$$\begin{pmatrix} f_1(x) \\ f_2(x) \\ \dots \\ f_m(x). \end{pmatrix}$$

Если мы хотим найти производную, нужно взять частную производную каждой функции по x и записать в виде вектора. Дифференциал также будет представлять из себя вектор, так как при приращении аргумента на какую-то величину изменяется каждая из функций

$$\mathrm{d}f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \dots \\ \frac{\partial f_m}{\partial x} \end{pmatrix} * \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}x \\ \dots \\ \mathrm{d}x \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x}\,\mathrm{d}x \\ \frac{\partial f_2}{\partial x}\,\mathrm{d}x \\ \dots \\ \frac{\partial f_m}{\partial x}\,\mathrm{d}x \end{pmatrix}.$$

Под символом $*$ имеется в виду поэлементное умножение. Если $n > 1$, то аргументов на вход в такой вектор из функций идёт несколько, на выходе получается матрица

$$\begin{pmatrix} f_1(x_1) & f_1(x_2) & \dots & f_1(x_n) \\ f_2(x_1) & f_2(x_2) & \dots & f_2(x_n) \\ \dots & \dots & \ddots & \dots \\ f_m(x_1) & f_m(x_2) & \dots & f_m(x_n) \end{pmatrix}$$

Производной такой многомерной функции будет матрица из частных производных каждой функции по каждому аргументу

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \dots & \dots & \ddots & \dots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Дифференциал снова будет представлять из себя вектор

$$\mathrm{d}f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \dots & \dots & \ddots & \dots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \\ \dots \\ \mathrm{d}x_n \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}\,\mathrm{d}x_1 + \frac{\partial f_1}{\partial x_2}\,\mathrm{d}x_2 + \ldots + \frac{\partial f_1}{\partial x_n}\,\mathrm{d}x_n \\ \frac{\partial f_2}{\partial x_1}\,\mathrm{d}x_1 + \frac{\partial f_2}{\partial x_2}\,\mathrm{d}x_2 + \ldots + \frac{\partial f_2}{\partial x_n}\,\mathrm{d}x_n \\ \dots \\ \frac{\partial f_m}{\partial x_1}\,\mathrm{d}x_1 + \frac{\partial f_m}{\partial x_2}\,\mathrm{d}x_2 + \ldots + \frac{\partial f_m}{\partial x_n}\,\mathrm{d}x_n \end{pmatrix}.$$

4. Функция бьёт **из матриц в скаляры** $f(A) : \mathbb{R}^{n\times k} \to \mathbb{R}$.В таком случае нам надо найти производную функции по каждому элементу матрицы, то есть дифференциал будет выглядеть как

$$\mathrm{d}f(A) = \frac{\partial f}{\partial a_{11}}\,\mathrm{d}a_{11} + \ldots + \frac{\partial f}{\partial a_{nk}}\,\mathrm{d}a_{nk}.$$

Его можно записать в компактном виде через след матрицы как

$$\mathrm{d}f(A) = \mathrm{tr}(\nabla_A f^T \,\mathrm{d}A),$$

Вполне естественен вопрос — а почему это можно записать именно так? Давайте попробуем увидеть этот факт на каком-нибудь простом примере. Пусть у нас есть две матрицы

$$A_{[2\times3]} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \qquad B_{[2\times3]} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix}.$$

Посмотрим на то, как выглядит $\mathrm{tr}(B^T \,\mathrm{d}A)$. Как это ни странно, он совпадает с дифференциалом

$$\mathrm{tr}(B^T \,\mathrm{d}A) = \mathrm{tr}\left(\begin{pmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \\ b_{13} & b_{23} \end{pmatrix} \begin{pmatrix} \mathrm{d}a_{11} & \mathrm{d}a_{12} & \mathrm{d}a_{13} \\ \mathrm{d}a_{21} & \mathrm{d}a_{22} & \mathrm{d}a_{23} \end{pmatrix} \right),$$

при произведении на выходе получаем матрицу размера 3×3

$$\begin{pmatrix} b_{11}\,\mathrm{d}a_{11} + b_{21}\,\mathrm{d}a_{21} & b_{11}\,\mathrm{d}a_{12} + b_{21}\,\mathrm{d}a_{22} & b_{11}\,\mathrm{d}a_{13} + b_{21}\,\mathrm{d}a_{23} \\ b_{12}\,\mathrm{d}a_{11} + b_{22}\,\mathrm{d}a_{21} & b_{12}\,\mathrm{d}a_{12} + b_{22}\,\mathrm{d}a_{22} & b_{12}\,\mathrm{d}a_{13} + b_{22}\,\mathrm{d}a_{23} \\ b_{13}\,\mathrm{d}a_{11} + b_{23}\,\mathrm{d}a_{21} & b_{13}\,\mathrm{d}a_{12} + b_{23}\,\mathrm{d}a_{22} & b_{13}\,\mathrm{d}a_{13} + b_{23}\,\mathrm{d}a_{23} \end{pmatrix}.$$

Когда мы берём её след, остаётся сумма элементов по диагонали. Это и есть требуемый дифференциал. Дальше мы периодически будем пользоваться таким приёмом.

Например, величину $||X - A||^2 = \sum_{i,j}(x_{ij} - a_{ij})^2$ можно записать в матричном виде как $\mathrm{tr}((X - A)^T(X - A))$.

5. В таблице осталось ещё несколько ситуаций, которые остались вне поля нашего зрения. Например, давайте посмотрим на ситуацию когда отображение бьёт из матриц в векторы $f(A) : \mathbb{R}^{n\times k} \to \mathbb{R}^m$.

   Тогда A матрица, а $f(A)$ вектор. Нам надо найти производную каждого элемента из вектора $f(A)$ по каждому элементу из матрицы A. Получается, что $\frac{\partial f}{\partial A}$ — это трёхмерная структура. Мы с такими ситуациями встречаться не будем, поэтому опустим их.

Свойства матричных дифференциалов очень похожи на свойства обычных. Надо только не забыть, что мы работаем с матрицами.

$\mathrm{d}(AB) = \mathrm{d}A B + A\,\mathrm{d}B, \quad \mathrm{d}A B \neq B\,\mathrm{d}A$
$\mathrm{d}(\alpha A + \beta B) = \alpha\,\mathrm{d}A + \beta\,\mathrm{d}B$
$\mathrm{d}(A^T) = (\,\mathrm{d}A)^T$
$\mathrm{d}C = 0, \quad C -$ матрица из констант

Чтобы доказать все эти свойства достаточно просто аккуратно расписать их. Кроме этих правил нам понадобится пара трюков по работе со скалярами. Если s — скаляр размера 1×1, тогда $s^T = s$ и $\operatorname{tr}(s) = s$.

С помощью этих преобразований мы будем приводить дифференциалы к каноническому виду и вытаскивать из них производные.

Задача 1.1. *Пусть $a \in \mathbb{R}^n$ — вектор параметров, а $x \in \mathbb{R}^n$ — вектор переменных. Рассмотрим функцию, которая представляет из себя их скалярное произведение $f(x) = a^T x$. Нужно найти её производную по вектору переменных $\nabla_x f(x)$.*

Решение. Функция $f(x)$ бьёт из векторов в скаляры $f(x) : \mathbb{R}^n \to \mathbb{R}$. Если мы хотим найти производную функции $f(x_1, x_2, \ldots, x_n)$, нам надо взять производную по каждому аргументу и выписать градиент. Можно расписать умножение одного вектора на другой в виде привычной нам формулы

$$\underset{[1\times 1]}{f(x)} = \underset{[1\times n]}{a^T} \cdot \underset{[n\times 1]}{x} = \begin{pmatrix} a_1 & a_2 & \ldots & a_n \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} = a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n.$$

Из неё чётко видно, что $\frac{\partial f}{\partial x_i} = a_i$

$$\nabla_x f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \ldots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \ldots \\ a_n \end{pmatrix} = a,$$

теперь можно записать дифференциал

$$\mathrm{d}f = a^T \,\mathrm{d}x = \frac{\partial f}{\partial x_1} \cdot \mathrm{d}x_1 + \ldots + \frac{\partial f}{\partial x_n} \cdot \mathrm{d}x_n = a_1 \cdot \mathrm{d}x_1 + \ldots + a_n \cdot \mathrm{d}x_n.$$

В то же самое время можно было бы просто воспользоваться правилами нахождения матричных дифференциалов

$$\mathrm{d}f = \mathrm{d}a^T x = a^T \,\mathrm{d}x = \nabla_x f^T \,\mathrm{d}x,$$

откуда $\nabla_x f = a$. При таком подходе нам не надо анализировать каждую частную производную по отдельности. Мы находим все производные за раз.

■

Задача 1.2. *Пусть $f(x) = x^T Ax$, где x вектор размера $1 \times n$, A матрица размера $n \times n$. Найдите производную $\nabla_x f(x)$.*

Решение. Функция бьёт из векторов в скаляры. Попробуем перемножить все матрицы и расписать её в явном виде по аналогии со скалярным произведением

$$\underset{[1\times 1]}{f(x)} = \underset{[1\times n]}{x^T} \cdot \underset{[n\times n]}{A} \cdot \underset{[n\times 1]}{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \cdot x_i \cdot x_j.$$

Если продолжить в том же духе, мы сможем найти все частные производные, а потом назад вернём их в матрицу. Однако это неудобно. Всё было записано в красивом компактном матричном виде, а мы это испортили. Более того, если множителей будет больше, тогда суммы станут совсем громоздкими, и мы легко запутаемся.

При этом, если воспользоваться правилами работы с матричными дифференциалами, мы легко получим ответ

$$\mathrm{d}f = \mathrm{d}x^T Ax = \mathrm{d}(x^T)Ax + x^T\,\mathrm{d}(Ax) = \mathrm{d}(x^T)Ax + x^T \underbrace{\mathrm{d}(A)}_{\mathrm{d}A=0} x + x^T A\,\mathrm{d}(x).$$

Заметим, что $\mathrm{d}(x^T)Ax$ это скаляр. Его транспонирование никак не повлияет на результат

$$\mathrm{d}f = \mathrm{d}(x^T)Ax + x^T A\,\mathrm{d}(x) = x^T A^T\,\mathrm{d}x + x^T A\,\mathrm{d}x = x^T(A^T + A)\,\mathrm{d}x.$$

Мы нашли матричный дифференциал и свели его к каноничной форме

$$\mathrm{d}f = \nabla_x^T f\,\mathrm{d}x = x^T(A^T + A)\,\mathrm{d}x$$

Получается, что искомая производная $\nabla_x f = (A + A^T)x$. Обратите внимание, что размерности не нарушены, и мы получили столбец из производных, то есть искомый градиент нашей функции $f(x)$.

■

Задача 1.3. *Пусть $f(x) = x^T Ax$, где $x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$. Найдите вторую производную по x.*

Решение. Чтобы найти вторую производную, надо продифференцировать первую производную. Первая производная $g(x) = (A + A^T)x$ бьёт из векторов в векторы. Приведём дифференциал к каноническому виду

$$\mathrm{d}g(x) = \mathrm{d}(A + A^T)x = (A + A^T)\,\mathrm{d}x.$$

Выходит, что матрица из вторых производных для функции $f(x)$ выглядит как $A + A^T$. Обратите внимание, что для этой ситуации в каноническом виде нет транспонирования.

■

Задача 1.4. *Пусть $f(X) = a^T X A X a$, где $a \in \mathbb{R}^n, X \in \mathbb{R}^{n\times n}$. Необходимо найти производную $\nabla_X f$.*

Решение. Функция бьёт из матриц в скаляры. Дифференциал будет по своей размерности совпадать со скаляром. Производная будет размера матрицы

$$\mathrm{d}f(X) = \mathrm{d}(a^T X A X a) = a^T \,\mathrm{d}(X) A X a + a^T X A \,\mathrm{d}(X) a.$$

Оба слагаемых, которые мы получаем после перехода к дифференциалу — скаляры. Мы хотим представить дифференциал в виде $\mathrm{tr}(\text{нечто}\,\mathrm{d}X)$. След скаляра — это снова скаляр. Получается, что мы бесплатно можем навесить над правой частью нашего равенство знак следа и воспользоваться его свойствами

$$\begin{aligned}\mathrm{d}f(X) = \mathrm{d}(a^T X A X a) &= \mathrm{tr}(a^T \,\mathrm{d}(X) A X a) + \mathrm{tr}(a^T X A \,\mathrm{d}(X) a) = \\ &= \mathrm{tr}(A X a a^T \,\mathrm{d}(X)) + \mathrm{tr}(a a^T X A \,\mathrm{d}(X)) = \\ &= \mathrm{tr}(A X a a^T \,\mathrm{d}(X) + a a^T X A \,\mathrm{d}(X)) = \mathrm{tr}((A X a a^T + a a^T X A)\,\mathrm{d}(X)).\end{aligned}$$

Производная найдена, оказалось что это

$$\nabla_X f = (A X a a^T + a a^T X A)^T = a a^T X^T A^T + A^T X a a^T.$$

■

Задача 1.5. *Пусть $f(x) = x x^T x$, где $x \in \mathbb{R}^n$. Необходимо найти производную $\nabla_x f$.*

Решение. Функция бьёт из векторов в векторы.

$$\underset{[n\times 1]}{f(x)} = \underset{[n\times 1]}{x}\,\underset{[1\times n]}{x^T}\,\underset{[n\times 1]}{x}.$$

Берём дифференциал

$$\mathrm{d}f(x) = \mathrm{d}x x^T x = \mathrm{d}x x^T x + x\,\mathrm{d}x^T x + x x^T \,\mathrm{d}x.$$

В первом слагаемом пользуемся тем, что $x^T x$ скаляр и его можно вынести перед дифференциалом. Этот скаляр умножается на каждый элемент вектора. Дальше мы захотим вынести дифференциал за скобку, чтобы не испортить матричное сложение, подчеркнём факт этого перемножения на каждый элемент единичной матрицей. Во втором слагаемом пользуемся тем, что $\mathrm{d}x^T x$ скаляр и транспонируем его

$$\mathrm{d}f(x) = \underset{[1\times 1]}{x^T x}\,\underset{[n\times n]}{I_n}\,\underset{[n\times 1]}{\mathrm{d}x} + x x^T \,\mathrm{d}x + x x^T \,\mathrm{d}x = (x^T x I_n + 2 x x^T)\,\mathrm{d}x.$$

Обратите внимание, что без единичной матрицы размерности у сложения поломаются. Получается, что наша производная выглядит как $\mathfrak{J} = x^T x I_n + 2 x x^T$

■

Найдём несколько табличных производных, которыми мы дальше будем активно пользоваться: производную обратной матрицы, определителя и следа.

Задача 1.6. *Пусть* $f(A) = A^{-1}$, *где* $A \in \mathbb{R}^{n\times n}$. *Необходимо найти* $\nabla_A f(A)$.

Решение. Вспомним, что производная константы равна нолю. Обратная матрица определяется как $A^{-1} \cdot A = I_n$, где I_n — единичная матрица. Берём дифференциал с обеих сторон нашего равенства

$$\mathrm{d}A^{-1}A + A^{-1}\,\mathrm{d}A = \mathrm{d}I_n = 0,$$

отсюда получаем что $\mathrm{d}A^{-1} = -A^{-1}\,\mathrm{d}AA^{-1}$. Везде, где мы будем встречать дифференциал обратной матрицы, мы будем использовать это значение.

■

Задача 1.7. *Пусть* $A \in \mathbb{R}^{n\times n}$. *Необходимо найти* $\nabla_A \det A$.

Решение. Определитель — это функция, которая бьёт из матриц в скаляры. 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 $\nabla_A \det A = (\det A)A^{-T}$. При этом, так как функция бьёт из матриц в скаляры дифференциал можно записать как $\mathrm{d}\det A = \mathrm{tr}(\det(A)A^{-1}\,\mathrm{d}A)$.

■

Задача 1.8. *Пусть* $A \in \mathbb{R}^{n\times n}$. *Необходимо найти* $\nabla_A \mathrm{tr}(A)$.

Решение. По аналогии с определителем след бьёт из пространства матриц в пространство скаляров представляет из себя сумму диагональных элементов. Получается, что $\mathrm{d}(\mathrm{tr}\, A) = \mathrm{tr}(I_n\,\mathrm{d}A)$ и $\nabla_A \mathrm{tr}\, A = I_n$.

■

Задача 1.9. *Пусть* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$. *Необходимо найти* $\nabla_A tr(AB)$.

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

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

получаем

$$\mathrm{d}\,\mathrm{tr}(AB) = \mathrm{tr}(\,\mathrm{d}AB) = \mathrm{tr}(B\,\mathrm{d}A),$$

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

■

Задача 1.10. *Пусть* $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, $y \in \mathbb{R}^m$. *Необходимо найти* $\nabla_A \operatorname{tr}(x^T A y)$.

Решение. Воспользовавшись циклическим свойством следа и результатом предыдущей задачи, получаем

$$\mathrm{d}\operatorname{tr}(x^T A y) = \operatorname{tr}(\mathrm{d}x^T A y) = \operatorname{tr}(y x^T \mathrm{d}A),$$

то есть $\nabla_A \operatorname{tr}(x^T A y) = x y^T$.

■

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

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

где $\mathfrak{J}_f(x) = \left(\frac{\partial f_i(x)}{\partial x_j}\right)_{i,j=1}^{m,n}$ — 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 1.11. *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. Воспользуемся правилом взятия производной сложной функции и производной скалярного произведение из задачи выше

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

■

§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 дифференциал функции ошибки по w:

$$\begin{aligned}
&\mathrm{d}_w Q = \mathrm{d}_w[(y - Xw)^T(y - Xw)] = \\
&= \mathrm{d}_w[(y - Xw)^T](y - Xw) + (y - Xw)^T \mathrm{d}_w[(y - Xw)] = \\
&= \mathrm{d}_w[(-Xw)^T](y - Xw) - (y - Xw)^T X \,\mathrm{d}w = \\
&\qquad = -\mathrm{d}w^T X^T(y - Xw) - (y - Xw)^T X \,\mathrm{d}w = -2(y - Xw)^T X \,\mathrm{d}w.
\end{aligned}$$

Тут мы воспользовались тем, что $\mathrm{d}w^T X^T(y - Xw)$ это скаляр и его можно транспонировать. Приравняем производную к нулю, чтобы найти минимум для w. Получается система уравнений

$$2X^T(y - Xw) = 0 \quad \Rightarrow \quad X^T y = X^T X w \quad \Rightarrow \quad w = (X^T X)^{-1} X^T y.$$

При решении системы мы сделали предположение, что матрица $X^T X$ обратима. Это так, если в матрице X нет линейно зависимых столбцов, а также наблюдений больше чем переменных.

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.

Если бы аналитического решения не существовало, мы могли бы найти точку оптимума с помощью градиентного спуска. Его шаг выглядел бы как

$$w_t = w_{t-1} + \gamma \cdot 2X^T(y - Xw), \quad \text{здесь } \gamma \text{ — это скорость обучения.}$$

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. Найдём вторую производную

$$\mathrm{d}_w[-2X^T(y - Xw)] = 2X^T X \,\mathrm{d}w.$$

Выходит, что

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

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.

Список литературы

[1] *Родоманов А.О.* (2017). Матрично-векторное дифференцирование. // http://www.machinelearning.ru/wiki/images/5/50/MOMO17_Seminar2.pdf.

[2] *Kaare B. Petersen, Michael S. Pedersen* (2012). The Matrix Cookbook. // https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.