

Семинар. Матрично-векторное дифференцирование

ММП ВМК МГУ

Осень 2025

Содержание

1	Мотивация	2
2	Общее определение дифференцируемости	3
2.1	Примеры для конкретных U, V	3
3	Аппарат матрично-векторного дифференцирования	4
3.1	Правила дифференцирования	4
3.1.1	Производная суммы / разности, умножения на число	4
3.1.2	Производная произведения	5
3.1.3	Производная частного	5
3.1.4	Производная композиции	5
3.2	Табличные производные	6
3.2.1	Дифференцирование линейной функции	6
3.2.2	Дифференцирование квадратичной функции	6
3.2.3	Задача. Линейная регрессия	7
3.2.4	Дифференцирование определителя	7
3.2.5	Дифференцирование обратной матрицы	9
3.2.6	Задача. ML-оценка для нормального распределения	9
4	О гессиане	10
4.1	Определение	11
4.2	Способы нахождения	11
4.3	Задача. Куб второй нормы	12
4.4	Задача. Линейная регрессия. Продолжение	12
5	Еще немного задач	13
5.1	Задача. Норма Фробениуса	13
5.2	Задача. Знакоопределенность гессиана	14

1 Мотивация

Представим, что нам удалось предложить для решения некоторой задачи модель машинного обучения. После этого требуется найти такие ее параметры, которые бы минимизировали некоторую функцию, имеющую смысл потерь:

$$\mathcal{L}(\mathbf{X}, \theta) \rightarrow \min_{\theta}$$

При ее решении (аналитическом или численном) почти наверное возникает потребность в нахождении градиента $\nabla_{\theta} \mathcal{L}$ функции. Какие существуют варианты решения этой задачи?

Из курса математического анализа нам известно, что компоненты вектора градиента совпадают с частными производными по соответствующим переменным:

$$(\nabla_{\theta} \mathcal{L})_i = \frac{\partial \mathcal{L}}{\partial \theta_i}$$

Попробуем, используя данный подход, получить выражение для градиента следующей функции векторного аргумента:

$$f(x_1, \dots, x_n) = f(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle, \quad \mathbf{A} \in \mathbb{R}^{n \times n}$$

Для этого найдем выражение для функции f через компоненты аргумента:

$$f(x_1, \dots, x_n) = \sum_{i,j} a_{ij} x_i x_j$$

Продифференцируем его по переменной x_k :

$$\begin{aligned} \frac{\partial f}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i,j} a_{ij} x_i x_j = \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} a_{ik} x_i x_k + \sum_{j \neq k} a_{kj} x_k x_j + a_{kk} x_k^2 + \sum_{i \neq k, j \neq k} a_{ij} x_i x_j \right] = \\ &= \sum_{i \neq k} a_{ik} x_i + \sum_{j \neq k} a_{kj} x_j + 2a_{kk} x_k = \left(\sum_{i \neq k} a_{ik} x_i + a_{kk} x_k \right) + \left(\sum_{j \neq k} a_{kj} x_j + a_{kk} x_k \right) = \\ &= \sum_i a_{ik} x_i + \sum_j a_{kj} x_j = (\mathbf{A}^{\top} \mathbf{x})_k + (\mathbf{A} \mathbf{x})_k = ([\mathbf{A} + \mathbf{A}^{\top}] \mathbf{x})_k \end{aligned}$$

Отсюда мы получаем следующее выражение для градиента:

$$\nabla f = (\mathbf{A} + \mathbf{A}^{\top}) \mathbf{x}$$

Обратим внимание на то, что выражение и для функции, и для градиента имеют очень простой вид в матричной форме, а при работе на уровне координат выражения получались сложнее.

Этот пример наталкивает нас на идею развития другого, более общего аппарата дифференцирования, который позволяет «общаться» с градиентом на более высоком уровне и как следствие получать для его вычисления выражения в терминах матрично-векторных операций, удобных для аналитической работы и вычислительно эффективных.

2 Общее определение дифференцируемости

Попробуем естественным образом обобщить понятие производной функции одной переменной. Для этого вспомним его

Определение. Говорят, что функция $f : B(a) \rightarrow \mathbb{R}$ дифференцируема в точке $a \in \mathbb{R}$, если ее приращение в достаточно малой окрестности с точностью до членов более высокого порядка малости описывается линейной функцией приращения аргумента:

$$f(a+h) - f(a) = Ah + \bar{o}(h), \quad h \rightarrow 0$$

Посмотрим на A не как на число, а как на линейный оператор $L_A(h) = Ah$. По аналогии продолжим определение на более абстрактные пространства. Мы бы хотели говорить, что отображение дифференцируемо в точке, если его приращение линейно зависит от приращения аргумента (для этого потребуем линейности пространства) с точностью до членов большего порядка малости (для этого – нормированности). Как итог, формулируем

Определение. Пусть U, V – линейные нормированные пространства. Говорят, что функция $f : B(x) \subset U \rightarrow V$ дифференцируема в точке $x \in U$, если

$$f(x+h) - f(x) = Lh + \bar{o}(\|h\|),$$

где $L : U \rightarrow V$ – линейный оператор, называемый *первой производной* функции, \bar{o} понимается в смысле нормированных пространств. Оператор L обозначается $df(x)$, а его действие на приращение – $df(x)[h]$.

Замечание 1. Существуют и другие обозначения для первой производной, в том числе иногда приращение для однородности записи обозначают за dx . Иногда приращение опускают для компактности записей, мы тоже будем так делать при решении задач, в общем случае важно помнить, что первая производная это оператор, что подчеркивается ее действием на аргумент. Так, формально то, что первая производная функции x^2 это $2x$ следует записать так:

$$d(x^2)[dx] = 2x dx,$$

однако при решении задач мы пишем

$$d(x^2) = 2x dx$$

Замечание 2. Существуют и другие названия для первой производной: производная по Фреше, дифференциал, дифференциальный оператор. Все они «об одном и том же», но в конкретной области математики / размерности могут обозначать чуть разные вещи. Сегодня наша цель познакомиться с двумя основными объектами: линейным оператором, отвечающим за линейное приближение, и градиентом. Иногда понятие первой производной будет заменяться синонимичными.

2.1 Примеры для конкретных U, V

Для лучшего понимания определения рассмотрим его для нескольких конкретных пространств U и V .

- $U = V = \mathbb{R}$. Определение совпадает с определением дифференцируемости одномерной функции действительного аргумента. $df(x)$ – число, dx – число.

- $U = \mathbb{R}^n, V = \mathbb{R}$. Определение совпадает с определением дифференцируемости одномерной функции многих переменных / векторного аргумента, рассмотренным в курсе математического анализа. В этом случае $df(\mathbf{x})$ – линейный функционал, действующий на $d\mathbf{x}$ – вектор. Линейный функционал во всяком конечномерном пространстве представим скалярным произведением на некоторый его элемент, называемый *градиентом*:

$$df(\mathbf{x})[d\mathbf{x}] = \langle \nabla f(\mathbf{x}), d\mathbf{x} \rangle = \nabla f(\mathbf{x})^\top d\mathbf{x} = d\mathbf{x}^\top \nabla f(\mathbf{x})$$

Итак, ∇f в этом случае вектор.

- $U = \mathbb{R}^{n \times m}, V = \mathbb{R}$. Учитывая изоморфность пространств $\mathbb{R}^{n \times m}$ и \mathbb{R}^{nm} , данный случай ничем не отличается от предыдущего, просто переменные организованы не в вектор, а в матрицу, из-за чего скалярное произведение принимает другой вид ($df(\mathbf{X})$ – все еще линейный функционал, уже действующий на $d\mathbf{X}$ – матрицу):

$$df(\mathbf{X})[d\mathbf{X}] = \langle \nabla f(\mathbf{X}), d\mathbf{X} \rangle = \text{tr} [d\mathbf{X}^\top \nabla f(\mathbf{X})] = \text{tr} [\nabla f(\mathbf{X})^\top d\mathbf{X}],$$

где $\text{tr}[\cdot]$ – след матрицы.

- $U = \mathbb{R}^n, V = \mathbb{R}^m$. Вектор-функцию с образами в \mathbb{R}^m в контексте дифференцируемости можно рассматривать как набор из m одномерных функций. Производная $df(\mathbf{x})$ – линейный оператор $\mathbb{R}^n \rightarrow \mathbb{R}^m$ – представляется в виде $m \times n$ -матрицы \mathbf{J}_f , называемой *якобианом*:

$$df(\mathbf{x})[d\mathbf{x}] = \mathbf{J}_f(\mathbf{x})d\mathbf{x}$$

3 Аппарат матрично-векторного дифференцирования

Как мы будем действовать, если нам нужно продифференцировать одномерную функцию действительного аргумента? Конечно, мы не будем делать это по определению, а воспользуемся правилами дифференцирования и таблицей стандартных производных. Если нам удастся получить их аналоги в матрично-векторном случае, то мы решим задачу аналитического вычисления градиента, превратив это занятие в такое же ремесло, как и дифференцирование функции одной переменной.

3.1 Правила дифференцирования

Еще немного «посидим» в абстракции и выведем правила дифференцирования, аналогичные тем, которые мы получали в 1 семестре в курсе математического анализа.

3.1.1 Производная суммы / разности, умножения на число

Первая производная линейна по функции, от которой она берется:

$$d(\alpha f + \beta g)(x)[h] = \alpha df(x)[h] + \beta dg(x)[h]$$

Не будем на этом подробно останавливаться, желающим еще раз себя проверить в понимании определения остается в качестве упражнения.

3.1.2 Производная произведения

Правило дифференцирования произведения аналогично одномерному случаю. Идея вывода та же: применить известный прием «добавить-вычесть». Важно обратить внимание на то, что вообще говоря, не понятно, о какой операции умножения идет речь. Дальнейшие рассуждения проведены при достаточно скромных ограничениях на ее структуру, что делает их верными, например, для произведения матриц (помним про отсутствие его коммутативности) и для скалярного произведения.

$$\begin{aligned} f(x+h)g(x+h) - f(x)g(x) &= \\ &= f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x) = \\ &= (f(x+h) - f(x))g(x+h) + f(x)(g(x+h) - g(x)) = \end{aligned}$$

Теперь расписываем выражения в скобках по определению первой производной, появляются $df(x)[h]$ и $dg(x)[h]$ соответственно.

$$= (df(x)[h] + \bar{o}(\|h\|))(g(x) + \bar{o}(1)) + f(x)(dg(x)[h] + \bar{o}(\|h\|)) =$$

Выделяем главную линейную часть:

$$= df(x)[h] g(x) + f(x) dg(x)[h] + \bar{o}(\|h\|)$$

3.1.3 Производная частного

Выводится для случая деления на скалярную функцию из правила дифференцирования произведения:

$$\begin{aligned} d\left(\frac{f}{\varphi}\right)(x)[h] &= d\left(f \cdot \frac{1}{\varphi}\right)(x)[h] = df(x)[h] \frac{1}{\varphi(x)} - f(x) \frac{1}{\varphi^2(x)} d\varphi(x)[h] = \\ &= \frac{df(x)[h] \varphi(x) - f(x) d\varphi(x)[h]}{\varphi^2(x)} \end{aligned}$$

3.1.4 Производная композиции

Остался последний и, возможно, самый мощный инструмент при дифференцировании – правило производной композиции. Итак, пусть $g : U \rightarrow V$ – дифференцируема в точке x , $f : V \rightarrow W$ – дифференцируема в точке $g(x)$, тогда первая производная функции $f \circ g : U \rightarrow W$ в точке x имеет следующий вид:

$$d(f \circ g)(x)[h] = df(g(x)) [dg(x)[h]]$$

Легко запоминается, ведь по записи почти в точности повторяет аналогичный результат в одномерном случае:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

3.2 Табличные производные

Перед выводом таблицы производных укажем два полезных способа самопроверки при выполнении операций матрично-векторного дифференцирования.

1. *Проверяйте размерность результата.* Например, при дифференцировании функции $\mathbb{R}^n \rightarrow \mathbb{R}$ в качестве градиента должен получиться вектор из \mathbb{R}^n , не вектор другой размерности, не матрица и не число.
2. *Проверяйте результат на одномерном случае.* Если размерность градиента сошлась, то полезно бывает посмотреть на то, совпадают ли полученная формула и производная изначальной функции в случае $n = 1$.

3.2.1 Дифференцирование линейной функции

Найдем, чему равна производная линейной функции. Итак, пусть f – линейное отображение, тогда:

$$f(x + h) - f(x) = f(x) + f(h) - f(x) = f(h) = f(h) + 0,$$

т.е. первая производная линейной функции действует на приращение аргумента, как и сама функция. Примеры:

$$\begin{aligned} d\langle \mathbf{c}, \mathbf{x} \rangle &= \langle \mathbf{c}, d\mathbf{x} \rangle \\ d \operatorname{tr} [\mathbf{A}\mathbf{X}\mathbf{B}] &= \operatorname{tr} [\mathbf{A} d\mathbf{X} \mathbf{B}] \end{aligned}$$

3.2.2 Дифференцирование квадратичной функции

Найдем, чему равна производная квадратичной функции

$$f(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$$

Распишем приращение

$$\begin{aligned} \langle \mathbf{A}(\mathbf{x} + \mathbf{h}), \mathbf{x} + \mathbf{h} \rangle - \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle &= \langle \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{h}, \mathbf{x} + \mathbf{h} \rangle - \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \\ &= \langle \mathbf{A}\mathbf{x}, \mathbf{h} \rangle + \langle \mathbf{A}\mathbf{h}, \mathbf{x} \rangle + \langle \mathbf{A}\mathbf{h}, \mathbf{h} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{h} \rangle + \langle \mathbf{h}, \mathbf{A}^\top \mathbf{x} \rangle + \langle \mathbf{A}\mathbf{h}, \mathbf{h} \rangle = \\ &= \langle \mathbf{A}\mathbf{x}, \mathbf{h} \rangle + \langle \mathbf{A}^\top \mathbf{x}, \mathbf{h} \rangle + \langle \mathbf{A}\mathbf{h}, \mathbf{h} \rangle = \\ &= \langle (\mathbf{A} + \mathbf{A}^\top) \mathbf{x}, \mathbf{h} \rangle + \langle \mathbf{A}\mathbf{h}, \mathbf{h} \rangle \end{aligned}$$

Если мы покажем, что второе слагаемое является \bar{o} от $\|\mathbf{h}\|$, то мы по определению сможем утверждать, что

$$df(\mathbf{x}) = d\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \langle (\mathbf{A} + \mathbf{A}^\top) \mathbf{x}, d\mathbf{x} \rangle$$

Действительно, по неравенству Коши-Буняковского и определению нормы:

$$|\langle \mathbf{A}\mathbf{h}, \mathbf{h} \rangle| \leq \|\mathbf{A}\mathbf{h}\| \|\mathbf{h}\| \leq \|\mathbf{A}\| \|\mathbf{h}\|^2 = \bar{o}(\|\mathbf{h}\|)$$

Самопроверка: градиент $(\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ – результат произведения квадратной матрицы на вектор, значит вектор той же размерности; в одномерном случае $f(x) = ax^2$, ее производная $f'(x) = 2ax = (a + a)x$.

3.2.3 Задача. Линейная регрессия

На предыдущем семинаре Вы подробно изучали линейную регрессию. Вы получили выражение для аналитического решения, используя сведения о псевдорешениях из линейной алгебры. Сейчас у нас достаточно результатов, чтобы получить этот результат с помощью средств матрично-векторного дифференцирования. Итак, рассмотрим функцию потерь для линейной регрессии с ℓ_2 -регуляризацией:

$$\mathcal{L}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$

Представляя квадрат второй нормы как скалярный квадрат, имеем следующее:

$$d\mathcal{L} = d\langle \mathbf{X}\mathbf{w} - \mathbf{y}, \mathbf{X}\mathbf{w} - \mathbf{y} \rangle + \lambda d\langle \mathbf{w}, \mathbf{w} \rangle =$$

Раскрываем скалярный квадрат в первом слагаемом

$$= d\langle \mathbf{X}\mathbf{w}, \mathbf{X}\mathbf{w} \rangle - 2d\langle \mathbf{X}\mathbf{w}, \mathbf{y} \rangle + d\langle \mathbf{y}, \mathbf{y} \rangle + \lambda d\langle \mathbf{w}, \mathbf{w} \rangle =$$

«Подгоняем» переносом через транспонирование функционалы к ранее вычисленным таблицам:

$$= d\langle \mathbf{X}^\top \mathbf{X}\mathbf{w}, \mathbf{w} \rangle - 2d\langle \mathbf{X}^\top \mathbf{y}, \mathbf{w} \rangle + \lambda d\langle \mathbf{I}\mathbf{w}, \mathbf{w} \rangle =$$

Применяем формулы:

$$\begin{aligned} &= 2\langle \mathbf{X}^\top \mathbf{X}\mathbf{w}, d\mathbf{w} \rangle - 2\langle \mathbf{X}^\top \mathbf{y}, d\mathbf{w} \rangle + 2\lambda \langle \mathbf{w}, d\mathbf{w} \rangle = \\ &= \langle 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}, d\mathbf{w} \rangle \end{aligned}$$

Как итог

$$\nabla \mathcal{L}(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}$$

Приравнивая его к нулю, получаем изученное в рамках семинара по линейной регрессии аналитическое решение:

$$\mathbf{w}_{\text{opt}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

Самопроверка: полученный градиент есть разность двух произведений матрицы на вектор – тоже вектор; в одномерном случае функция потерь линейной регрессии имеет вид

$$l(w) = (xw - y)^2 + \lambda w^2$$

Ее производная:

$$l'(w) = 2(xw - y)x + 2\lambda w = 2(x \cdot x + \lambda \cdot 1) - 2xy$$

3.2.4 Дифференцирование определителя

Мы уже упоминали то, что переменные могут быть собраны в матрицу. Попробуем продифференцировать функцию матричного аргумента.

В качестве примера рассмотрим определитель матрицы. Мы приведем 2 решения, одно будет аналитическими в смысле соответствия идеологии подхода, который мы развиваем, другое – в одну строчку через факт из курса линейной алгебры, используем его для проверки. Итак, рассмотрим приращение:

$$\det(\mathbf{X} + \mathbf{H}) - \det \mathbf{X} =$$

Будем рассматривать случай обратимых \mathbf{X} , они образуют открытое множество.

$$\begin{aligned} &= \det(\mathbf{X}(\mathbf{I} + \mathbf{X}^{-1}\mathbf{H})) - \det \mathbf{X} = \det \mathbf{X} \det(\mathbf{I} + \mathbf{X}^{-1}\mathbf{H}) - \det \mathbf{X} = \\ &= \det \mathbf{X} [\det(\mathbf{I} + \mathbf{X}^{-1}\mathbf{H}) - 1] = \end{aligned}$$

Пусть $\{\lambda_i\}$ – собственные числа матрицы $\mathbf{X}^{-1}\mathbf{H}$ с учетом кратности, тогда $\{1 + \lambda_i\}$ – это собственные числа матрицы $\mathbf{I} + \mathbf{X}^{-1}\mathbf{H}$. Из курса линейной алгебры нам известно, что определитель матрицы есть произведение ее собственных значений. Используя это, продолжим наши выкладки:

$$\begin{aligned} &= \det \mathbf{X} \left[\prod_i (1 + \lambda_i) - 1 \right] = \det \mathbf{X} \left[1 + \sum_i \lambda_i + \sum_{i < j} \lambda_i \lambda_j + \sum_{i < j < k} \lambda_i \lambda_j \lambda_k + \dots - 1 \right] = \\ &= \det \mathbf{X} \left[\sum_i \lambda_i + \sum_{i < j} \lambda_i \lambda_j + \sum_{i < j < k} \lambda_i \lambda_j \lambda_k + \dots \right] = \end{aligned}$$

Используя следующий факт из линейной алгебры:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \Rightarrow |\lambda| \|\mathbf{v}\| = \|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\| \|\mathbf{v}\| \Rightarrow |\lambda| \leq \|\mathbf{A}\|$$

несложно видеть, что все члены, начиная со второго будут $\bar{o}(\|\mathbf{H}\|)$, осталось вспомнить, что сумма собственных значений есть след матрицы, и завершить вывод:

$$= \det \mathbf{X} \operatorname{tr} [\mathbf{X}^{-1}\mathbf{H}] + \bar{o}(\|\mathbf{H}\|) = \det \mathbf{X} \langle \mathbf{X}^{-\top}, \mathbf{H} \rangle + \bar{o}(\|\mathbf{H}\|)$$

В первом решении мы использовали общие матричные объекты (собственные значения), не обращаясь к элементной структуре, как и хотели в начале семинара, но в конкретно этой задаче есть путь короче, который лежит через поэлементное рассмотрение.

Итак, разложим определитель матрицы $\mathbf{X} = (x_{ij})$ по теореме Лапласа по i строке:

$$\det \mathbf{X} = \sum_j x_{ij} A_{ij},$$

где A_{ij} – соответствующее алгебраическое дополнение. Возьмем частную производную по x_{ij} :

$$\frac{\partial \det \mathbf{X}}{\partial x_{ij}} = A_{ij}$$

Для завершения вывода осталось вспомнить, как через алгебраические дополнения можно выразить элементы обратной матрицы:

$$(\mathbf{X}^{-1})_{ij} = \frac{A_{ji}}{\det \mathbf{X}} \Rightarrow A_{ij} = \det \mathbf{X} (\mathbf{X}^{-1})_{ji} = \det \mathbf{X} (\mathbf{X}^{-\top})_{ij}$$

Как итог получаем то же выражение, что и в первом решении:

$$\frac{\partial \det \mathbf{X}}{\partial x_{ij}} = \det \mathbf{X} (\mathbf{X}^{-\top})_{ij} \Rightarrow d \det \mathbf{X} = \det \mathbf{X} \langle \mathbf{X}^{-\top}, d\mathbf{X} \rangle$$

3.2.5 Дифференцирование обратной матрицы

В курсе математического анализа мы часто дифференцировали выражения, из которых интересующая нас функция не выражалась явно, с целью нахождения ее производной. Воспользуемся аналогичным трюком для нахождения первой производной обратной матрицы:

$$\begin{aligned} d\mathbf{X}\mathbf{X}^{-1} &= d\mathbf{I} \\ d\mathbf{X}\mathbf{X}^{-1} + \mathbf{X}d\mathbf{X}^{-1} &= 0 \\ d\mathbf{X}^{-1} &= -\mathbf{X}^{-1}d\mathbf{X}\mathbf{X}^{-1} \end{aligned}$$

3.2.6 Задача. ML-оценка для нормального распределения

Воспользуемся полученными результатами для вычисления ML-оценки на параметры многомерного нормального распределения. Пусть $\mathbf{x}_1, \dots, \mathbf{x}_n$ – реализация независимых одинаково распределенных случайных величин из многомерного нормального распределения со средним $\boldsymbol{\mu} \in \mathbb{R}^d$ и матрицей ковариаций $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Напомним, как выглядит его функция плотности.

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Выпишем правдоподобие выборки:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}^d \sqrt{\det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} = \\ &= \frac{1}{\left[\sqrt{2\pi}^d \sqrt{\det \boldsymbol{\Sigma}} \right]^n} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \langle \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle \right\} \end{aligned}$$

Теперь перейдем к его логарифму с обратным знаком:

$$-\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{n}{2} \log \det \boldsymbol{\Sigma} + \frac{1}{2} \sum_{i=1}^n \langle \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle + \text{const}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Найдем градиенты по параметрам распределения, для этого найдем производную и представим ее в следующем виде:

$$d(-\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \langle \nabla_{\boldsymbol{\mu}} \mathcal{L}, d\boldsymbol{\mu} \rangle + \langle \nabla_{\boldsymbol{\Sigma}} \mathcal{L}, d\boldsymbol{\Sigma} \rangle$$

Итак:

$$\begin{aligned} d(-\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) &= \frac{n}{2} d(\log \det \boldsymbol{\Sigma}) + \frac{1}{2} \sum_{i=1}^n d \langle \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle = \\ &= \frac{n}{2} \frac{1}{\det \boldsymbol{\Sigma}} \det \boldsymbol{\Sigma} \langle \boldsymbol{\Sigma}^{-\top}, d\boldsymbol{\Sigma} \rangle + \frac{1}{2} \sum_{i=1}^n [\langle d\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle + \langle \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), d(\mathbf{x}_i - \boldsymbol{\mu}) \rangle] \end{aligned}$$

Рассмотрим подробнее слагаемые в сумме:

$$\langle d\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle + \langle \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), d(\mathbf{x}_i - \boldsymbol{\mu}) \rangle =$$

$$\begin{aligned}
&= \langle -\Sigma^{-1} d\Sigma \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{x}_i - \boldsymbol{\mu} \rangle + \langle -\Sigma^{-1} d\boldsymbol{\mu}, \mathbf{x}_i - \boldsymbol{\mu} \rangle + \langle \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), d(\mathbf{x}_i - \boldsymbol{\mu}) \rangle = \\
&= -\langle d\Sigma, \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} \rangle - \langle d\boldsymbol{\mu}, \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \rangle - \langle \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), d\boldsymbol{\mu} \rangle = \\
&= -\langle \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}, d\Sigma \rangle - 2\langle \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), d\boldsymbol{\mu} \rangle =
\end{aligned}$$

Подставляем в исходное выражение:

$$d(-\log \mathcal{L}(\boldsymbol{\mu}, \Sigma)) = \frac{1}{2} \left\langle n\Sigma^{-1} - \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}, d\Sigma \right\rangle - \left\langle \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}), d\boldsymbol{\mu} \right\rangle$$

Получаем следующие выражения для градиентов:

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}} \mathcal{L} &= -\Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}) \\
\nabla_{\Sigma} \mathcal{L} &= \frac{n}{2} \Sigma^{-1} - \frac{1}{2} \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}
\end{aligned}$$

Приравниваем их к нулю, используем то, что Σ положительно определена, а значит обратима. Как итог, имеем следующие оценки максимального правдоподобия на параметры $\boldsymbol{\mu}$ и Σ – выборочные среднее и дисперсию:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{\text{ML}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \\
\hat{\Sigma}_{\text{ML}} &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top
\end{aligned}$$

Замечание. Вообще говоря, мы решали задачу оптимизации с ограничением

$$\Sigma = \Sigma^\top > 0$$

То, что ответ, полученный методом приравнивания градиента нулю, оказался подходящим – не более, чем удача. Чтобы сделать решение «честным», можно воспользоваться специальными методами решения задач условной оптимизации, о которых Вам расскажут в течение ближайших семинаров, но в нашем конкретном случае можно схитрить. Если мы перепараметризуем матрицу ковариаций через ее квадратный корень

$$\Sigma = \mathbf{L}\mathbf{L}^\top, \det \mathbf{L} \neq 0$$

то сможем решать безусловную задачу оптимизации относительно новых переменных $\boldsymbol{\mu}$ и \mathbf{L} . Сделать это предлагается предлагается в качестве упражнения.

4 О гессiane

В курсе математического анализа понятие дифференцируемости в точке обобщалось на понятие производной функции, которое в свою очередь позволяло определить производные произвольных порядков, оказавшиеся крайне полезным инструментом как для анализа, так и для создания численных методов.

Посмотрим на эти идеи в матрично-векторном случае. Остановимся на второй производной. Производные порядков выше второго почти не используются на практике, к тому же аналитически представляются сильно сложнее младших братьев.

4.1 Определение

Определение строится на следующем наблюдении: «Вторая производная это первая производная первой производной». Суждение безусловно философское ☺, однако, кроме шуток, оно помогает не путаться с тем, кто есть кто.

Итак, первая производная функции f это линейный функционал $df(x)[h]$, зависящий от точки x , в которой берется производная, действующий на приращение h . Такие функционалы сами по себе образуют линейное пространство (его называют *сопряженным*), а значит, если $df[\cdot]$ определен(а) в некоторой окрестности точки x , то в ней мы можем рассмотреть отображение: $x \mapsto df(x)[\cdot]$.

Теперь предположим, что оно имеет первую производную. Запишем это:

$$df(x+h)[\cdot] - df(x)[\cdot] = d(df[\cdot])(x)[h] + \bar{o}(\|h\|)[\cdot]$$

Использование значка $[\cdot]$ еще раз обращает внимание на то, что df это оператор. В дальнейшем мы будем записывать это по-другому, вводя «как бы» два приращения, h_2 имеет смысл приращения с точки зрения дифференцирования, h_1 – приращения, на котором определены первые производные $df(x)$:

$$df(x+h_2)[h_1] - df(x)[h_1] = d^2f(x)[h_1, h_2] + \bar{o}(\|h_2\|)[h_1]$$

Теперь подумаем над тем, как выглядит вторая производная в «хороших» случаях, например, когда $U = \mathbb{R}^n, V = \mathbb{R}$. Мы знаем, что в таком случае первая производная есть скалярное произведение с вектором градиента:

$$\begin{aligned} df(\mathbf{x} + \mathbf{h}_2)[\mathbf{h}_1] - df(\mathbf{x})[\mathbf{h}_1] &= \langle \nabla f(\mathbf{x} + \mathbf{h}_2), \mathbf{h}_1 \rangle - \langle \nabla f(\mathbf{x}), \mathbf{h}_1 \rangle = \\ &= \langle \nabla f(\mathbf{x} + \mathbf{h}_2) - \nabla f(\mathbf{x}), \mathbf{h}_1 \rangle = \end{aligned}$$

Градиент это отображение $\mathbb{R}^n \rightarrow \mathbb{R}^n$, поэтому его первая производная – умножение на некоторую матрицу, которая как раз и называется *гессианом*, ее элементы – вторые частные производные функции f :

$$= \langle \nabla^2 f(\mathbf{x}) \mathbf{h}_2, \mathbf{h}_1 \rangle + \langle \bar{o}(\|\mathbf{h}_2\|), \mathbf{h}_1 \rangle$$

4.2 Способы нахождения

Рецепт поиска гессиана при найденном градиенте заключается в следующем: объявить ранее использованный dx за dx_1 и, пользуясь ранее выведенными правилами, продифференцировать градиент, обозначая новое приращение за dx_2 . Далее следует привести полученное выражение к каноническому виду

$$d^2f = \langle \nabla^2 f dx_2, dx_1 \rangle,$$

откуда и получить выражение для гессиана. Решим несколько задач по описанному алгоритму.

4.3 Задача. Куб второй нормы

Найдем градиент и гессиан следующей функции:

$$f(\mathbf{x}) = \frac{1}{3} \|\mathbf{x}\|^3$$

Для этого представим ее в следующем виде:

$$f(\mathbf{x}) = \frac{1}{3} \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{3}{2}}$$

Воспользуемся правилами дифференцирования композиции и квадратичной функции для нахождения градиента:

$$df = \frac{1}{2} \|\mathbf{x}\| d\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\| \langle \mathbf{x}, d\mathbf{x} \rangle$$

Приступим к поиску гессиана

$$d^2 f = d \|\mathbf{x}\| \langle \mathbf{x}, d\mathbf{x}_1 \rangle = d\|\mathbf{x}\| \langle \mathbf{x}, d\mathbf{x}_1 \rangle + \|\mathbf{x}\| \langle d\mathbf{x}_2, d\mathbf{x}_1 \rangle =$$

Первую норму дифференцируем аналогично через переход к скалярному квадрату:

$$= \frac{1}{\|\mathbf{x}\|} \langle \mathbf{x}, d\mathbf{x}_2 \rangle \langle \mathbf{x}, d\mathbf{x}_1 \rangle + \|\mathbf{x}\| \langle d\mathbf{x}_2, d\mathbf{x}_1 \rangle =$$

Преобразуем в каноническую форму, для этого рассмотрим подробнее первое слагаемое:

$$\langle \mathbf{x}, d\mathbf{x}_2 \rangle \langle \mathbf{x}, d\mathbf{x}_1 \rangle = \langle \mathbf{x}, d\mathbf{x}_1 \rangle \langle \mathbf{x}, d\mathbf{x}_2 \rangle = d\mathbf{x}_1^\top \mathbf{x} \mathbf{x}^\top d\mathbf{x}_2 = \langle \mathbf{x} \mathbf{x}^\top d\mathbf{x}_2, d\mathbf{x}_1 \rangle$$

Осталось собрать общий канонический вид

$$d^2 f = \left\langle \left[\frac{1}{\|\mathbf{x}\|} \mathbf{x} \mathbf{x}^\top + \|\mathbf{x}\| \mathbf{I} \right] d\mathbf{x}_2, d\mathbf{x}_1 \right\rangle$$

И получить итоговый ответ на задачу:

$$\nabla f = \|\mathbf{x}\| \mathbf{x}$$

$$\nabla^2 f = \frac{1}{\|\mathbf{x}\|} \mathbf{x} \mathbf{x}^\top + \|\mathbf{x}\| \mathbf{I}$$

4.4 Задача. Линейная регрессия. Продолжение

Найдем гессиан функции потерь линейной регрессии. Для этого вспомним выражение для градиента:

$$\nabla \mathcal{L}(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) \mathbf{w} - 2\mathbf{X}^\top \mathbf{y}$$

Тогда первая производная выглядит следующим образом:

$$d\mathcal{L}(\mathbf{w})[d\mathbf{w}_1] = \langle 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) \mathbf{w} - 2\mathbf{X}^\top \mathbf{y}, d\mathbf{w}_1 \rangle$$

Дифференцируем градиент:

$$d^2 \mathcal{L}(\mathbf{w})[d\mathbf{w}_1, d\mathbf{w}_2] = \langle d[2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) \mathbf{w} - 2\mathbf{X}^\top \mathbf{y}], d\mathbf{w}_1 \rangle = \langle 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) d\mathbf{w}_2, d\mathbf{w}_1 \rangle$$

Получаем выражение для гессиана:

$$\nabla^2 \mathcal{L}(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$$

Посмотрим на него повнимательней: в случае, если коэффициент регуляризации λ равен 0, гессиан пропорционален матрице $\mathbf{X}^\top \mathbf{X}$ – матрице ковариаций признаков (положим выборочное математическое ожидание признака равным 0).

Вспомним, что гессиан это оператор, характеризующий изменение градиента функции в точке. Если матрица ковариаций имеет существенное диагональное преобладание, что соответствует попарной статистической независимости признаков, то ее собственные значения будут слабо отличаться от диагональных элементов – дисперсий признаков. Иными словами, будут существовать такие направления, вдоль которых градиент будет масштабироваться в дисперсия признаков раз. Если дисперсии признаков существенно различаются (например, по порядку значений), то это приведет к плохой обусловленности задачи и как следствие замедлению сходимости градиентных методов оптимизации, поэтому стандартным подходом является нормировка признаков.

5 Еще немного задач

Поработаем еще немного с производными функций от матриц, а также решим задачу на исследование свойств гессиана. Похожие номера будут предложены Вам в домашнем задании.

5.1 Задача. Норма Фробениуса

Найдем градиент следующей функции:

$$f(\mathbf{X}) = \|\mathbf{AX} - \mathbf{B}\|_F$$

Несложно видеть, что данная функция есть композиция линейной функции и нормы Фробениуса. Для начала отдельно найдем производную последней. С похожими рассуждениями мы уже сталкивались, когда дифференцировали векторную норму. Норма Фробениуса – корень из скалярного квадрата в пространстве матриц.

$$d\|\mathbf{X}\|_F = d\langle \mathbf{X}, \mathbf{X} \rangle^{\frac{1}{2}} = \frac{1}{2} \frac{1}{\|\mathbf{X}\|_F} d\langle \mathbf{X}, \mathbf{X} \rangle = \frac{1}{\|\mathbf{X}\|_F} \langle \mathbf{X}, d\mathbf{X} \rangle$$

Используем правило производной композиции, дифференцируем линейную функцию:

$$\begin{aligned} df &= \frac{1}{\|\mathbf{AX} - \mathbf{B}\|_F} \langle \mathbf{AX} - \mathbf{B}, d(\mathbf{AX} - \mathbf{B}) \rangle = \frac{1}{\|\mathbf{AX} - \mathbf{B}\|_F} \langle \mathbf{AX} - \mathbf{B}, \mathbf{A} d\mathbf{X} \rangle = \\ &= \frac{1}{\|\mathbf{AX} - \mathbf{B}\|_F} \langle \mathbf{A}^\top (\mathbf{AX} - \mathbf{B}), d\mathbf{X} \rangle = \left\langle \frac{\mathbf{A}^\top (\mathbf{AX} - \mathbf{B})}{\|\mathbf{AX} - \mathbf{B}\|_F}, d\mathbf{X} \right\rangle \end{aligned}$$

Записываем ответ:

$$\nabla f(\mathbf{X}) = \frac{\mathbf{A}^\top (\mathbf{AX} - \mathbf{B})}{\|\mathbf{AX} - \mathbf{B}\|_F}$$

5.2 Задача. Знакоопределенность гессиана

Рассмотрим матричную функцию

$$f(\mathbf{X}) = \log \det \mathbf{X},$$

определенную на множестве симметричных матриц. Такую функцию мы уже дифференцировали в задаче на поиск ML-оценки параметров многомерного нормального распределения. Тогда мы получили, что

$$df = \langle \mathbf{X}^{-\top}, d\mathbf{X} \rangle = \langle \mathbf{X}^{-1}, d\mathbf{X} \rangle$$

Найдем вторую производную:

$$d^2 f = d\langle \mathbf{X}^{-1}, d\mathbf{X}_1 \rangle = \langle d\mathbf{X}^{-1}, d\mathbf{X}_1 \rangle = \langle -\mathbf{X}^{-1} d\mathbf{X}_2 \mathbf{X}^{-1}, d\mathbf{X}_1 \rangle$$

Теперь покажем, что если \mathbf{X} положительно определена, то выражение $d^2 f[\mathbf{H}, \mathbf{H}]$ имеет отрицательный знак:

$$d^2 f[\mathbf{H}, \mathbf{H}] = \langle -\mathbf{X}^{-1} \mathbf{H} \mathbf{X}^{-1}, \mathbf{H} \rangle =$$

В силу положительной определенности \mathbf{X} положительно определена и \mathbf{X}^{-1} и, кроме того, из нее можно извлечь положительно определенный корень $\mathbf{X}^{-\frac{1}{2}}$. Распределим этот множитель по аргументам скалярного произведения и завершим доказательство:

$$= \langle -\mathbf{X}^{-\frac{1}{2}} \mathbf{H} \mathbf{X}^{-\frac{1}{2}}, \mathbf{X}^{-\frac{1}{2}} \mathbf{H} \mathbf{X}^{-\frac{1}{2}} \rangle = -\|\mathbf{X}^{-\frac{1}{2}} \mathbf{H} \mathbf{X}^{-\frac{1}{2}}\|_F^2 < 0, \mathbf{H} \neq \mathbf{0}$$