Матрично-векторное дифференцирование

14 октября 2020 г.

Здесь и далее приняты обозначения: a — скаляр, $\mathbf{a}$ — вектор, A — матрица, $\langle \mathbf{x}, \mathbf{a} \rangle$ — скалярное произведение объектов.

Зачем нужно матричное дифференцирование? Почти все задачи машинного обучения ставятся как задачи оптимизации. А при оптимизации часто используют градиентный спуск или аппроксимацию целевой функции:

$$f(\mathbf{x} + d\mathbf{x}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), d\mathbf{x} \rangle + \frac{1}{2}\langle H d\mathbf{x}, d\mathbf{x} \rangle + \dots$$

Чаще всего используют аппроксимацию первого порядка, но иногда используют и второго (причем обычно не Гессиан в явном виде, а его продукты умножения $\langle H d\mathbf{x}, d\mathbf{x} \rangle$ или $H d\mathbf{x}$)

Векторные и матричные функции — удобное обощение скалярных функций и функций многих переменных.

1. Скалярная функция:

$$f(x) : \mathbb{R}^1 \to \mathbb{R}^1, \quad x \in \mathbb{R}^1$$

   Пример:

$$f(x) = x, \quad f'(x) = 1$$

2. Функция многих переменных:

$$f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^1, \quad \mathbf{x} \in \mathbb{R}^n$$

   Градиент функции — вектор частных производных:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

   Градиент часто обозначают как $\frac{\partial}{\partial \mathbf{x}}$.
   Пример:

$$f(\mathbf{x}) = x_1 + \dots + x_n, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}; \quad \nabla f(\mathbf{x}) = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

3. Векторная функция:

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}, \quad \mathbb{R}^n \to \mathbb{R}^m, \quad \mathbf{x} \in \mathbb{R}^n$$

Матрица Якоби — матрица частных производных:

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Обратите внимание, что i-ая строка отвечает за дифференцирование i-ой компоненты функции f_i, а j-ый столбец отвечает за дифференцирование по j-ой компоненте x_j. То есть в покомпонентном виде, эту матрицу можно записать следующим образом:

$$J_{ij} = \frac{\partial f_i}{\partial x_j}, \quad i = \overline{1, m}, \quad j = \overline{1, n}$$

Пример:

$$\mathbf{f}(\mathbf{x}) = \mathbf{x}; \quad J = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

4. Матричная функция скалярного аргумента:

$$F(x) = \begin{pmatrix} f_{11}(x) & \dots & f_{1m}(x) \\ \vdots & & \vdots \\ f_{n1}(x) & \dots & f_{nm}(x) \end{pmatrix}, \quad x \in \mathbb{R}^1, \quad F : \mathbb{R}^1 \to \mathbb{R}^{n \times m}$$

матрица частных производных $\frac{\partial}{\partial x} F(x)$:

$$\frac{\partial}{\partial x} F(x) = \begin{pmatrix} \frac{\partial f_{11}(x)}{\partial x} & \dots & \frac{\partial f_{1m}(x)}{\partial x} \\ \vdots & & \vdots \\ \frac{\partial f_{n1}(x)}{\partial x} & \dots & \frac{\partial f_{nm}(x)}{\partial x} \end{pmatrix}$$

Пример:

$$F(x) = \begin{pmatrix} x & \dots & x \\ \vdots & & \vdots \\ x & \dots & x \end{pmatrix}, \quad \frac{\partial}{\partial x} F(x) = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

Большинство правил преобразования производных обобщается и на векторный случай, в силу линейности операции дифференцирования.

В простых случаях можно действовать по определению.

Пример 1

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{a} \rangle, \quad \mathbf{x}, \mathbf{a} \in \mathbb{R}^n, \quad \mathbf{a} - const$$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1} \sum\limits_{i=1}^{n} x_i a_i \\ \vdots \\ \frac{\partial}{\partial x_n} \sum\limits_{i=1}^{n} x_i a_i \end{pmatrix} = \begin{pmatrix} a_1 + 0 + \cdots + 0 \\ \vdots \\ 0 + 0 + \cdots + a_n \end{pmatrix} = \mathbf{a}$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{x}} \langle \mathbf{x}, \mathbf{a} \rangle = \mathbf{a}$$

Пример 2

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

$$\langle A\mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\intercal A \mathbf{x} = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} \sum\limits_{i,j=1}^{n} a_{ij} x_i x_j \\ \vdots \\ \frac{\partial}{\partial x_n} \sum\limits_{i,j=1}^{n} a_{ij} x_i x_j \end{pmatrix} = \begin{pmatrix} \sum\limits_{j=1}^{n} a_{1j} x_j + \sum\limits_{i=1}^{n} a_{i1} x_i \\ \vdots \\ \sum\limits_{j=1}^{n} a_{nj} x_j + \sum\limits_{i=1}^{n} a_{in} x_i \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{n} (a_{1i} + a_{i1}) x_i \\ \vdots \\ \sum\limits_{i=1}^{n} (a_{ni} + a_{in}) x_i \end{pmatrix} = (A + A^\intercal) \cdot \mathbf{x}$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{x}} \langle A\mathbf{x}, \mathbf{x} \rangle = (A + A^\intercal) \cdot \mathbf{x}$$

Удобнее всего оказывается работать в терминах «дифференциала» – с ним можно не задумываться о промежуточных размерностях, а просто применять стандартные правила.

Правила преобразования

X, Y - произвольные дифференцируемые матричные функции, согласованные по размерностям.

1. $dA = 0$
2. $d(\alpha X) = \alpha(dX)$
3. $d(AXB) = A(dX)B$
4. $d(X + Y) = dX + dY$
5. $d(X^T) = (dX)^T$
6. $d(XY) = (dX)Y + X(dY)$
7. $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$

Стандартные дифференциалы

1. $d\langle A, X\rangle = \langle A, dX\rangle$
2. $\langle Ax, x\rangle = \langle (A + A^T)x, dx\rangle$
3. $d(\det(X)) = \det(X)\langle X^{-1}, dX\rangle$
4. $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

Чтобы из выражения для дифференциала получить градиент, нужно это выражение привести к канонической форме:

$$df(x) = \langle \nabla f(x), dx\rangle$$

Для разных функций она выглядит по разному:

- Для функции скалярнго аргумента: $df(x) = f'(x)dx$
- Для функции векторного аргумента: $df(\mathbf{x}) = \nabla f(\mathbf{x})^T d\mathbf{x}$
- Для функции матричного аргумента: $df(X) = \mathrm{tr}(X^T dX)$

Примеры для разбора

Пример 3

$$f(\mathbf{x}) = \langle A\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{x}, \mathbf{b}\rangle + c, \quad \frac{\partial}{\partial \mathbf{x}} - ?$$

$$df(\mathbf{x}) = \langle (A + A^T)\mathbf{x}, d\mathbf{x}\rangle - \langle \mathbf{b}, d\mathbf{x}\rangle + 0 = \langle (A + A^T)\mathbf{x} - \mathbf{b}, d\mathbf{x}\rangle$$

$$\Rightarrow \nabla f(\mathbf{x}) = (A + A^T)\mathbf{x} - \mathbf{b}$$

Пример 4

$$f(\mathbf{x}) = ||A\mathbf{x} - \mathbf{b}||_2^2, \quad \frac{\partial}{\partial \mathbf{x}} - ?$$

$$||A\mathbf{x}-\mathbf{b}||_2^2 = (A\mathbf{x}-\mathbf{b})^\intercal(A\mathbf{x}-\mathbf{b}) = (\mathbf{x}^\intercal A^\intercal A\mathbf{x}-\mathbf{x}^\intercal A^\intercal\mathbf{b}-\mathbf{b}^\intercal A\mathbf{x}+\mathbf{b}^\intercal\mathbf{b}) = \{\mathbf{b}^\intercal A\mathbf{x} = \mathbf{x}^\intercal A^\intercal\mathbf{b}\} =$$

$$= \mathbf{x}^\intercal A^\intercal A\mathbf{x} - 2\mathbf{x}^\intercal A^\intercal\mathbf{b} + \mathbf{b}^\intercal\mathbf{b} = \langle \mathbf{x}, A^T A\mathbf{x}\rangle - 2\langle \mathbf{x}, A^T\mathbf{b}\rangle + 0$$

$$\Rightarrow \nabla f(\mathbf{x}) = (A^T A + AA^T)\mathbf{x} - 2A^T\mathbf{b} = 2A^T A - 2A^T\mathbf{b}$$

Дифференциал сложной функции

Одним из самых важных является правило производной композиции. Пусть $g(Y)$ и $f(X)$ - две дифференцирумые функции, и мы знаем выражения для их дифференциалов: $dg(Y)$ и $df(X)$. Чтобы посчитать дифференциал композиции $g(f(X))$, как и в скалярном случае, нужно:

- взять выражение посчитанного дифференциала $dg(Y)$
- подставить в него вместо Y значение $f(X)$, а вместо dY - значение $df(X)$.

Пример 5

$$f(\mathbf{x}) = \log(1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}) = g\left(h(\mathbf{x})\right), \quad \frac{\partial}{\partial x} - ?$$

$$df(\mathbf{x}) = \frac{d(1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle})}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}} = \frac{e^{\langle \mathbf{a}, \mathbf{x} \rangle} d\langle \mathbf{a}, \mathbf{x} \rangle}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}} = \frac{e^{\langle \mathbf{a}, \mathbf{x} \rangle} \langle \mathbf{a}, d\mathbf{x} \rangle}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}} = \left\langle \frac{e^{\langle \mathbf{a}, \mathbf{x} \rangle}}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}} \cdot \mathbf{a}, d\mathbf{x} \right\rangle$$

$$\Rightarrow \nabla f(\mathbf{x}) = \frac{e^{\langle \mathbf{a}, \mathbf{x} \rangle}}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}} \cdot \mathbf{a} = \frac{\mathbf{a}}{1 + e^{-\langle \mathbf{a}, \mathbf{x} \rangle}}$$

Пример 6

$$f(X) = -\log\det(X) + \operatorname{Tr}(AX), \ \nabla f(X) - ?, \ \langle \nabla^2 f(X) S, S \rangle - ?$$

$$df(X) = -\frac{\det(X) \langle X^{-1}, dX \rangle}{\det(X)} + \langle A, dX \rangle = -\langle X^{-1}, dX \rangle + \langle A, dX \rangle = \langle A - X^{-1}, dX \rangle$$

$$\Rightarrow \nabla f(X) = A - X^{-1}$$

Выразим произведение гессиана на произвольную матрицу S. Чтобы выразить второй дифференциал, нужно принять первое приращение за константу (обозначим его dX_1) и проварьировать снова.

$$d^2 f(X) = -\langle d(X^{-1}), dX_1 \rangle = -\langle -X^{-1} dX_2 X^{-1}, dX_1 \rangle = \langle X^{-1} dX_2 X^{-1}, dX_1 \rangle$$

И, подставив вместо приращений произвольный вектор, можно получить искомое:

$$\langle \nabla^2 f(X) S, S \rangle = \langle X^{-1} S X^{-1}, S \rangle$$