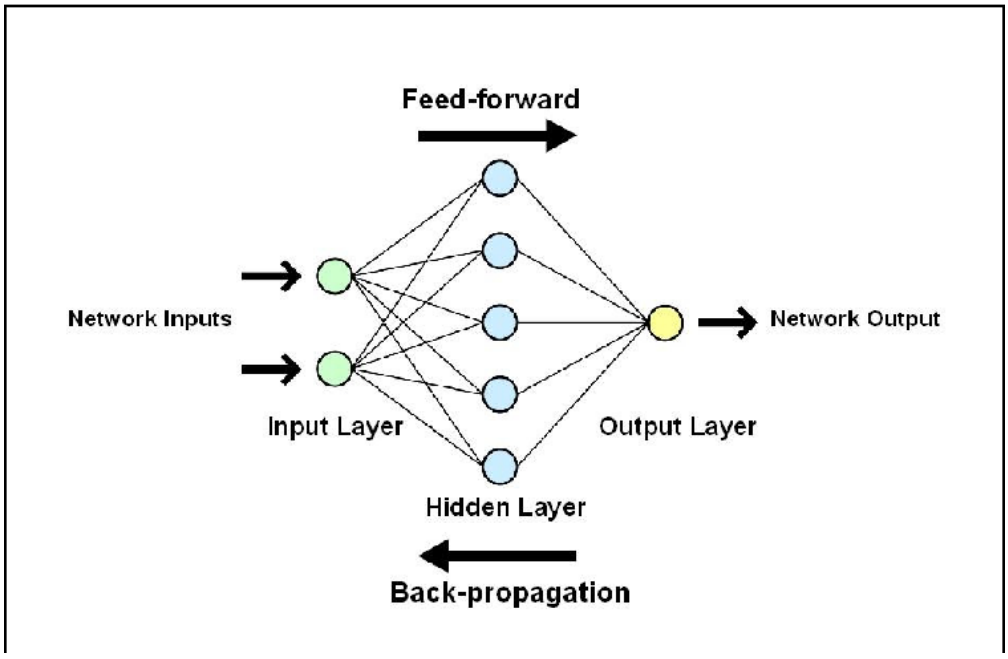
Факультет
математики
и компьютерных
наук
СПбГУ

Машинное обучение

Семинар 2. Матрично-векторное дифференцирование

9 сентября 2021

- на лекции мы осознаем важность градиентного спуска
- ещё умение считать производные (градиенты) пригодится для понимания и написания нейронных сетей (обратное распространение ошибки)



Определения

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

  $\nabla_x f(x) =

  ```
  \bigg[
      \frac{\partial f}{\partial x_1},
      \dots,
      \frac{\partial f}{\partial x_n}
  \bigg]^T
  ```

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

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

Задача 1

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

Решение

$\frac{\partial}{\partial x_i} a^T x = \frac{\partial}{\partial x_i} \sum\limits_j a_j x_j = a_i$, 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 2

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

Решение

$\frac{\partial}{\partial x_i} x^T A x = \frac{\partial}{\partial x_i} \sum\limits_j x_j (Ax)_j = \frac{\partial}{\partial x_i} \sum\limits_j x_j \left(\sum\limits_k a_{jk} x_k \right) =$

$= \frac{\partial}{\partial x_i} \sum\limits_{j,k} a_{jk} x_j x_k =$

$= \sum\limits_{j \neq i} a_{ji} x_j + \sum\limits_{k \neq i} a_{ik} x_k + 2a_{ii} x_i = \sum\limits_j a_{ji} x_j + \sum\limits_k a_{ik} x_k = \sum\limits_j (a_{ji} + a_{ij}) x_j$

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

Задача 3

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

Решение

Воспользуемся следствием теоремы Лапласа о разложении определителя по строке:

$\frac{\partial}{\partial a_{ij}} \det A = \frac{\partial}{\partial a_{ij}} \left[\sum\limits_k (-1)^{i+k} a_{ik} M_{ik} \right] = (-1)^{i+j} M_{ij},$ 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 4

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

Решение

$\frac{\partial}{\partial a_{ij}} \mathrm{tr}(AB) = \frac{\partial}{\partial a_{ij}} \sum\limits_k (AB)_{kk} = \frac{\partial}{\partial a_{ij}} \sum\limits_{k,l} a_{kl} b_{lk} = b_{ji}.$

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

Задача 5

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

Решение

Воспользовавшись выполняющимся для скаляра равенством $a = tr(a)$, циклическим свойством следа матрицы (для матриц подходящего размера):

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

и результатом предыдущей задачи, получаем

$$\nabla_A x^T A y = \nabla_A \mathrm{tr}(x^T A y) = \nabla_A \mathrm{tr}(A y x^T) = x y^T$$

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

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

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 6

Вычислите градиент логистической функции потерь для линейной модели по параметрам этой модели:
$\nabla_w \log(1 + \exp(-y \langle w, x \rangle))$

Решение

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

```
= \\
=
\frac{
    1
}{
    1
    +
    \exp(-y \langle w, x \rangle)
}
\nabla_w \left(
    1
    +
    \exp(-y \langle w, x \rangle)
\right)
=\\
=
\frac{
    1
}{
    1
    +
    \exp(-y \langle w, x \rangle)
}
\exp(-y \langle w, x \rangle)
\nabla_w \left(
    -y \langle w, x \rangle
\right)
=\\
=
```

Многомерная линейная регрессия

Задача найти $\hat{w} = \arg\min\limits_{w \in R^n} \frac{1}{l} \sum\limits_{i=1}^{l} (w^T x_i - y_i)^2$

Для решения два подхода:

1. Численно (градиентный спуск)
2. Аналитически

Перепишем минимизируемую функцию в матричном виде:

$$\sum_{i=1}^{l}(w^T x_i - y_i)^2 = (Xw - y)^T(Xw - y) =$$

$$w^T X^T Xw - y^T Xw - w^T X^T y + y^T y = w^T X^T Xw - 2y^T Xw + y^2$$

$(y^T Xw = w^T X^T y)$, т.к. транспонированный скаляр равен себе

Замечание (что нам нужно знать про градиент):

$\frac{\partial}{\partial w} w^T a = a$ (проверятся покоординатно)

$\frac{\partial}{\partial w} a^T w = a$ (проверятся покоординатно)

$\frac{\partial}{\partial w} w^T w = 2w$ (проверяется покоординатно)

$\frac{\partial}{\partial w} f(\vec{g}(w)) = \frac{\partial(\vec{g})}{\partial(w)}(\frac{\partial}{\partial w} f)(\vec{g}(w))$, где $\frac{\partial(\vec{g})}{\partial(w)} g(w)$ — матрица производных $\vec{g}(w)$,

т.е. $\begin{pmatrix} \frac{\partial g_1}{\partial w_1}(w) & \frac{\partial g_1}{\partial w_2}(w) & \cdots & \frac{\partial g_1}{\partial w_n}(w) \\ \frac{\partial g_2}{\partial w_1}(w) & \frac{\partial g_2}{\partial w_2}(w) & \cdots & \frac{\partial g_2}{\partial w_n}(w) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_n}{\partial w_1}(w) & \frac{\partial g_n}{\partial w_2}(w) & \cdots & \frac{\partial g_n}{\partial w_n}(w) \end{pmatrix}$ (проверяется покоординатно)

Посчитаем градиент:

$\frac{\partial}{\partial w} w^T X^T Xw - 2y^T Xw + y^2 = \frac{\partial}{\partial w}(Xw)^T Xw - \frac{\partial}{\partial w} 2y^T Xw = 2X^T Xw - 2X^T y$ (см. задачи 2 и 1)

Условие минимумма:

$(2X^T Xw - 2X^T y) = 0,$

значит:

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

Замечание.

Найденная точка — минимум, если матрица $X^T X$ обратима. Из курса математического анализа известно, что если матрица Гессе функции положительно определёна в точке, градиент которой равен нулю, то эта точка является локальным минимумом.

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

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.