

第4章 梯度下降的改进与超越

第3章介绍了梯度下降法。梯度下降法基于函数局部一阶特性。一阶近似是粗糙的，这种粗糙带来了一些问题。本章将介绍函数在局部的二阶特性。基于二阶特性分析函数在局部的性质。

本章首先回顾一些矩阵的相关知识，之后介绍如何在局部对函数进行二阶近似。有了函数的二阶近似就可以确定驻点的类型：极小点、极大点或者鞍点。之后本章介绍对原始梯度下降法的一些改进，这些改进有助于提高收敛速度，防止震荡或发散，规避局部极小。

最后，本章介绍两个基于函数二阶特性的优化算法：牛顿法和共轭方向法。然后介绍用牛顿法训练逻辑回归模型。二阶算法虽然不常用在神经网络和深度学习的训练中。阅读完本章，读者应该对函数的局部形态有更深刻的理解。

4.1 矩阵

首先回顾一些矩阵的基本知识。矩阵是实数构成的2维阵列。以一个 3×3 矩阵为例：

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = (\mathbf{a}_{*1} \quad \mathbf{a}_{*2} \quad \mathbf{a}_{*3}) = \begin{pmatrix} \mathbf{a}_{1*}^T \\ \mathbf{a}_{2*}^T \\ \mathbf{a}_{3*}^T \end{pmatrix} \quad (4.1)$$

式(4.1)囊括了本书用到的对矩阵的各种表示。本书用大写粗斜体字母表示矩阵，例如 \mathbf{A} 。 a_{ij} 是实数，是矩阵 \mathbf{A} 的第*i*行、第*j*列元素。 \mathbf{a}_{*j} 是矩阵的第*j*列，它是一个列向量：

$$\mathbf{a}_{*j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{pmatrix} \quad (4.2)$$

\mathbf{a}_{i*} 是矩阵的第*i*行，它是一个行向量：

$$\mathbf{a}_{i*} = \begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} \quad (4.3)$$

式(4.1)中对 \mathbf{a}_{i*} 进行了转置，以表示一行。矩阵的行数和列数不一定相等，可以是 $m \times n$ ， $m \neq n$ 。表示成：

$$\mathbf{A}_{m \times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad (4.4)$$

一般可省略下标 $m \times n$ 。两个相同形状的矩阵可以相加：

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix} \quad (4.5)$$

矩阵相加就是把相应元素相加。可以用实数（标量）乘一个矩阵：

$$k\mathbf{A} = \begin{pmatrix} ka_{11} & \cdots & ka_{1n} \\ \vdots & \ddots & \vdots \\ ka_{m1} & \cdots & ka_{mn} \end{pmatrix} \quad (4.6)$$

$-\mathbf{A}$ 就是 $(-1)\mathbf{A}$ 。显然有 $\mathbf{A} - \mathbf{A} = \mathbf{A} + (-\mathbf{A}) = \mathbf{O}$ 。 \mathbf{O} 是所有元素都为 0 的矩阵——零矩阵。矩阵 \mathbf{A} 的转置定义为：

$$\mathbf{A}^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix} = (\mathbf{a}_{1*} \quad \mathbf{a}_{2*} \quad \mathbf{a}_{3*}) = \begin{pmatrix} \mathbf{a}_{*1}^T \\ \mathbf{a}_{*2}^T \\ \mathbf{a}_{*3}^T \end{pmatrix} \quad (4.7)$$

\mathbf{A}^T 把 \mathbf{A} 的行当做列，列当做行。如果 \mathbf{A} 是 $m \times n$ 的，那么 \mathbf{A}^T 就是 $n \times m$ 的。

如果矩阵 \mathbf{A} 是 $m \times n$ 的，它可以与一个 n 维向量 \mathbf{x} 相乘：

$$\mathbf{Ax} = \sum_{j=1}^n x_j \mathbf{a}_{*j} \quad (4.8)$$

矩阵 \mathbf{A} 乘向量 \mathbf{x} ，使用 \mathbf{x} 的元素对矩阵的列进行线性组合。所以 \mathbf{A} 的列数和 \mathbf{x} 的维数必须相同。得到的结果是一个 m 维向量。容易看出 \mathbf{Ax} 的第 i 个元素是 $\sum_{j=1}^n x_j a_{ij} = \mathbf{a}_{i*}^T \mathbf{x}$ ，即 \mathbf{A} 的第 i 行与 \mathbf{x} 的内积。

有了矩阵和向量相乘的定义，就可以定义矩阵与矩阵相乘：

$$\mathbf{AB} = (\mathbf{Ab}_{*1} \quad \mathbf{Ab}_{*2} \quad \dots \quad \mathbf{Ab}_{*k}) \quad (4.9)$$

\mathbf{A} 与 \mathbf{B} 的乘积是矩阵 \mathbf{AB} 。 \mathbf{AB} 的第 j 列是 \mathbf{A} 与 \mathbf{B} 的第 j 列 \mathbf{b}_{*j} 的乘积。如果 \mathbf{A} 是 $m \times n$ 的，那么 \mathbf{b}_{*j} 必须是 n 维向量，即 \mathbf{B} 必须为 n 行。 \mathbf{B} 的列数任意，例如 k 。所以要能够与 $m \times n$ 的 \mathbf{A} 相乘， \mathbf{B} 的形状必须是 $n \times k$ ， k 任意。结果 \mathbf{AB} 的形状是 $m \times k$ 。 \mathbf{AB} 的第 i 行、第 j 列元素是：

$$\mathbf{a}_{i*}^T \mathbf{b}_{*j} = \sum_{s=1}^n a_{is} b_{sj} \quad (4.10)$$

仅从形状上看 \mathbf{B} 与 \mathbf{A} 不一定能够相乘，因为 k 不一定等于 m 。就算 $k = m$ ， \mathbf{BA} 也不一定等于 \mathbf{AB} 。即矩阵乘法不满足交换律。一个反例就可以证明这一点。这里不再赘述。

矩阵的乘法满足结合率：

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \quad (4.11)$$

矩阵的数乘满足分配率：

$$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}, \quad (k + h)\mathbf{A} = k\mathbf{A} + h\mathbf{A} \quad (4.12)$$

式（4.11）和（4.12）的证明很简单，只要检查一下矩阵元素的表达式。

向量 \mathbf{x} 的转置 \mathbf{x}^T 可以乘一个矩阵 \mathbf{A} ：

$$\mathbf{x}^T \mathbf{A} = (\mathbf{A}^T \mathbf{x})^T \quad (4.13)$$

行数和列数相同的矩阵是方阵。如果一个 $n \times n$ 的方阵的对角线元素为 1，其余元素都是 0，那么它是单位阵：

$$\mathbf{I}_{n \times n} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} \quad (4.14)$$

容易验证对于任何矩阵 $\mathbf{A}_{m \times n}$, $\mathbf{I}_{m \times m} \mathbf{A}_{m \times n} = \mathbf{A}_{m \times n} \mathbf{I}_{n \times n} = \mathbf{A}_{m \times n}$ 。在上下文很清晰时一般省略 \mathbf{I} 的下标。

如果 \mathbf{A} 是 $n \times n$ 方阵, 假如存在 $n \times n$ 方阵 \mathbf{A}^{-1} 满足:

$$\mathbf{A} \mathbf{A}^{-1} = \mathbf{A}^{-1} \mathbf{A} = \mathbf{I} \quad (4.15)$$

则称 \mathbf{A} 是可逆的, \mathbf{A}^{-1} 是 \mathbf{A} 的逆矩阵。 \mathbf{A} 的逆矩阵是唯一的。因为假如任何一个矩阵 \mathbf{B} 是 \mathbf{A} 的逆矩阵, 有:

$$\mathbf{B} = \mathbf{I} \mathbf{B} = \mathbf{A}^{-1} \mathbf{A} \mathbf{B} = \mathbf{A}^{-1} \quad (4.16)$$

方阵 \mathbf{A} 的对角线元素之和称为它的迹 (trace):

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} \quad (4.17)$$

方阵 \mathbf{A} 和 \mathbf{B} 的乘积 \mathbf{AB} 的迹等于 \mathbf{BA} 的迹:

$$\text{tr}(\mathbf{AB}) = \sum_{i=1}^n \mathbf{a}_{i*} \mathbf{b}_{*i} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ji} = \sum_{j=1}^n \sum_{i=1}^n b_{ji} a_{ij} = \sum_{j=1}^n \mathbf{b}_{j*} \mathbf{a}_{*j} = \text{tr}(\mathbf{BA}) \quad (4.18)$$

4.1.1 特征值与特征向量