第 4 章 超越梯度下降

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

首先回顾一下矩阵。这不是一个关于矩阵的全面介绍，例如行列式这个概念就没有出现。本节只介绍一下后文讨论中用得上的相关知识。

4.1.1 矩阵基础

矩阵是实数构成的 2 维阵列。以一个3×3矩阵为例：

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

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$\boldsymbol{A}$。a_{ij}是实数，是矩阵$\boldsymbol{A}$的第i行、第j列元素。$\boldsymbol{a}_{*j}$是矩阵的第j列，它是一个列向量：

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

$\boldsymbol{a}_{i*}$是矩阵的第i行，它是一个列向量：

$$\boldsymbol{a}_{i*} = \begin{pmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \end{pmatrix} \qquad (4.3)$$

式（4.1）中对$\boldsymbol{a}_{i*}$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：

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

一般可省略下标$m \times n$。两个相同形状的矩阵可以相加：

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

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：

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

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

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

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

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

$$\boldsymbol{Ax} = \sum_{j=1}^{n} x_j\, \boldsymbol{a}_{*j} \qquad (4.8)$$

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

$$\boldsymbol{AB} = (\boldsymbol{Ab}_{*1} \quad \boldsymbol{Ab}_{*2} \quad \cdots \quad \boldsymbol{Ab}_{*k}) \qquad (4.9)$$

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

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

仅从形状上看$\boldsymbol{B}$与$\boldsymbol{A}$不一定能够相乘，因为k不一定等于m。就算$k = m$，$\boldsymbol{BA}$也不一定等于$\boldsymbol{AB}$。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：

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \qquad (4.11)$$

矩阵的数乘满足分配率：

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

矩阵乘积的转置是：

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T \qquad (4.13)$$

式（4.11）、（4.12）和（4.13）的证明很简单，只需要检查一下矩阵元素的表达式。

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

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}} \qquad (4.14)$$

行数和列数相同的矩阵是方阵。方阵$\boldsymbol{A}$的对角线元素之和称为它的迹（trace）：

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \qquad (4.15)$$

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

$$tr(\boldsymbol{AB}) = \sum_{i=1}^{n} \boldsymbol{a}_{i*}\boldsymbol{b}_{*i} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \sum_{j=1}^{n} \boldsymbol{b}_{j*}\boldsymbol{a}_{*j} = tr(\boldsymbol{BA}) \qquad (4.16)$$

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：

$$\boldsymbol{I}_{n\times n} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \qquad (4.17)$$

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

4.1.2 矩阵的逆

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

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \qquad (4.18)$$

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

$$\boldsymbol{B} = \ \boldsymbol{IB} = \boldsymbol{A}^{-1}\boldsymbol{AB} = \boldsymbol{A}^{-1} \qquad (4.19)$$

如果方阵$\boldsymbol{A}$的逆矩阵是$\boldsymbol{A}^{\mathrm{T}}$，则称$\boldsymbol{A}$为正交矩阵：

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{I} \qquad (4.20)$$

从（4.20）可以看出$\boldsymbol{A}$的列$\boldsymbol{a}_{*j=1\dots n}$是单位向量且两两正交：

$$\begin{cases} {\boldsymbol{a}_{*i}}^{T}\boldsymbol{a}_{*j} = 0,\ i \neq j \\ {\boldsymbol{a}_{*i}}^{T}\boldsymbol{a}_{*j} = 1,\ i = j \end{cases} \qquad (4.21)$$

也就是说，正交矩阵$\boldsymbol{A}$的列都是单位向量，$\|\boldsymbol{a}_{*j=1\dots n}\| = 1$,。任意两列是正交的（夹角为$\pi/2$）。$\boldsymbol{A}$的列构成一组线性无关的向量，因为假如某列$\boldsymbol{a}_{*j}$可以被其他$\boldsymbol{a}_{*i\neq j}$线性表出：

$$\boldsymbol{a}_{*j} = \sum_{i\neq j} w_i \boldsymbol{a}_{*i} \qquad (4.22)$$

根据式（4.21），有：

$$ {\boldsymbol{a}_{*j}}^{\boldsymbol{T}}\boldsymbol{a}_{*j} = \sum_{i\neq j} w_i {\boldsymbol{a}_{*j}}^{\boldsymbol{T}}\boldsymbol{a}_{*i} = 0 \qquad (4.23)$$

这与式（4.21）矛盾。所以$\boldsymbol{a}_{*j=1\dots n}$是线性无关的。根据第 1 章的介绍，$\boldsymbol{A}$的列构成$n$维线性空间$\mathbb{R}^n$的一组基。因为$\boldsymbol{a}_{*j=1\dots n}$是两两正交的单位向量，所以它们被称为$\mathbb{R}^n$的一组标准正交基。之前提到的$\boldsymbol{e}_{i=1\dots n}$也是标准正交基，它们也是两两正交的单位向量。

4.1.3　特征值与特征向量

特征值和特征向量的概念不局限于方阵，但本书主要关注方阵。用方阵$\boldsymbol{A}$乘向量$\boldsymbol{x}$是在$\mathbb{R}^n$中进行一个变换，将$\boldsymbol{x}$变换成$\boldsymbol{A}\boldsymbol{x}$。例如矩阵：

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \qquad (4.24)$$

用$\boldsymbol{R}$乘向量$\boldsymbol{x}$等于将$\boldsymbol{x}$逆时针旋转θ度。这可以自行验证。