**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。以一个矩阵为例：

（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。是实数，是矩阵的第行、第列元素。是矩阵的第列，它是一个列向量：

（4.2）

是矩阵的第行，它是一个列向量：

（4.3）

式（4.1）中对进行了转置，以表示一行。矩阵的行数和列数不一定相等，可以是，。表示成：

（4.4）

一般可省略下标。两个相同形状的矩阵可以相加：

（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：

（4.6）

就是。显然有。是所有元素都为0的矩阵——零矩阵。矩阵的转置定义为：

（4.7）

把的行当做列，列当做行。如果是的，那么就是的。

如果矩阵是的，它可以与一个维向量相乘：

（4.8）

矩阵乘向量，使用的元素对矩阵的列进行线性组合。所以的列数和的维数必须相同。得到的结果是一个维向量。容易看出的第个元素是，即的第行与的内积。

有了矩阵和向量相乘的定义，就可以定义矩阵与矩阵相乘：

（4.9）

与的乘积是矩阵。的第列是与的第列的乘积。如果是的，那么必须是维向量，即必须为行。的列数任意，例如。所以要能够与的相乘，的形状必须是，任意。结果的形状是。的第行、第列元素是：

（4.10）

仅从形状上看与不一定能够相乘，因为不一定等于。就算，也不一定等于。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：

（4.11）

矩阵的数乘满足分配率：

（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。

向量的转置可以乘一个矩阵：

（4.13）

行数和列数相同的矩阵是方阵。如果一个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：

（4.14）

容易验证对于任何矩阵，。在上下文很清晰时一般省略的下标。

如果是方阵，假如存在方阵满足：

（4.15）

则称是可逆的，是的逆矩阵。的逆矩阵是唯一的。因为假如任何一个矩阵是的逆矩阵，有：

（4.16）

方阵的对角线元素之和称为它的迹（trace）：

（4.17）

方阵和的乘积的迹等于的迹**：**

（4.18）

**4.1.1 特征值与特征向量**