**第3章 逻辑回归模型的训练**

训练是给定训练集的前提下寻找最优模型参数的过程。借助损失函数，训练转化为在参数空间中寻找使损失函数最小的参数值的问题。这是函数优化问题。

本章首先回顾多元微积分基础。阐述描述函数局部特性的偏导数、方向导数、梯度、赫森矩阵以及泰勒展开等概念。在这些概念基础上介绍函数的驻点、局部极小点、鞍点和全局最小点。

梯度下降法是基于函数局部一阶特性——梯度的优化算法，它是神经网络和深度学习中最重要的训练算法。本文介绍梯度下降法的原理及其各种变体。

函数在某点局部的二阶特性体现在赫森矩阵中，本章介绍基于函数二阶特性的优化算法牛顿法和共轭方向法。

最后，本章将上述优化算法具体应用到逻辑回归模型的训练中。阅读完本章，读者应能理解逻辑回归、神经网络和深度学习的训练原理。

**3.1 多元微积分**

本节名为“多元微积分”，其实我们主要关注多元微分。它刻画了函数的局部特性。寻找函数的最小点就利用了这些局部特性。

**3.1.1 梯度**

回忆一下一元函数的可导性以及导数：

（3.1）

如果极限（3.1）存在，则在可导。是自变量的某一点。是一个小变化量，决定了自变量的另一点。用一条直线连接函数图像在和的两点，式（3.1）极限里的比例是这条直线的斜率。这条直线称为割线。随着趋近于0，割线趋近于在的切线。割线斜率的极限就是切线的斜率。如图3-1所示。

图3-1 一元函数的导数

也可以视作自变量从变化到时，的平均变化（速）率。是平均变化（速）率的极限，就是在的瞬时变化率。

在一元情况下，自变量只能沿着一个方向（前后）运动。用的瞬时变化率定义导数。如果是多元函数，自变量是向量，它可以向无数方向运动，此时不能类似式（3.1）那样定义的导数。

对于一元函数，构造一个对的仿射变换：

（3.2）

令，容易看出。根据式（3.1）有：

（3.3）

所以可以写成一个仿射变换加上余项：

（3.4）

其中有：

（3.5）

如果满足式（3.5），称是变化幅度的高阶无穷小。用语言描述就是当变化幅度趋近于0，即向靠近时，也随着消失（趋近于0）。且消失得比向靠近得更快。

反过来，如果在附近的微小变化可以写成一个仿射变换加余项：，其中是的一阶无穷小，那么：

（3.6）

式（3.6）的极限存在说明在可导。所以一元函数在等价于其在附近的变化量等价于存在仿射函数，该仿射函数与的差是的高阶无穷小。就是导数。函数在可导即在附近可有一个仿射函数（直线、平面，超平面）近似。近似的误差随着局部范围的缩小而消失，且消失得比局部范围的缩小更快。

这个可导的第二种定义可以扩展到多元函数。假设一个变化向量。变化的距离是的长度。如果作为的函数可以被一个仿射变换近似：

（3.7）

其中是的高阶无穷小：

（3.8）

式（3.7）中的是一个向量，就是多元函数在的梯度（gradient）。的近似仿射变换是：

（3.9）

梯度是该仿射变化等高线的法向量。也就是说，如果忽略近似误差，在附近认为图像就是仿射的图像。上述概念体现在图3-2中。

图3-2 多元函数的导——梯度

是函数在附近的一阶近似，它的特性就是在附近的一阶特性。如果自变量是n维，则的图像是n+1维空间中一张超平面。该超平面的法向量是n+1维向量，即给梯度添加一维常量-1。第1章曾经说过，仿射函数的全部特性体现在中：的方向决定超平面的朝向，决定超平面的倾角。所以的一阶特性都包含在梯度中。