**第3章 梯度下降法**

函数优化就是寻找使函数值最小的自变量。在模型训练的语境下，就是寻找使损失函数最小的模型参数值。梯度下降法是基于函数局部一阶特性的优化算法。它是神经网络和深度学习中最主要的训练算法。

本章首先回顾多元微积分基础。介绍多元函数梯度、方向导数、偏导数等概念。在某点附近函数可以由它在该点的切面近似。切面的朝向和倾斜程度信息蕴含在函数在该点的梯度之中。这些信息就是函数在该点局部的一阶信息。具备了多元微分的相关知识后，理解梯度下降算法就非常自然了。

由于梯度下降算法只利用函数局部的一阶特性，所以它是短视的。本章阐释这种短视所带来的种种问题。这些问题的规避和改进，将在下一章介绍函数二阶特性后加以说明。

最后，介绍运用梯度下降法训练逻辑回归模型。阅读完本章，读者应能透彻理解梯度下降法原理和局限。

**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及其导数：

（3.1）

如果极限（3.1）存在则在可导。是自变量空间的某一点。是一个变化量，决定了另一点。在的图像中用一条直线连接和两点，称为割线。式（3.1）极限里的商是割线的斜率。随着趋近于0，割线的极限是在的切线。割线斜率的极限是切线的斜率。如图3-1所示。

图3-1 一元函数的割线、切线和斜率

也可以视作自变量从变化到过程中的平均变化（速）率。是平均变化（速）率的极限——在的瞬时变化（速）率。

在一元情况下，自变量只能沿着一个方向（轴）前后运动。可以用瞬时变化（速）率定义导数。如果是多元函数，自变量是向量，它可以沿无数方向运动。这种情况下不能以类似式（3.1）那样定义的导数。

对一元函数，在点构造一个以为自变量的仿射变换：

（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。当向靠近，即趋近于0时，也随之消失（趋近于0）。且消失得比更快。

反过来，如果在附近的变化可以写成一个仿射变换加上余项：，其中是的高阶无穷小，那么：

（3.6）

式（3.6）的极限存在说明在可导。所以在可导等价于它在附近的值可以被一个仿射函数近似。该近似与的误差是的高阶无穷小。仿射函数的斜率就是。

可导的仿射近似定义可以扩展到多元函数。假设一个变化向量。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）

如果忽略近似误差，在附近可认为图像就是仿射的图像——超平面。如图3-2所示。

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局部一阶特性。如果自变量是n维，则的图像是n+1维空间中一张超平面，称为在的切平面。切平面面的法向量是n+1维向量，即给梯度添加一维常量-1。

第1章曾经介绍，仿射函数的全部特性体现在中：的方向决定超平面的朝向，决定超平面的倾斜程度。所以的局部一阶特性都包含在梯度中。

在的梯度是唯一的（如果在可导的话）。我们以几何方式证明这一点。如果在有两个不同的梯度向量和，则它们决定了两个切平面和。

和都经过点。当某一个自变量趋近于时，和之间的距离是一个以为高的三角形的底边。该距离与成固定比例。而与之间的距离是的高阶无穷小，所以与之间的距离不可能是的高阶无穷小，引出矛盾。所以在的梯度一定是唯一的。如图3-3所示。

图3-3 梯度唯一性的几何证明

**3.1.2 方向导数**

如何在多元情况下讨论在的瞬时变化率呢？在自变量空间中指定一条直线，然后讨论当自变量沿着这条直线运动时在的瞬时变化率。自变量空间中的直线可定义为：

（3.10）

式（3.10）定义了一条经过的直线。其中是单位向量。它的方向决定了直线的走向。是实数。决定了离的距离。将该直线看作自变量空间中一个坐标轴，以为原点，以的方向为正方向。的值是坐标轴上的坐标。

定义复合函数：

（3.11）

它以为自变量的一元函数。在0的导数是：

（3.12）

式（3.12）称为在沿的方向导数（directional derivative）。是在沿方向的瞬时变化率。如图3-4所示。

图3-4 方向导数

根据式（3.7）有：

（3.13）

其中是的高阶无穷小：

（3.14）

因为，所以当趋近于0时趋近于0。这时有：

（3.15）

由式（3.15）可知：是的高阶无穷小。式（3.13）表明在0的导数，即等于：

（3.16）

其中是与之间的夹角。

由式（3.16）可知：方向导数等于梯度向的投影长度。当与同向，即时最大。也就是说是变化率最大的方向，其变化率是。相反，沿着与相反的方向，即时最小，为。是变化率最小，即下降最快的方向

在2维的情况下可以用切平面阐述梯度与方向导数的关系。的法向量是。第3维-1说明该指向平面的下方。在平面的投影是，它指向切平面的上坡方向，指向切平面的下坡方向。沿着任意方向的运动可分以解成沿的分量和垂直于的分量。垂直于的方向上，所以的变化率就打了折扣，折扣系数正是沿的分量所占的“份额”——。如图3-5所示。

图3-5 梯度与方向导数

**3.1.3 偏导数**

在点对其第分量的偏导数是把其他分量当作常数时对的导数。这时候将看作关于的一元函数。根据导数的定义：

（3.17）

其中是第个标准基向量。保持不变，只有发生变化，变化量是。有n个偏导数：。

根据式（3.12），有：

（3.18）

对的偏导数就是沿的方向导数。偏导数是方向导数的特例，它们的方向是各个坐标轴正方向。

与的内积是的第分量。根据式（3.16），有：

（3.19）

所以梯度的第分量是对的偏导数。于是就有了梯度的计算式：

（3.20）

偏导数是唯一的，所以也是唯一的。

3.1.4 **驻点**

函数的驻点（stationary point）是梯度为零向量的点。在驻点的切平面的法向量是：

（3.21）

法向量垂直指向下方，即切平面是水平的。如图3-6所示。

图3-6 驻点的切平面

在驻点沿任意方向的方向导数是，所以在驻点向任意方向的变化率都为0。

**3.1.5 局部极小点**

如果是的局部极小点（local minima），则在周围存在一个半径为的邻域，该邻域所有点的函数值都不小于。用公式表示就是：

（3.22）

如果自变量空间中所有都有，则是的全局最小点（global minima）。很显然，全局最小点是局部极小点。但是局部极小点不一定是全局最小点。类似还可以定义局部极大点（local minima）和全局最大点（global maxima）。如图3-7所示。

图3-7 局部极小点和全局最小点

局部极小点一定是驻点。假如是的局部极小点，但是。从点出发沿方向产生一个位移， 是正实数。根据在的可导性，有：

（3.23）

是的高阶无穷小。于是：

（3.24）

这说明式（3.23）等号右边的后两项当趋近于0时的极限是负值。所以对于足够小的，当时有：

（3.25）

随着趋近于0，无限靠近的同时保持。这与是的局部极小点矛盾。所以一定是零向量，即是驻点。类似可以证明，局部极大点也一定是驻点。

驻点是局部极小点的必要非充分条件。驻点也有可能是局部极大点，还有可能是鞍点（saddle point）。鞍点的梯度为零向量，但在任意一个领域内都同时存在函数值更大和更小的点。如图3-7所示。

图3-7 鞍点

仅靠函数一阶特性难以判断驻点的类型。第4章介绍赫森矩阵后会知道：驻点的类型由赫森矩阵特征值的符号决定。

**3.2 梯度下降法**