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$f(x)$的可导性及其导数$f'(x)$：

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \qquad (3.1)$$

如果极限（3.1）存在则$f(x)$在x可导。x是自变量空间的某一点。h是一个变化量，决定了另一点$x+h$。在$f(x)$的图像中用一条直线连接$\big(x, f(x)\big)$和$\big(x+h, f(x+h)\big)$两点，称为割线。式（3.1）极限里的商$\big(f(x+h) - f(x)\big)/h$是割线的斜率。随着$h$趋近于 0，割线的极限是$f(x)$在$x$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

$\big(f(x+h) - f(x)\big)/h$也可以视作自变量从$x$变化到$x+h$过程中$f(x)$的平均变化（速）率。$f'(x)$是平均变化（速）率的极限——$f(x)$在$x$的瞬时变化（速）率。

在一元情况下，自变量只能沿着一个方向（x轴）前后运动。可以用瞬时变化（速）率定义导数。如果$f(\boldsymbol{x})$是多元函数，自变量$\boldsymbol{x}$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$f(\boldsymbol{x})$的导数。

对一元函数$f(x)$，在x点构造一个以h为自变量的仿射变换：

$$g(h) = f(x) + hf'(x) \qquad (3.2)$$

令$\mathcal{R}(h) = f(x+h) - g(h)$，容易看出$\mathcal{R}(0) = 0$。根据式（3.1）有：

$$\lim_{h\to 0}\left|\frac{\mathcal{R}(h)}{h}\right| = \lim_{h\to 0}\left|\frac{f(x+h)-f(x)}{h} - f'(x)\right| = 0 \qquad (3.3)$$

所以$f(x+h)$可以写成一个仿射变换加上余项：

$$f(x+h) = g(h) + \mathcal{R}(h) = f(x) + hf'(x) + \mathcal{R}(h) \qquad (3.4)$$

其中有：

$$\lim_{h\to 0}\frac{\mathcal{R}(h)}{|h|} = 0 \qquad (3.5)$$

如果$\mathcal{R}(h)$满足式（3.5），称$\mathcal{R}(h)$是变化幅度$|h|$的高阶无穷小。当$x+h$向x靠近，即$|h|$趋近于 0 时，$\mathcal{R}(h)$也随之消失（趋近于 0）。且$\mathcal{R}(h)$消失得比$|h|$更快。

反过来，如果$f(x)$在x附近的变化$f(x+h)$可以写成一个仿射变换加上余项：$f(x+h) = f(x) + ha + \mathcal{R}(h)$，其中$\mathcal{R}(h)$是$|h|$的高阶无穷小，那么：

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{ha+\mathcal{R}(h)}{h} = a + \lim_{h\to 0}\frac{\mathcal{R}(h)}{h} = a \qquad (3.6)$$

式（3.6）的极限存在说明$f(x)$在x可导。所以$f(x)$在x可导等价于它在x附近的值$f(x+h)$可以被一个仿射函数$f(x)+ha$近似。该近似与$f(x+h)$的误差是$|h|$的高阶无穷小。仿射函数

的斜率a就是$f'(x)$。

可导的仿射近似定义可以扩展到多元函数$f(\boldsymbol{x})$。假设一个变化向量$\boldsymbol{h}$。如果$f(\boldsymbol{x}+\boldsymbol{h})$作为$\boldsymbol{h}$的函数可以被一个仿射变换近似：

$$f(\boldsymbol{x}+\boldsymbol{h})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}+\mathcal{R}(\boldsymbol{h}) \qquad (3.7)$$

其中$\mathcal{R}(\boldsymbol{h})$是$\|\boldsymbol{h}\|$的高阶无穷小：

$$\lim_{\|\boldsymbol{h}\|\to 0}\frac{\mathcal{R}(\boldsymbol{h})}{\|\boldsymbol{h}\|}=0 \qquad (3.8)$$

式（3.7）中的$\nabla f(\boldsymbol{x})$是一个向量，就是多元函数$f(\boldsymbol{x})$在$\boldsymbol{x}$的梯度（gradient）。$f(\boldsymbol{x}+\boldsymbol{h})$的近似仿射变换是：

$$g(\boldsymbol{h})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h} \qquad (3.9)$$

如果忽略近似误差$\mathcal{R}(\boldsymbol{h})$，在$\boldsymbol{x}$附近可认为$f(\boldsymbol{x}+\boldsymbol{h})$图像就是仿射$g(\boldsymbol{h})$的图像——超平面。如图 3-2 所示。

图 3-2　多元函数的导——梯度

$g(\boldsymbol{h})$是函数$f(\boldsymbol{x})$在$\boldsymbol{x}$附近的一阶近似。它的特性就是$f(\boldsymbol{x})$在$\boldsymbol{x}$附近的局部一阶特性。如果自变量$\boldsymbol{x}$是 n 维，则$g(\boldsymbol{h})$的图像是 n+1 维空间中一张超平面，称为$f(\boldsymbol{x})$在$\boldsymbol{x}$的切平面。切平面面的法向量是 n+1 维向量$(\nabla f(\boldsymbol{x})^{\mathrm{T}},-1)^{T}$，即给梯度$\nabla f(\boldsymbol{x})$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$\nabla f(\boldsymbol{x})$中：$\nabla f(\boldsymbol{x})$的方向决定超平面$g(\boldsymbol{h})$的朝向，$\|\nabla f(\boldsymbol{x})\|$决定超平面$g(\boldsymbol{h})$的倾斜程度。所以$f(\boldsymbol{x})$的局部一阶特性都包含在梯度$\nabla f(\boldsymbol{x})$中。

$f(\boldsymbol{x})$在$\boldsymbol{x}$的梯度是唯一的（如果$f(\boldsymbol{x})$在$\boldsymbol{x}$可导的话）。我们以几何方式证明这一点。如果$f(\boldsymbol{x})$在$\boldsymbol{x}$有两个不同的梯度向量$\nabla f(\boldsymbol{x})$和$\nabla f(\boldsymbol{x})'$，则它们决定了两个切平面$g$和$g'$。
g和g'都经过点$\big(\boldsymbol{x},f(\boldsymbol{x})\big)$。当某一个自变量$\boldsymbol{x}'$趋近于$\boldsymbol{x}$时，$g$和$g'$之间的距离是一个以$\boldsymbol{x}'-\boldsymbol{x}$为高的三角形的底边。该距离与$\|\boldsymbol{x}'-\boldsymbol{x}\|$成固定比例。而$f$与$g$之间的距离是$\|\boldsymbol{x}'-\boldsymbol{x}\|$的高阶无穷小，所以$f$与$g'$之间的距离不可能是$\|\boldsymbol{x}'-\boldsymbol{x}\|$的高阶无穷小，引出矛盾。所以$f(\boldsymbol{x})$在$\boldsymbol{x}$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$f(\boldsymbol{x})$在$\boldsymbol{x}$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$f(\boldsymbol{x})$在$\boldsymbol{x}$的瞬时变化率。自变量空间中的直线可定义为：

$$l(t) = \boldsymbol{x} + t\boldsymbol{d} \quad t \in \mathbb{R},\ \ \|\boldsymbol{d}\| = 1 \qquad (3.10)$$

式（3.10）定义了一条经过$\boldsymbol{x}$的直线。其中$\boldsymbol{d}$是单位向量。它的方向决定了直线的走向。t是实数。$|t|$决定了$\boldsymbol{x} + t\boldsymbol{d}$离$\boldsymbol{x}$的距离。将该直线看作自变量空间中一个坐标轴$l$，以$\boldsymbol{x}$为原点，以$\boldsymbol{d}$的方向为正方向。$t$的值是坐标轴$l$上的坐标。

定义复合函数$(f \oplus l)(t)$：

$$(f \oplus l)(t) = f\big(l(t)\big) = f(\boldsymbol{x} + t\boldsymbol{d}) \qquad (3.11)$$

它以t为自变量的一元函数。$(f \oplus l)(t)$在 0 的导数是：

$$\nabla_{\boldsymbol{d}} f(\boldsymbol{x}) = \frac{d(f \oplus l)}{dt}(0) = \lim_{h \to 0} \frac{f\big(l(\mathrm{h})\big) - f\big(l(0)\big)}{h} = \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{d}) - f(\boldsymbol{x})}{h} \qquad (3.12)$$

式（3.12）称为$f(\boldsymbol{x})$在$\boldsymbol{x}$沿$\boldsymbol{d}$的方向导数（directional derivative）。$\nabla_{\boldsymbol{d}} f(\boldsymbol{x})$是$f(\boldsymbol{x})$在$\boldsymbol{x}$沿方向$\boldsymbol{d}$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：

$$(f \oplus l)(\mathrm{h}) = f(\boldsymbol{x} + \mathrm{h}\boldsymbol{d}) = (f \oplus l)(0) + h\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d} + \mathcal{R}(h\boldsymbol{d}) \qquad (3.13)$$

其中$\mathcal{R}(h\boldsymbol{d})$是$\|h\boldsymbol{d}\|$的高阶无穷小：

$$\lim_{\|h\boldsymbol{d}\|\to 0}\frac{\mathcal{R}(h\boldsymbol{d})}{\|h\boldsymbol{d}\|}=0 \qquad (3.14)$$

因为$\|h\boldsymbol{d}\|=|h|\|\boldsymbol{d}\|=|h|$，所以当$h$趋近于 0 时$\|h\boldsymbol{d}\|$趋近于 0。这时有：

$$\lim_{h\to 0}\frac{\mathcal{R}(h\boldsymbol{d})}{h}=\lim_{h\to 0}\frac{\mathcal{R}(h\boldsymbol{d})}{\|h\boldsymbol{d}\|}\frac{\|h\boldsymbol{d}\|}{h}=\lim_{h\to 0}\pm\frac{\mathcal{R}(h\boldsymbol{d})}{\|h\boldsymbol{d}\|}=\pm\lim_{\|h\boldsymbol{d}\|\to 0}\frac{\mathcal{R}(h\boldsymbol{d})}{\|h\boldsymbol{d}\|}=0 \qquad (3.15)$$

由式（3.15）可知：$\mathcal{R}(\mathrm{h}\boldsymbol{d})$是$h$的高阶无穷小。式（3.13）表明$(f\oplus l)(t)$在 0 的导数，即$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$等于$\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$：

$$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})=\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}=\|\nabla f(\boldsymbol{x})\|\cdot\|\boldsymbol{d}\|\cdot\boldsymbol{\cos}\,\theta=\|\nabla f(\boldsymbol{x})\|\,\boldsymbol{\cos}\,\theta \qquad (3.16)$$

其中θ是$\nabla f(\boldsymbol{x})$与$\boldsymbol{d}$之间的夹角。

由式(3.16)可知：方向导数$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$等于梯度$\nabla f(\boldsymbol{x})$向$\boldsymbol{d}$的投影长度。当$\nabla f(\boldsymbol{x})$与$\boldsymbol{d}$同向，即$\theta=0$时$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$最大。也就是说$\nabla f(\boldsymbol{x})$是$f(\boldsymbol{x})$变化率最大的方向，其变化率是$\|\nabla f(\boldsymbol{x})\|\geq 0$。相反，沿着与$\nabla f(\boldsymbol{x})$相反的方向，即$\theta=\pi$时$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$最小，为$-\|\nabla f(\boldsymbol{x})\|\leq 0$。$-\nabla f(\boldsymbol{x})$是$f(\boldsymbol{x})$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g阐述梯度$\nabla f(\boldsymbol{x})$与方向导数$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$的关系。$g$的法向量是$\boldsymbol{w}=(\nabla f(\boldsymbol{x})^{\mathrm{T}},-1)^{T}$。第 3 维-1 说明该$\boldsymbol{w}$指向$x_1x_2$平面的下方。$\boldsymbol{w}$在$x_1x_2$平面的投影是$\nabla f(\boldsymbol{x})$，它指向切平面$g$的上坡方向，$-\nabla f(\boldsymbol{x})$指向切平面$g$的下坡方向。沿着任意方向的运动可分以解成沿$\nabla f(\boldsymbol{x})$的分量和垂直于$\nabla f(\boldsymbol{x})$的分量。垂直于$\nabla f(\boldsymbol{x})$的方向上$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})=0$，所以$f(\boldsymbol{x})$的变化率就打了折扣，折扣系数正是沿$\nabla f(\boldsymbol{x})$的分量所占的“份额”——$\boldsymbol{\cos}\,\theta$。如图 3-5 所示。

图 3-5　梯度与方向导数

3.1.3　偏导数

$f(\boldsymbol{x})$在$\boldsymbol{x}$点对其第i分量x_i的偏导数是把其他分量$x_{j\neq i}$当作常数时$f(\boldsymbol{x})$对x_i的导数。这时候将$f(\boldsymbol{x})$看作关于x_i的一元函数。根据导数的定义：

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x})=\lim_{h\to 0}\frac{f(x+he_i)-f(x)}{h} \qquad (3.17)$$

其中$\boldsymbol{e}_i$是第i个标准基向量。$\boldsymbol{x}+h\boldsymbol{e}_i$保持$x_{j\neq i}$不变，只有$x_i$发生变化，变化量是$h$。$f(\boldsymbol{x})$有 n 个偏导数：$\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2},\dots,\frac{\partial f}{\partial x_n}$。

根据式（3.12），有：

$$\nabla_{\boldsymbol{e}_i} f(\boldsymbol{x})=\frac{\partial f}{\partial x_i}(\boldsymbol{x})=\lim_{h\to 0}\frac{f(x+he_i)-f(x)}{h} \qquad (3.18)$$

$f(\boldsymbol{x})$对x_i的偏导数就是$f(\boldsymbol{x})$沿$\boldsymbol{e}_i$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。

$\nabla f(\boldsymbol{x})$与$\boldsymbol{e}_i$的内积是$\nabla f(\boldsymbol{x})$的第i分量。根据式（3.16），有：

$$\nabla f(\boldsymbol{x})_i=\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{e}_i=\frac{\partial f}{\partial x_i}(\boldsymbol{x}),\quad i=1\dots\mathrm{n} \qquad (3.19)$$

所以梯度$\nabla f(\boldsymbol{x})$的第i分量是$f(\boldsymbol{x})$对x_i的偏导数。于是就有了梯度的计算式：

$$\nabla f(\boldsymbol{x})=\begin{pmatrix}\frac{\partial f}{\partial x_1}(\boldsymbol{x})\\ \frac{\partial f}{\partial x_2}(\boldsymbol{x})\\ \vdots\\ \frac{\partial f}{\partial x_n}(\boldsymbol{x})\end{pmatrix} \qquad (3.20)$$

偏导数是唯一的，所以$\nabla f(\boldsymbol{x})$也是唯一的。

3.1.4 驻点

函数$f(\boldsymbol{x})$的驻点（stationary point）是梯度为零向量的点。$f(\boldsymbol{x})$在驻点的切平面的法向量是：

$$\boldsymbol{w}=\begin{pmatrix}\nabla f(\boldsymbol{x})\\-1\end{pmatrix}=\begin{pmatrix}0\\0\\\vdots\\0\\-1\end{pmatrix}\qquad(3.21)$$

法向量$\boldsymbol{w}$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

$f(\boldsymbol{x})$在驻点沿任意方向$\boldsymbol{d}$的方向导数是$\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}=0$，所以$f(\boldsymbol{x})$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$\boldsymbol{x}^*$是$f(\boldsymbol{x})$的局部极小点（local minima），则在$\boldsymbol{x}^*$周围存在一个半径为$\varepsilon>0$的邻域，该邻域所有点的函数值都不小于$f(\boldsymbol{x}^*)$。用公式表示就是：

$$f(\boldsymbol{x})\geq f(\boldsymbol{x}^*)\quad\|\boldsymbol{x}-\boldsymbol{x}^*\|<\varepsilon\qquad(3.22)$$

如果自变量空间中所有$\boldsymbol{x}$都有$f(\boldsymbol{x})\geq f(\boldsymbol{x}^*)$，则$\boldsymbol{x}^*$是$f(\boldsymbol{x})$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$\boldsymbol{x}^*$是$f(\boldsymbol{x})$的局部极小点，但是$\nabla f(\boldsymbol{x}^*)\neq\boldsymbol{o}$。从$\boldsymbol{x}^*$点出发沿$-\nabla f(\boldsymbol{x}^*)$方向产生一个位移$-t\nabla f(\boldsymbol{x}^*)$，$t$ 是正实数。根据$f(\boldsymbol{x})$在$\boldsymbol{x}^*$的可导性，有：

$$f\big(\boldsymbol{x}^*-t\nabla f(\boldsymbol{x}^*)\big)=f(\boldsymbol{x}^*)-t\|\nabla f(\boldsymbol{x}^*)\|^2+\mathcal{R}\big(-t\nabla f(\boldsymbol{x}^*)\big)\qquad(3.23)$$

$\mathcal{R}\big(-t\nabla f(\boldsymbol{x}^*)\big)$是$\|-t\nabla f(\boldsymbol{x}^*)\|$的高阶无穷小。于是：

$$\lim_{t\to0}\frac{-t\|\nabla f(\boldsymbol{x}^*)\|^2+\mathcal{R}\big(-t\nabla f(\boldsymbol{x}^*)\big)}{\|-t\nabla f(\boldsymbol{x}^*)\|}=-\|\nabla f(\boldsymbol{x}^*)\|+\lim_{t\to0}\frac{\mathcal{R}\big(-t\nabla f(\boldsymbol{x}^*)\big)}{\|-t\nabla f(\boldsymbol{x}^*)\|}=-\|\nabla f(\boldsymbol{x}^*)\|<0\qquad(3.24)$$

这说明式（3.23）等号右边的后两项当t趋近于 0 时的极限是负值。所以对于足够小的$\varepsilon > 0$，当$t < \varepsilon$时有：

$$f\left(\boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*)\right) - f(\boldsymbol{x}^*) = -t\|\nabla f(\boldsymbol{x}^*)\|^2 + \mathcal{R}\left(-t\nabla f(\boldsymbol{x}^*)\right) < 0 \tag{3.25}$$

随着t趋近于 0，$\boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*)$无限靠近$\boldsymbol{x}^*$的同时保持$f\left(\boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*)\right) < f(\boldsymbol{x}^*)$。这与$\boldsymbol{x}^*$是$f(\boldsymbol{x})$的局部极小点矛盾。所以$\nabla f(\boldsymbol{x}^*)$一定是零向量，即$\boldsymbol{x}^*$是驻点。类似可以证明，局部极大点$\boldsymbol{x}^*$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