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是自变量的一个变化量。在$f(x)$的图像中用一个线段连接$(x \quad f(x))^{\mathrm{T}}$和$(x+h \quad f(x+h))^{\mathrm{T}}$两点。这个线段称为割线。式（3.1）极限里的商$\big(f(x+h)-f(x)\big)/h$是割线的斜率。随着变化量$h$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

$\big(f(x+h)-f(x)\big)/h$也是自变量从x变化到$x+h$过程中$f(x)$的平均变化率。$f'(x)$是平均变化率的极限——$f(x)$在x的瞬时变化率。

在一元情况下，自变量只能沿着x轴前后运动。可以用瞬时变化率定义导数。在多元情况下，自变量$\boldsymbol{x}$是向量，它可以沿无数方向运动。这时就不能以类似式（3.1）那样定义$f(\boldsymbol{x})$的导数。一元函数的导数还有一种等价定义，即在x附近用直线近似表示$f(x)$。这种定义可以扩展到多维的情况。现在考察这种定义。

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$\mathcal{R}(h) = f(x+h) - g(h)$，称为余项。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$f(x+h)$作为h的函数，可以写成一个仿射变换加余项$\mathcal{R}(h)$：

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

根据式（3.3），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$|h|$的高阶无穷小。当$|h|$趋近于 0，也就是当$x+h$向x靠近时，$\mathcal{R}(h)$也趋近于 0（消失）。且$\mathcal{R}(h)$消失得比$|h|$还快。

仿射变换$g(h)$的图像经过点$(0 \quad f(x))^{\mathrm{T}}$，是一条截距为$f(x)$，斜率为$f'(x)$的直线。如果将$g(h)$的图像平移，使原来的$(0 \quad f(x))^{\mathrm{T}}$点移动到$(x \quad f(x))^{\mathrm{T}}$点，平移后的图像就是过$(x \quad f(x))^{\mathrm{T}}$点的斜率为$f'(x)$的直线——$f(x)$在$x$的切线。如图 3-2 所示。

图 3-2　可导函数的仿射近似——切线

反过来，如果$f(x)$在x附近的变化$f(x+h)$可以写成：$f(x+h) = f(x) + ha + \mathcal{R}(h)$，$\mathcal{R}(h)$是

$|h|$的高阶无穷小，那么有：

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

式（3.6）的极限存在说明$f(x)$在x可导，导数是$f'(x)=a$。

综上所述，$f(x)$在x可导等价于$f(x+h)$作为h的函数可以被一个仿射函数$f(x)+ha$近似表示。并且$f(x+h)$与$f(x)+ha$的误差是$|h|$的高阶无穷小。该仿射函数的斜率a就是$f'(x)$。

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：

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

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

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

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

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

$g(\boldsymbol{h})$称为$f(\boldsymbol{x})$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 维空间中一个超平面。该超平面经过点$(\boldsymbol{o}^{\mathrm{T}}\quad f(\boldsymbol{x}))^{\mathrm{T}}$。将图像平移，使得$(\boldsymbol{o}^{\mathrm{T}}\quad f(\boldsymbol{x}))^{\mathrm{T}}$移动到$(\boldsymbol{x}^{\mathrm{T}}\quad f(\boldsymbol{x}))^{\mathrm{T}}$，平移后的超平面称为$f(\boldsymbol{x})$在$\boldsymbol{x}$的切平面。根据第 1 章的介绍，切平面的法向量是 n+1 维向量$(\nabla f(\boldsymbol{x})^{\mathrm{T}}\quad -1)^{\mathrm{T}}$，即给梯度$\nabla f(\boldsymbol{x})$添加一维常量-1。

仿射函数$g(\boldsymbol{h})$的全部特性体现在$\nabla f(\boldsymbol{x})$中：$\nabla f(\boldsymbol{x})$的方向决定超平面的朝向，$\|\nabla f(\boldsymbol{x})\|$的大小决定超平面的倾斜程度。所以$f(\boldsymbol{x})$的局部一阶特性都包含在梯度$\nabla f(\boldsymbol{x})$中。如图 3-3 所示。

图 3-3　多元函数的切平面

3.1.2　方向导数

如果$f(\boldsymbol{x})$在$\boldsymbol{x}$可导，如何讨论$f(\boldsymbol{x})$在$\boldsymbol{x}$的瞬时变化率呢？指定一条经过$\boldsymbol{x}$的直线，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。经过$\boldsymbol{x}$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$\boldsymbol{x}$点的直线。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

因为$f(\boldsymbol{x})$在$\boldsymbol{x}$可导，根据式（3.7）有：

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

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$\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$：

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

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，即函数值下降最快的方向。

在 2 维的情况下可以用切平面阐述梯度$\nabla f(\boldsymbol{x})$与方向导数$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$的关系。切平面的法向量是$\boldsymbol{w}=(\nabla f(\boldsymbol{x})^{\mathrm{T}}\quad -1)^{\mathrm{T}}$。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$\nabla f(\boldsymbol{x})$，它指向切平面的上坡方向，$-\nabla f(\boldsymbol{x})$指向切平面的下坡方向。沿着任意方向的运动可分解成沿$\nabla f(\boldsymbol{x})$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”——$\cos\theta$。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：

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

$f(\boldsymbol{x})$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：

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

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：

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

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：

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

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$\|\nabla f(\boldsymbol{x}^*)\|\neq 0$。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：

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

（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：

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

随着t趋近于 0，$\boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*)$在无限靠近$\boldsymbol{x}^*$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，但在任意一个邻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

为了寻找函数的全局最小点，可以先找到满足全局最小点必要条件的点——驻点。但大多数时候这并不可行。举个简单的例子，有一个二次函数：

$$f(x_1, x_2, \dots, x_n) = \frac{1}{2}\sum_{i,j} w_{ij} x_i x_j \qquad (3.26)$$

该二次函数对每一个自变量x_i的偏导数是：

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{n} w_{ij} x_j \qquad (3.27)$$

令梯度是零向量，求$x_1, x_2, \dots, x_n$，相当于解 n 元 1 次方程组：

$$\sum_{j=1}^{n} w_{ij} x_j = 0 \quad i = 1,2,\dots,n \qquad (3.28)$$

一般情况下，这需要$O(n^3)$的时间复杂度。当模型参数个数 n 非常大时（这在神经网络和深度学习中是必然的），求驻点解析解是不可接受的。例子（3.27）还仅仅是简单的二次函数的情况。当情况更复杂时，梯度为零向量的解析解甚至是不存在的。这就需要一些迭代的数值解法。

3.2.1　反向梯度场

如果$f(\boldsymbol{x})$是 n 元函数，则$\nabla f(\boldsymbol{x})$是 n 维向量。自变量空间中每一个点都存在一个反向梯度向量$-\nabla f(\boldsymbol{x})$，指向$f(\boldsymbol{x})$下降最快的方向。这就形成了一个向量场，或者说速度场。可以将$-\nabla f(\boldsymbol{x})$画成箭头，尾部移到$\boldsymbol{x}$的位置，将该向量场呈现出来。如图 3-8 所示。

图 3-8　反向梯度场

向量场在每一点指定了该位置的速度——方向和速率。在反向梯度场情况下，向量场指向的是函数值下降最快的方向，速率大小是函数值的下降速率。

将一个粒子（particle）从任意位置放入向量场中，它就会按照场指定的方向和速率运动。在反向梯度场情况下，粒子就是超函数值下降最快的方向运动。

局部极小点的梯度为零向量。它们是向量场的静止点。如果一个粒子处于稳定点上，它将不发生运动。同理局部极大点也是静止点。但是，局部极小点是稳定静止点，或者说吸引子（attractor）。当粒子偏离局部极小点一个小位移，它将被吸引向局部极小点。局部极大点和鞍点是不稳定静止点，或者说排斥子（repeller）。粒子位于局部极大点或鞍点时，它也是静止的，但是一旦有一个微小的扰动使它发生极小的位移，它将被推得远离该局部极大点或鞍点。如图 3-9 所示。

图 3-9　吸引子和排斥子

反向梯度场的性质，使我们可以从任意位置开始模拟粒子的运动。除非粒子初始位置刚好是不稳定静止点，否则粒子将向函数值下降的方向运动，并最终无穷逼近吸引子——局部极小点。

从理论上，反向梯度场只能保证解收敛到局部极小点而不能保证收敛到全局最小点，且不保证收敛的速度。实践中无法精确模拟粒子的运动。而只能以一种离散的方式近似模拟，会将带来更多问题，甚至不收敛。

3.2.2　梯度下降法及其问题

在计算机中模拟粒子在向量场中的运动，属于数值积分。梯度下降是一个简单的数值积分算法。伪代码如下：

$\boldsymbol{x} \leftarrow$ randomly initialized
while $\|\nabla f(\boldsymbol{x})\| \geq \varepsilon$ ：
　　$\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta \nabla f(\boldsymbol{x})$

ε是一个预设的阈值。当$\|\nabla f(\boldsymbol{x})\| < \varepsilon$时认为$\nabla f(\boldsymbol{x})$已经足够接近零向量，算法停止。也可以采用其他的停止标准。例如循环次数达到预设的最大值，或者函数值的下降幅度小于某个阈值。

η是另一个预设值，称为学习率（learning rate, LR）。每一步迭代中，自变量向$-\nabla f(\boldsymbol{x})$运动，运动的距离是$\eta \cdot \|\nabla f(\boldsymbol{x})\|$。$\eta$是梯度下降算法的一个关键超参数，它控制着粒子运动的行为。因为$\nabla f(\boldsymbol{x})$是$f(\boldsymbol{x})$的局部近似特性，在距离$\boldsymbol{x}$过远的地方$f(\boldsymbol{x})$的形状会有大的变化，这是从$\nabla f(\boldsymbol{x})$无法体现的。所以η不能设置得过大。但是如果η，收敛的速度会很慢。如图 3-10 所示。

图 3-10　学习率对梯度下降算法的影响

$f(\boldsymbol{x})$的“地形”有可能千奇百怪，例如“悬崖”，“峡谷”，“广袤的平原”都是一些对梯度下降形成负面干扰的病态情况。

“悬崖”的情形如图 3-11 所示。如果悬底是局部极小点所在位置，当粒子靠近崖底时，稍大的η会使粒子跨过崖底，爬上了对面的崖顶。崖顶有非常大的梯度，一下将粒子弹回了很远的地方。如图 3-11 所示。

图 3-11　“悬崖”对梯度下降法的影响

“峡谷”的情形如图 3-12 所示。局部极小点在谷底。这种情形与赫森矩阵的特征值大小有关，第 4 章会进行讲解。在“峡谷”情形下，梯度下降会发生震荡，轻则延缓收敛速度，重则导致不收敛。

在第 4 章介绍了函数二阶特性后，我们会知道在高维情况下局部极小点和局部极大点在理论上是稀少的，大部分驻点是鞍点。而更多是“广袤的平原”，在这样的区域里梯度非常小，这将导致收敛缓慢。如图 3-13 所示。

图 3-13 “广袤的平原”对梯度下降法的影响

本节讲解了梯度下降法最朴素的形式以及它会遇到的一些（非全部）问题。第 4 章介绍函数二阶特性——赫森矩阵后，将会对例如“峡谷”的成因进行理论上的说明，同时介绍一些梯度下降法的改进和变体。

3.3 运用梯度下降法训练逻辑回归

运用梯度下降法训练逻辑回归需要首先计算交叉熵损失函数对逻辑回归的参数$w_{i,i=1,2,\dots,n}$和b的偏导数。

交叉熵损失函数是每一个训练样本上的交叉熵损失之和。对第i个样本$\{\boldsymbol{x}^i, y^i\}$的交叉熵损失是：

$$loss(\boldsymbol{w}, b|\boldsymbol{x}^i, y^i) = -y^i log\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} - (1-y^i)log\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} \tag{3.29}$$

考虑$loss(\boldsymbol{w}, b|\boldsymbol{x}^i, y^i)$对$w_j$的偏导数。分两种情况考虑$y^i=1$和$y^i=0$。当$y^i=1$时：

$$\frac{\partial loss(\boldsymbol{w}, b|\boldsymbol{x}^i, y^i)}{\partial w_j} = \frac{e^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}x_j^i = (1-\hat{y}^i)(-x_j^i) = -(y^i-\hat{y}^i)x_j^i \tag{3.30}$$

x_j^i是$\boldsymbol{x}^i$的第j分量。式（3.30）中的$\hat{y}^i$是逻辑回归对$\boldsymbol{x}^i$的输出。当$y^i=0$时：

$$\frac{\partial loss(\boldsymbol{w}, b|\boldsymbol{x}^i, y^i)}{\partial w_j} = \frac{e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}x_j^i = (\hat{y}^i)x_j^i = -(y^i-\hat{y}^i)x_j^i \tag{3.31}$$

喜闻乐见的事情发生了，两种情况统一成一种情况：

$$\frac{\partial loss(\boldsymbol{w}, b|\boldsymbol{x}^i, y^i)}{\partial w_j} = -(y^i-\hat{y}^i)x_j^i \tag{3.32}$$

在继续进行之前先观察一下式（3.32）。括号中的$y^i-\hat{y}^i$是真实标签（1/0）与预测概率之差。这个差可以看做是模型在样本$\{\boldsymbol{x}^i, y^i\}$上产生的误差。

式（3.32）前面的负号控制更新方向：$y^i=1$表示当前是一个 A 类样本，希望$\hat{y}^i$越接近 1 越好。这时$-\left(y^i-\hat{y}^i\right)<0$。如果$x_j^i>0$，则$-\left(y^i-\hat{y}^i\right)x_j^i<0$，对$w_i$向梯度反方向更新是对$w_j$减去一个负数，使$w_j$变大。$w_j$变大会导致$\hat{y}^i$更接近 1。如果$x_j^i<0$，则$-\left(y^i-\hat{y}^i\right)x_j^i>0$，对$w_j$向梯度反方向更新是对$w_j$减去一个正数，使$w_j$变小。因为$x_j^i$是负数，所以$w_j$变小会导致$\hat{y}^i$更接近 1。对$y^i=0$的情况可做类似分析。这里不再赘述。式(3.32)可是视为将模型对训练样本$\left\{\boldsymbol{x}^i,y^i\right\}$的误差$-\left(y^i-\hat{y}^i\right)$以$x_j^i$为权重分配到每一个$w_j$上。

损失函数$loss(\boldsymbol{w},b)$是对全部训练样本的损失做平均，所以$loss(\boldsymbol{w},b)$对w_j的偏导数是每一个$loss\left(\boldsymbol{w},b|\boldsymbol{x}^i,y^i\right)$对$w_j$的偏导数的平均：

$$\frac{\partial loss(\boldsymbol{w},b)}{\partial w_j}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial loss\left(\boldsymbol{w},b|\boldsymbol{x}^i,y^i\right)}{\partial w_j}=-\frac{1}{m}\sum_{i=1}^{m}\left(y^i-\hat{y}^i\right)x_j^i \qquad (3.33)$$

现在考察$loss\left(\boldsymbol{w},b|\boldsymbol{x}^i,y^i\right)$对$b$的偏导数。类似的计算可得，$y^i=1$和$y^i=0$两种情况统一到一个表达式：

$$\frac{\partial loss\left(\boldsymbol{w},b|\boldsymbol{x}^i,y^i\right)}{\partial b}=-\left(y^i-\hat{y}^i\right) \qquad (3.34)$$

损失函数$loss(\boldsymbol{w},b)$对b的偏导数就是：

$$\frac{\partial loss(\boldsymbol{w},b)}{\partial b}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial loss\left(\boldsymbol{w},b|\boldsymbol{x}^i,y^i\right)}{\partial b}=-\frac{1}{m}\sum_{i=1}^{m}\left(y^i-\hat{y}^i\right) \qquad (3.35)$$

有了各个偏导数就可以计算$loss(\boldsymbol{w},b)$对$w_{i,i=1,2,\ldots,n}$和b的梯度了：

$$\nabla loss(\boldsymbol{w},b)=-\frac{1}{m}\sum_{i=1}^{m}\left(y^i-\hat{y}^i\right)\begin{pmatrix}x_1^i\\x_2^i\\\vdots\\x_n^i\\1\end{pmatrix} \qquad (3.36)$$

有了损失函数的梯度，就可以应用梯度下降法训练逻辑回归模型了。

小结：

本章首先回顾了多元微积分的相关知识，尤其是梯度这个概念以及它的内涵。之后介绍了最原始的梯度下降法，讨论了它的种种问题。本章没有讨论改善这些问题的办法。在第 4 章介绍了函数的局部二阶特性，以及对函数图形的病态情况的成因进行分析后，会介绍一些针对原始梯度下降法的改进措施。

本章最后介绍了如何运用梯度下降法训练逻辑回归模型。阅读完本章，读者应该对逻辑回归的训练过程有了一个较透彻的认识。