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。在机器学习模型训练的语境下，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。介绍多元函数的梯度、方向导数、偏导数等概念。在某一点附近可以用切平面近似表示函数本身。切平面的朝向和倾斜程度蕴含在梯度之中。这些信息就是函数的局部的一阶信息。

随后，本章介绍梯度下降法。梯度下降法利用梯度确定自变量空间中使函数值下降最快的方向，然后向该方向前进一段距离。迭代地重复此步骤，希望使函数值不断下降，乃至找到函数的全局最小点。具备了多元微积分的相关知识后，能深刻地理解梯度下降法。

由于梯度下降法只利用了局部一阶特性，所以它是短视的。梯度下降法本身的离散性质也会带来的种种问题。本章举例几个梯度下降法存在的问题。这些问题的成因以及规避和改进办法将在第 4 章介绍函数二阶特性后加以说明。

最后阐述如何运用梯度下降法训练逻辑回归模型。阅读完本章，读者应能透彻理解梯度下降法的原理和局限，并完整地理解逻辑回归模型。

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。微分刻画了函数的局部近似特性。寻找函数的最小点就利用了这些局部近似特性。

3.1.1 梯度

首先，回忆一下一元函数$f(x)$的可导性及其导数$f'(x)$：

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

如果极限（3.1）存在，则称$f(x)$在x可导。h是一个变化量。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）里的商$\big(f(x+h)-f(x)\big)/h$是割线的斜率。随着$h$趋近于 0，割线趋近于$f(x)$在$x$的切线。割线斜率的极限是切线的斜率。如图 3-1 所示。

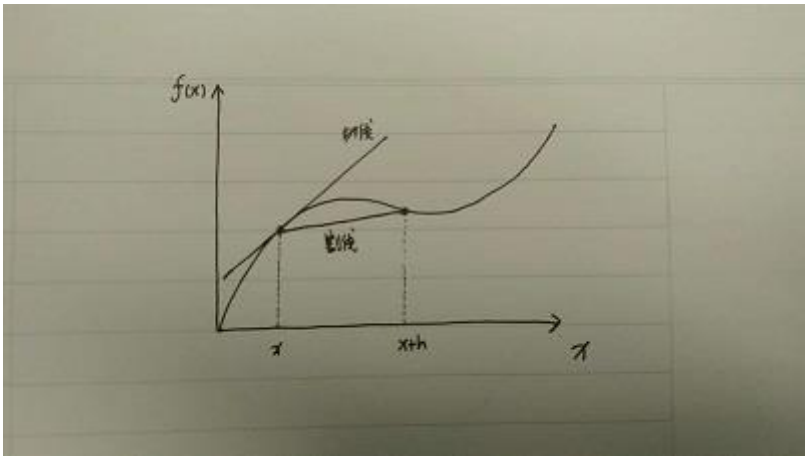

图 3-1　一元函数的割线、切线和斜率

$\big(f(x+h)-f(x)\big)/h$也是自变量从x变化到$x+h$时，函数值$f(x)$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。可以像式（3.1）那样用瞬时变化率定义导数。但在多元情况下自变量是向量，它可以沿无数方向运动。这时就不能用瞬时变化率定义多元函数$f(\boldsymbol{x})$的导数。

一元函数的可导性还有另一种定义，即在x附近是否能用直线近似表示$f(x)$。这种定义可以扩展到多维的情况。构造一个以变化量h为自变量的仿射变换$g(h)$：

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

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

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

所以作为h的函数的$f(x+h)$可以写成一个仿射变换加余项的形式：

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

$\mathcal{R}(h)$是两个连续函数的差，它也是连续的。因为$\mathcal{R}(0)=0$，所以有$\lim_{h\to 0}\mathcal{R}(h)=0$。当变化量$h$趋向于消失时，$\mathcal{R}(h)$也趋向于消失。这还不够。根据式（3.3），有：

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

变化量h趋近于 0 时，$\mathcal{R}(h)$与h之比$\mathcal{R}(h)/\mathrm{h}$趋近于 0。这就是说随着变化量的消失，$\mathcal{R}(h)$也消失，而且比变化量消失得更快。这种情况称$\mathcal{R}(h)$是$h$的高阶无穷小。

当$h=0$时仿射变换$g(h)$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。如果将该直线平移，使原来的$(0 \quad f(x))^{\mathrm{T}}$移动到$(x \quad f(x))^{\mathrm{T}}$。平移后的直线经过$(x \quad f(x))^{\mathrm{T}}$，斜率为$f'(x)$。它就是$f(x)$在$x$的切线。如图 3-2 所示。

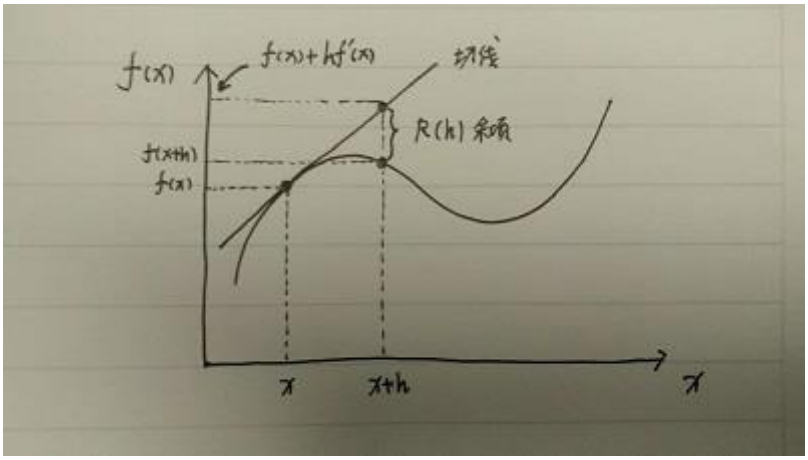

图 3-2　可导函数的仿射近似——切线

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：

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

式（3.6）说明$f(x)$在x可导，导数$f'(x)=a$。

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

现在将这种可导性的定义扩展到多元函数$f(\boldsymbol{x})$。将变化量$\boldsymbol{h}$作为自变量，如果$f(\boldsymbol{x}+\boldsymbol{h})$可以被一个仿射变换近似表示：

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

其中$\mathcal{R}(\boldsymbol{h})$是变化量的长度$\|\boldsymbol{h}\|$的高阶无穷小：

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

这时称多元函数$f(\boldsymbol{x})$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：

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

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，经过点$(\boldsymbol{o}^{\mathrm{T}} \quad f(\boldsymbol{x}))^{\mathrm{T}}$。将图像平移，使$(\boldsymbol{o}^{\mathrm{T}} \quad f(\boldsymbol{x}))^{\mathrm{T}}$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。切平面是$f(\boldsymbol{x})$在$\boldsymbol{x}$附近的一阶近似。切平面的特性就是$f(\boldsymbol{x})$的局部一阶特性。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。

切平面的全部特性体现在$\nabla f(\boldsymbol{x})$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。

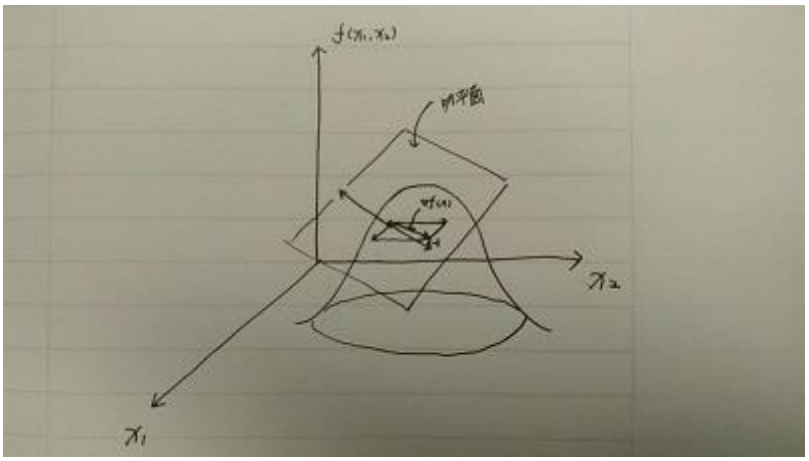

图 3-3　多元函数的切平面

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。$\boldsymbol{d}$是单位向量。$\boldsymbol{d}$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，$l(0) = \boldsymbol{x}$。可以把该直线当作自变量空间中一个以$\boldsymbol{x}$为原点，以$\boldsymbol{d}$的方向为正方向的坐标轴。$l(t)$的$t$值就是在这个坐标轴上的坐标。

接下来在$l(t)$基础上定义一个复合函数$(f \oplus l)(t)$：

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

$(f \oplus l)(t)$是自变量为t的一元函数。根据式（3.1），$(f \oplus l)(t)$在 0 的导数是：

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

$\nabla_{\boldsymbol{d}} f(\boldsymbol{x})$是$f(\boldsymbol{x})$在$\boldsymbol{x}$沿$\boldsymbol{d}$的方向导数（directional derivative）。上文提到过，直线$l(t)$定义了一个坐标轴，$\boldsymbol{x}$在该坐标轴上。如果限制自变量只能在这条坐标轴上变化，这就相当于定义了一个一元函数。方向导数是这个一元函数在$\boldsymbol{x}$瞬时变化率。如图 3-4 所示。坐标轴是有方向的，其方向由$\boldsymbol{d}$的方向确定。所以$-\boldsymbol{d}$确定了方向相反的坐标轴，函数的变化率相反。即$\nabla_{-\boldsymbol{d}} f(\boldsymbol{x}) = -\nabla_{\boldsymbol{d}} f(\boldsymbol{x})$。

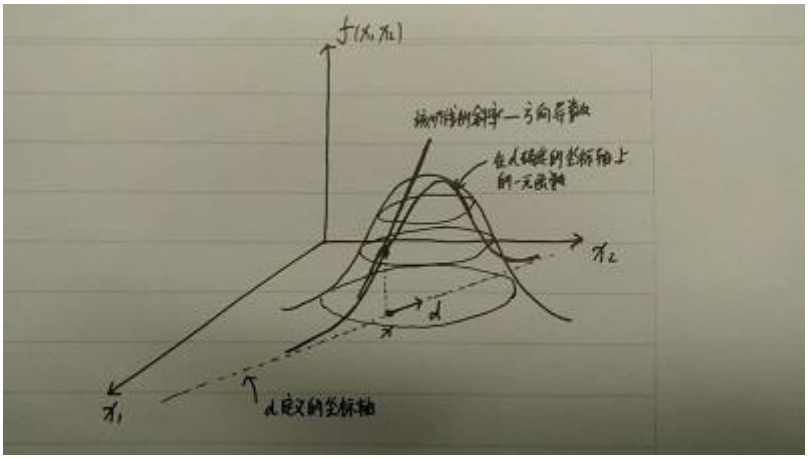

图 3-4　方向导数

$f(\boldsymbol{x})$在$\boldsymbol{x}$沿各个方向的方向导数都蕴含在梯度$\nabla f(\boldsymbol{x})$中。因为$f(\boldsymbol{x})$在$\boldsymbol{x}$可导，根据式（3.7）有：

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

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$\|h\boldsymbol{d}\|$也趋近于 0。这时有：

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

所以$\mathcal{R}(\mathrm{h}\boldsymbol{d})$是$h$的高阶无穷小。另外因为$f(\boldsymbol{x})=(f\oplus l)(0)$，所以式（3.13）表明$(f\oplus l)(t)$在 0 的导数是$\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$，即$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})=\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$。现在有：

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

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（$\theta=0$）时$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$最大。所以$\nabla f(\boldsymbol{x})$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$\nabla f(\boldsymbol{x})$的反方向（$\theta=\pi$）$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})$最小，为$-\|\nabla f(\boldsymbol{x})\|\leq 0$。$-\nabla f(\boldsymbol{x})$是$f(\boldsymbol{x})$变化率为负且最小的方向，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 维是-1 说明$\boldsymbol{w}$向下指向$x_1x_2$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。

自变量沿任意方向$\boldsymbol{d}$的运动可分解为两个分量：沿$\nabla f(\boldsymbol{x})$方向的分量和垂直于$\nabla f(\boldsymbol{x})$的分量。垂直于$\nabla f(\boldsymbol{x})$的方向上$\nabla_{\boldsymbol{d}}f(\boldsymbol{x})=0$。自变量沿$\boldsymbol{d}$的运动一部分摊在了垂直分量上，这部分运动不会导致$\nabla f(\boldsymbol{x})$变化。所以$f(\boldsymbol{x})$的变化率就打了折扣，折扣系数正是自变量运动沿$\nabla f(\boldsymbol{x})$方向分量所占的“份额”——$\cos\theta$。如图 3-5 所示。

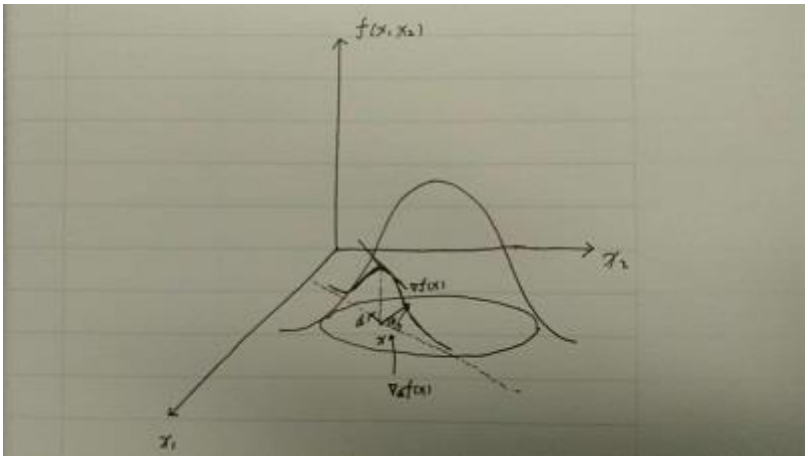

图 3-5　梯度与方向导数

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：

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

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：$\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2},\cdots,\frac{\partial f}{\partial x_n}$。

根据式（3.12），有：

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

$f(\boldsymbol{x})$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。它们的方向分别是n个坐标轴的正方向。

$\nabla f(\boldsymbol{x})$与$\boldsymbol{e}_i$的内积是$\nabla f(\boldsymbol{x})$的第i分量。根据式（3.17），有：

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

所以梯度$\nabla f(\boldsymbol{x})$的第i分量是$f(\boldsymbol{x})$对自变量第i分量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。

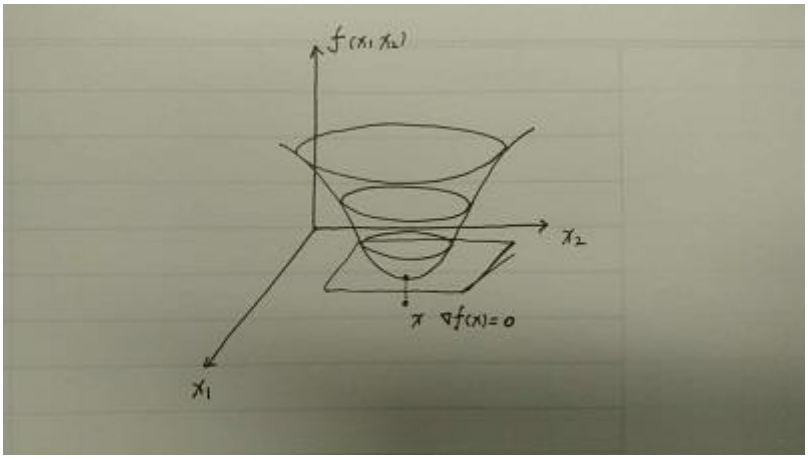

图 3-6　驻点的切平面

$f(\boldsymbol{x})$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方向导数都为 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，该邻域内所有点的函数值都不小于$f(\boldsymbol{x}^*)$。用公式表示就是：

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

如果对于全部$\boldsymbol{x}$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。

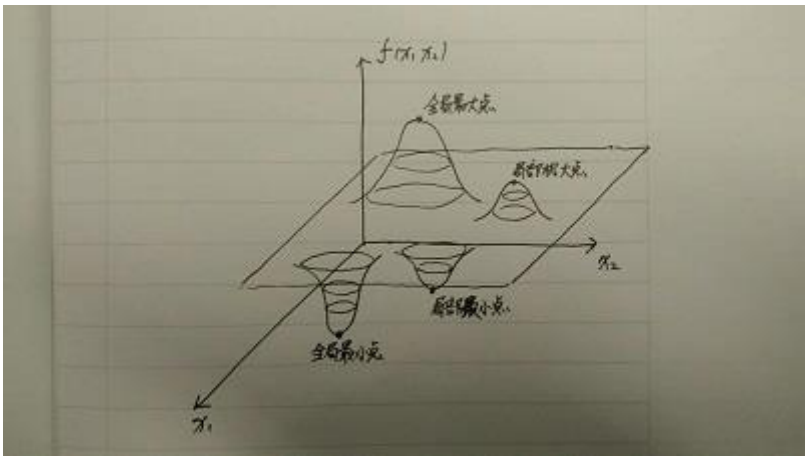

图 3-7　局部极小点和全局最小点

局部极小点$\boldsymbol{x}^*$一定是驻点，即$\|\nabla f(\boldsymbol{x}^*)\| = 0$。运用反证法。假设$\|\nabla f(\boldsymbol{x}^*)\| \neq 0$。从$\boldsymbol{x}^*$点出发沿$-\nabla f(\boldsymbol{x}^*)$方向产生一个位移$-t\nabla f(\boldsymbol{x}^*)$，$t > 0$。根据$f(\boldsymbol{x})$在$\boldsymbol{x}^*$的可导性，有：

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

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

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

（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$\nabla f(\boldsymbol{x}^*)$只能是零向量，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或者鞍点（saddle point）。鞍点的梯度也是零向量，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 3-8 所示。

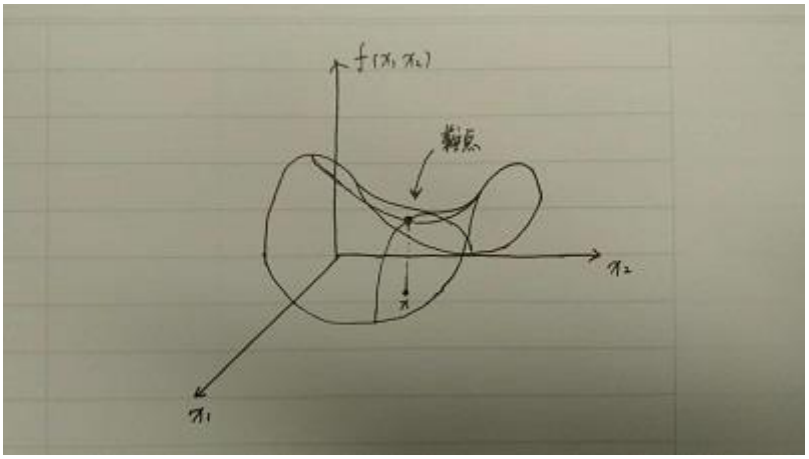

图 3-8　鞍点

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揭示：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，有一个二次型（quadratic form）函数：

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

满足$w_{ij} = w_{ji}$。该二次型对每一个自变量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$x_1, x_2, \cdots, x_n$，相当于解n元 1 次方程组：

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

一般情况下，这需要$O(n^3)$的时间复杂度。当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型的情况。当情况更复杂时驻点的解析解可能不存在。这就需要数值解法。

3.2.1 反梯度场

如果$f(\boldsymbol{x})$是n元函数，则$\nabla f(\boldsymbol{x})$是n维向量。可导函数$f(\boldsymbol{x})$在自变量空间中每一个点都有一个反梯度向量$-\nabla f(\boldsymbol{x})$，指向$f(\boldsymbol{x})$下降最快的方向。这形成了一个速度场。可以将$-\nabla f(\boldsymbol{x})$画成箭头，尾部移到$\boldsymbol{x}$的位置，把这个速度场呈现出来。如图 3-9 所示。

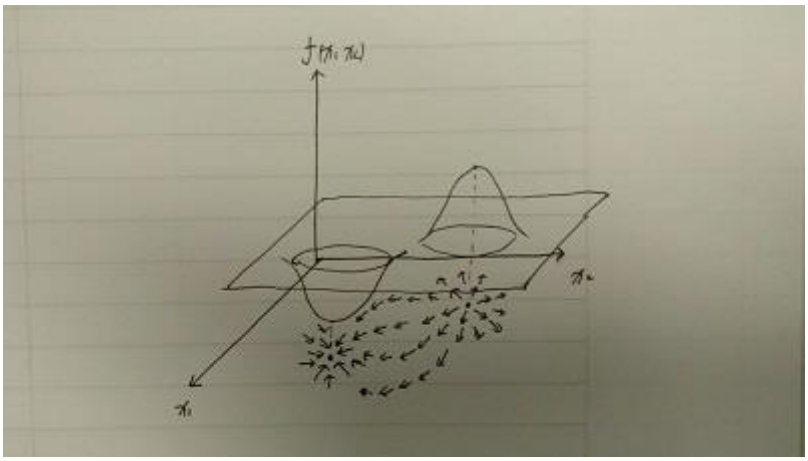

图 3-9　反梯度场

速度场在每一点指定了该位置的速度——方向和速率。在反梯度场情况下，速度指向函数下降最快的方向，速率大小是函数的下降速率。将一个粒子（particle）从任意位置放入速度场中，它就会按照场指定的方向和速率运动。在反梯度场情况下，粒子就是朝函数下降最快的方向运动。

这里澄清一下反梯度场的物理意义。在 2 维情况下，将$f(\boldsymbol{x})$的图像看作一幅起伏不平的地形，设重力加速度是g。有一个质量是$m = 1/g$的小球被放在任意位置。地面对小球的支持力垂直于坡面指向上方。即该支持力是沿着$f(\boldsymbol{x})$在该点向上指的法向量的方向，即$-w = (-\nabla f(\boldsymbol{x})_1 \quad -\nabla f(\boldsymbol{x})_2 \quad 1)^T$的方向。$-w$与$x_1x_2$平面的夹角$\theta$的正弦是$\tan\theta = 1/\|\nabla f(\boldsymbol{x})\|$。该支持力的垂直分量抵消重力$mg = 1$，水平分量的大小则是$mg/\tan\theta = \|\nabla f(\boldsymbol{x})\|$。水平分量的方向是$-w$向$x_1x_2$平面的投影$-\nabla f(\boldsymbol{x})$的方向。所以小球在地形上受到的水平作用力正是$-\nabla f(\boldsymbol{x})$。小球的水平加速度就是$-g\nabla f(\boldsymbol{x})$。所以粒子在反梯度场中的运动并非模拟小球在函数地形上自然滚落。如图 3-10 所示。

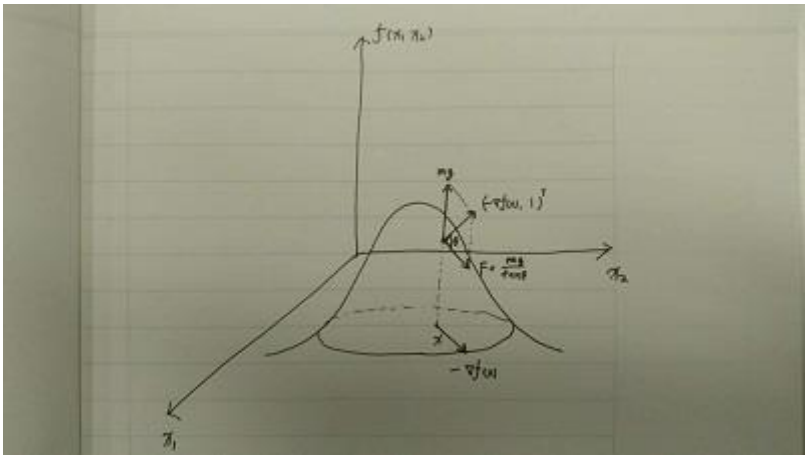

图 3-10　反梯度场的物理意义

局部极小点的梯度为零向量。它们是反梯度场的静止点。如果一个粒子处于静止点上，它将不发生运动。同理局部极大点也是静止点。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 3-11 所示。

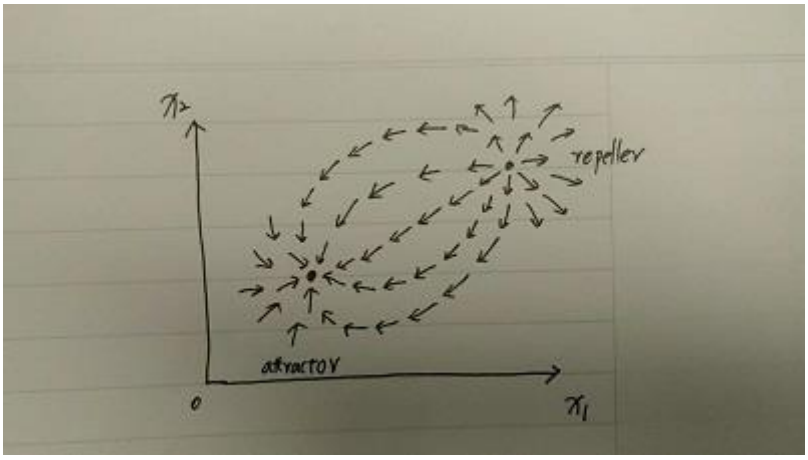

图 3-11　吸引子和排斥子

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静止点，否则粒子将向函数值下降的方向运动，并最终逼近吸引子——局部极小点。

从理论上，反梯度场只能保证粒子运动到局部极小点，不能保证运动到全局最小点。收敛的速度也没有保证。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。这会将带来更多问题，甚至不收敛。

3.2.2　梯度下降法

在计算机中模拟粒子在速度场中的运动，属于数值积分问题。梯度下降法（gradient descent, GD）是一个简单的数值积分算法。伪代码如下：

$\boldsymbol{x} \leftarrow$ randomly initialized

while $\|\nabla f(\boldsymbol{x})\| \geq \varepsilon$:

$\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta \nabla f(\boldsymbol{x})$

$\varepsilon > 0$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阈值。

$\eta > 0$是另一个预设值，称为学习率（learning rate, LR）或步长。每一次迭代自变量向$-\nabla f(\boldsymbol{x})$方向运动，运动的距离是$\eta \cdot \|\nabla f(\boldsymbol{x})\|$。$\eta$是梯度下降法的一个关键参数。可以保证的是，在某个点$\boldsymbol{x}$，能够找到一个合适的步长η_x使得$f\big(\boldsymbol{x} - \eta_x \nabla f(\boldsymbol{x})\big) < f(\boldsymbol{x})$。也就是从$\boldsymbol{x}$出发移动$-\eta_x \nabla f(\boldsymbol{x})$可以使函数值下降。这是因为：

$$f\big(\boldsymbol{x} - \eta_x \nabla f(\boldsymbol{x})\big) = f(\boldsymbol{x}) - \eta_x \|\nabla f(\boldsymbol{x})\|^2 + \mathcal{R}\big(-\eta_x \nabla f(\boldsymbol{x})\big) \qquad (3.29)$$

$\mathcal{R}\big(-\eta_x \nabla f(\boldsymbol{x})\big)$是$\|-\eta_x \nabla f(\boldsymbol{x})\|$的高阶无穷小。式（3.29）与式（3.23）相似，可以证明存在一个$\varepsilon > 0$，当$\eta_x < \varepsilon$时有$f\big(\boldsymbol{x} - \eta_x \nabla f(\boldsymbol{x})\big) < f(\boldsymbol{x})$。所以称$-\nabla f(\boldsymbol{x})$是确保下降的方向。只是对于不同的$\boldsymbol{x}$，$\eta_x$也不同，而且无法计算出$\eta_x$的值。所以只能选择固定的步长$\eta$。

3.2.3　梯度下降法的问题

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、未知的变化，

这是$\nabla f(\boldsymbol{x})$无法体现的。所以如果η设置得过大，函数值有可能不降反升。较小的η更有可能保证函数值下降。但是如果η过小，收敛的速度会很慢。

函数图像的“地形”千奇百怪，有很多病态情况都会对梯度下降法的效果产生负面影响。本节试举几例。

“悬崖”如图 3-12 所示。如果悬崖底是局部极小点所在位置，当解靠近悬崖底时，过大的η会使解一步跨过崖底，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后方。

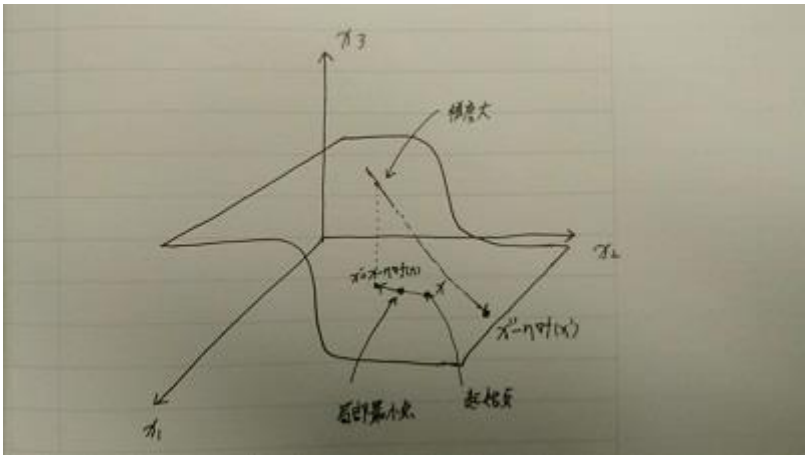

图 3-12　“悬崖”对梯度下降法的影响

“峡谷”如图 3-13 所示。局部极小点在谷底。在峡谷里，梯度下降过程会发生震荡，轻则延缓收敛速度，重则导致不收敛。第 4 章介绍函数二阶特性后，我们会知道峡谷的成因与赫森矩阵的各个特征值相对大小有关。并且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。

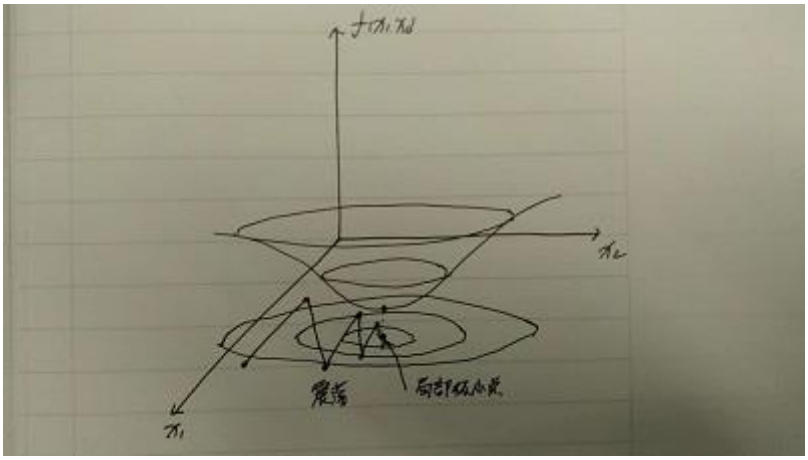

图 3-13　“峡谷”对梯度下降法的影响

“广袤的平原”如图 3-14 所示。在这样的区域里梯度非常小，这将导致收敛缓慢。

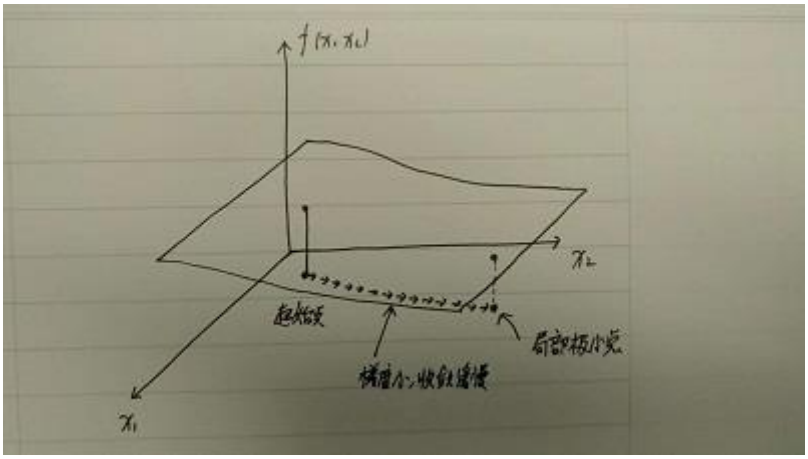

图 3-14　“广袤的平原”对梯度下降法的影响

本节列举了原始梯度下降法会遇到的一些（非全部）问题。第 4 章将介绍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$w_{i=1\ldots n}$和b的梯度。根据式（3.21），计算梯度需要计算各个偏导数。

交叉熵损失是每一个训练样本上的损失的平均。第i个训练样本$\{\boldsymbol{x}^i, y^i\}$的损失是：

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

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：

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

x_j^i是$\boldsymbol{x}^i$的第j分量。式（3.31）中的$\hat{y}^i$是逻辑回归对$\boldsymbol{x}^i$的输出。当$y^i = 0$时：

$$\frac{\partial loss(\boldsymbol{w},b|\boldsymbol{x}^i,y^i)}{\partial w_j}=\frac{e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}x_j^i=\left(\hat{y}^i\right)x_j^i=-\left(y^i-\hat{y}^i\right)x_j^i \qquad (3.32)$$

喜闻乐见的事情发生了，两种情况统一成一种情况：

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

在继续之前先观察一下式（3.33）。括号中的$y^i-\hat{y}^i$是真实标签（1/0）与预测概率之差。这个差可以看作模型在样本$\{\boldsymbol{x}^i,y^i\}$上的误差。更新参数时加上负梯度，所以$\left(y^i-\hat{y}^i\right)x_j^i$会被加到$w_j$上。$y^i-\hat{y}^i$是真实值与预测值的差距。这个差距以$x_j^i$为权重分配到$w_j$上。误差分配是看待梯度下降的一个视角。在后文介绍神经网络的反向传播算法时会看到误差不仅在同一层内部分配，还要在层与层之间分配。

回到梯度计算中来。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：

$$\frac{\partial loss(\boldsymbol{w},b)}{\partial w_j}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial loss(\boldsymbol{w},b|\boldsymbol{x}^i,y^i)}{\partial w_j}=-\frac{1}{m}\sum_{i=1}^{m}\left(y^i-\hat{y}^i\right)x_j^i \qquad (3.34)$$

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

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

损失函数$loss(\boldsymbol{w},b)$对b的偏导数就是：

$$\frac{\partial loss(\boldsymbol{w},b)}{\partial b}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial loss(\boldsymbol{w},b|\boldsymbol{x}^i,y^i)}{\partial b}=-\frac{1}{m}\sum_{i=1}^{m}\left(y^i-\hat{y}^i\right) \qquad (3.36)$$

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

$$\nabla loss(\boldsymbol{w}, b) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^i - \hat{y}^i\right)\begin{pmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_n^i \\ 1 \end{pmatrix} \qquad (3.37)$$

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。逻辑回归模型训练的伪代码如下：

$\boldsymbol{w} \leftarrow$ randomly initialized
$b \leftarrow$ randomly initialized

while $\|\nabla loss(\boldsymbol{w}, b)\| \geq \varepsilon$:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \frac{1}{m}\eta\sum_{i=1}^{m}\left(y^i - \hat{y}^i\right)\begin{pmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_n^i \end{pmatrix}$$

$$b \leftarrow b + \frac{1}{m}\eta\sum_{i=1}^{m}\left(y^i - \hat{y}^i\right)$$

上述训练过程本质上是根据交叉熵损失函数的梯度更新模型参数。但我们可以抛开损失函数和梯度来观察一下$\boldsymbol{w}$和b的更新公式。

对于 A 类训练样本，$y^i = 1$。模型的输出$\hat{y}^i$是 0-1 之间的一个概率值。我们希望提高$\hat{y}^i$使它更接近 1。更新公式中$y^i - \hat{y}^i > 0$。如果$x_j^i > 0$，对w_j的更新是加上一个正值；如果$x_j^i < 0$，对w_j的更新是加上一个负值。两种情况都是增加$w_j x_j^i$，也就是增加$\hat{y}^i$。

对于 B 类训练样本，$y^i = 0$。我们希望降低$\hat{y}^i$使它更接近 0。更新公式中$y^i - \hat{y}^i < 0$。如果$x_j^i > 0$，对w_j的更新是加上负值；如果$x_j^i < 0$，算法对w_j的更新是加上正值。两种情况都是减小$w_j x_j^i$，也就是减小$\hat{y}^i$。

对b的更新不依赖模型输入$\boldsymbol{x}$。A 类样本增大b，B 类样本压低b。变化的幅度与模型输出概率与理想值（1 或 0）的差距有关。某样本上模型输出概率与理想值差距越大，则对b的影响越大。

每一个训练样本都将模型参数向对自己来说更理想的方向拉。原始梯度下降算法取所有训练样本的更新的平均。所有训练样本的“合力”将模型参数拉向能更好地分类训练集的方向。最早的感知机模型的更新规则也是从类似的视角出发。但是每一次更新取一个训练样本，将模型参数拉向它更理想的方向。之后再取下一个训练样本，如此循环往复。在优化损失函数的语境下，这其实是随机梯度下降。随机梯度下降将在第 4 章详细介绍。

3.4 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。之后介绍了原始的梯度下降法，讨论了它的种种问题。本章没有涉及改善这些问题的办法。在第 4 章讨论函数的局部二阶特性后，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，读者应该对梯度下降法和逻辑回归的训练过程有了较透彻的认识。