第 3 章 逻辑回归模型的训练

本章首先回顾多元微积分基础。阐述描述多元函数局部特性的梯度、偏导数、方向导数、赫森矩阵等概念。之后介绍多元函数的驻点、局部极小点、全局最小点和鞍点。

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。赫森矩阵包含函数的二阶特性。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，将上述优化算法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$f(x)$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：

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

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：

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

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

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

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”——$\mathbf{cos}\,\theta$。如图 3-5 所示。

图 3-5　梯度与方向导数