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。该超平面的法向量是 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$\nabla f(\boldsymbol{x})^{\mathrm{T}}$中：$\nabla f(\boldsymbol{x})^{\mathrm{T}}$的方向决定超平面$g(\boldsymbol{h})$的朝向，$\left\|\nabla f(\boldsymbol{x})^{\mathrm{T}}\right\|$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$\nabla f(\boldsymbol{x})^{\mathrm{T}}$中。