—武大本科生课程

MLPR-第3讲数学知识补充

武汉大学计算机学院 袁志勇

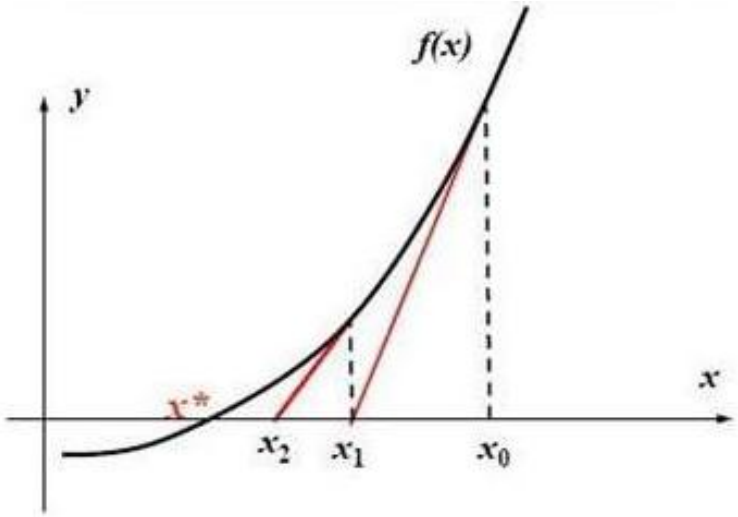
Email: yuanzywhu@163.com

一. 牛顿迭代法

牛顿迭代法是牛顿在17世纪提出的一种在实数域和复数域上近似求解方程根的方法。

多数方程不存在求根公式，因此求精确根非常困难，甚至不可能，从而寻找方程的近似根就显得特别重要。方法是使用函数f(x)的Taylor级数的前面几项来寻找方程f(x) = 0的根。

牛顿迭代法(*Newton iterative method*)是求方程根的重要方法之一，其最大优点是在方程f(x) = 0的单根附近具有平方收敛(*quadratic convergence*)，而且该法还可以用来求方程的重根、复根。该方法广泛应用于科学计算的计算机编程中。



设x^*是$f(x)=0$的根，选取$x0$作为x^*初始近似值，过点$(x0,f(x0))$做曲线$y=f(x)$的切线L，L的方程为$y=f(x0)+f'(x0)(x-x0)$，求出L与x轴交点的横坐标 $x1=x0-f(x0)/f'(x0)$，称$x1$为x^*的一次近似值。过点$(x1,f(x1))$做曲线$y=f(x)$的切线，并求该切线与x轴交点的横坐标 $x2=x1-f(x1)/f'(x1)$，称$x2$为x^*的二次近似值。重复以上过程，得x^*的近似值序列，其中：

$$x(k+1)=x(k)-f(x(k))/f'(x(k))$$

称$x(k+1)$为x^*的$k+1$次近似值，上式称为**牛顿迭代公式**。

解非线性方程f(x)=0的牛顿法是把非线性方程线性化的一种近似方法。把f(x)在x0点附近展开成Taylor级数 **f(x) = f(x0)+(x-x0)f’(x0)+(x-x0)²*f”(x0)/2! +…** 取其线性部分，作为非线性方程f(x) = 0的近似方程，即Taylor展开的前两项，则有f(x0)+f’(x0)(x-x0)=f(x)=0 。设f‘(x0)≠0，则其解为x1=x0-f(x0)/f’(x0) 。这样，得到牛顿法的一个迭代序列：**x(k+1)=x(k)-f(x(k))/f'(x(k))**。

二. 梯度(Gradient)

在单变量的实值函数的情况，梯度就是导数。若实值函数为线性函数，梯度就是该直线的斜率。

以二元函数为例说明，设函数$z=f(x,y)$在平面区域D内具有一阶连续偏导数，则$f(x,y)$在对应点$P(x,y)\in D$上的梯度为一个二维向量：

$$\nabla f(x,y)=Grad[f(x,y)]=\begin{bmatrix}\dfrac{\partial f}{\partial x}\\ \dfrac{\partial f}{\partial y}\end{bmatrix}$$

$$\text{梯度的模/幅度}\left\|\nabla f(x,y)\right\|=\sqrt{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2}$$

满足梯度$\nabla f(x^*,y^*)=0$的点称为驻点/平稳点. 在区域内部，极值点必为驻点，而驻点不一定为极值点。

记号：多元函数的偏导——梯度

$$f(\boldsymbol{X}) = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_{n+1} = \boldsymbol{W}_0^{\mathrm{T}} \boldsymbol{X} + w_{n+1}$$

$$\nabla f(\boldsymbol{X}) = \left[\frac{\partial f}{\partial x_i}\right]_{n\times 1} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \mathrm{L} \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \mathrm{L} \\ w_n \end{bmatrix} = \boldsymbol{W}$$

在工程应用中，为了方便起见，有时将梯度的幅度简称为梯度。

在图像处理与分析中，为了降低图像的运算量，常用绝对值代替平方根运算，即：

$$\|G[f(x,y)]\| \approx \left|\frac{\partial f}{\partial x}\right| + \left|\frac{\partial f}{\partial y}\right|$$

在图像处理及图像模式识别中，有一些典型的梯度快速求解方法。

如，一种方法是将微分$\frac{\partial f}{\partial x}$和$\frac{\partial f}{\partial x}$近似用差分来代替，

沿x和y方向的一阶差分可写成：

$$\frac{\partial f}{\partial x} \approx \Delta_x f(i,j) = f(i+1,j) - f(i,j)$$

$$\frac{\partial f}{\partial y} \approx \Delta_y f(i,j) = f(i,j+1) - f(i,j)$$

梯度定义的一般形式:

设函数f(**X**)是n元实函数, 其中向量**X**=$[x_1, x_2, \ldots, x_n]^T$

则f(**X**)在**X**处的梯度/一阶导数定义为:

$$\nabla f(\mathbf{X}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{X}} f(\mathbf{X}) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_i}, \ldots, \frac{\partial f}{\partial x_n} \right]^T$$

梯度是一个向量, 它的分量$\frac{\partial f}{\partial x_i}$表示自变量分量$x_i$方向上的变化速率.

梯度方向是自变量增加时f(**X**)增长最快的方向, 因此负梯度方向是f(**X**)减小最快的方向.

启发: 在求某函数极大值时, 若沿梯度的方向走, 就可以最快地到达极大值; 若沿负梯度方向走, 则可以最快地求得最小值. 梯度(下降)法就是根据这一思想得出的.

在梯度法对***W***权向量的优化中, ∇J的方向是***W***增加时*J*增长最快的方向, 因此($-\nabla J$)的方向是***W***增加时*J*减小最快的方向。梯度法就是用这个负梯度向量的值对权向量***W***进行修正, 实现准则函数达到极小值的目的。

三. Hessian矩阵(赫森/海色矩阵)

n元函数f(**X**)在**X**处的二阶偏导数或Hessian(赫森)矩阵定义如下:

$$\nabla^2 f(\mathbf{X}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

若梯度$\nabla f(\mathbf{X})$的每个分量函数在$\mathbf{X}$处都连续，则称$f(\mathbf{X})$在$\mathbf{X}$处一阶连续可微.

若Hessian矩阵$\nabla^2 f(\mathbf{X})$的各个分量函数都连续，则称$f(\mathbf{X})$在$\mathbf{X}$处二阶连续可微.

若f在区域D的每一点都连续可微，则称f在D上一阶连续可微. 若f在区域D的每一点都二阶连续可微，则称f在D上二阶连续可微.

由上述定义不难发现，若f在$\mathbf{X}$上二阶连续可微，则：

$$\frac{\partial f(\mathbf{X})}{\partial x_i \partial x_j} = \frac{\partial f(\mathbf{X})}{\partial x_j \partial x_i} \qquad i, j = 1, 2, \ldots, n$$

即Hessian矩阵$\nabla^2 f(\mathbf{X})$是对称阵.

Hessian矩阵被应用于采用牛顿法解决的大规模优化问题.

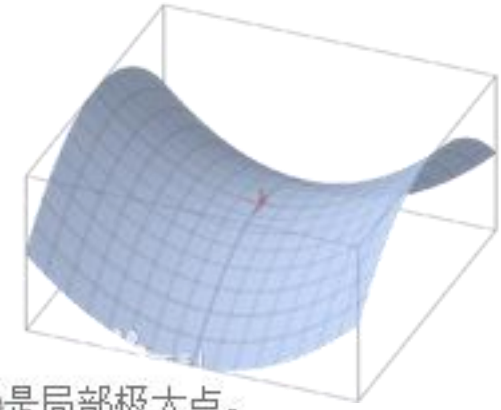
HESSIAN矩阵基本应用举例：

在 R^2→R 的函数的应用

给定二阶导数连续的函数$f:\mathbb{R}^2\to\mathbb{R}$，海色矩阵的行列式，可用于分辨f的临界点是属于鞍点还是极值。

对于f的临界点(x0,y0)一点，有 $\frac{\partial f(x_0,y_0)}{\partial x}=\frac{\partial f(x_0,y_0)}{\partial y}=0$，然而凭一阶导数不能判断它是鞍点、局部极大点还是局部极小点。海色矩阵可能解答这个问题。

$$H=\begin{vmatrix}\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\,\partial y}\\ \frac{\partial^2 f}{\partial y\,\partial x} & \frac{\partial^2 f}{\partial y^2}\end{vmatrix}=\frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2}-(\frac{\partial^2 f}{\partial y\,\partial x})^2$$



- H > 0：若$\frac{\partial^2 f}{\partial x^2}>0$，则(x0,y0)是局部极小点；若$\frac{\partial^2 f}{\partial x^2}<0$，则(x0,y0)是局部极大点。
- H < 0：(x0,y0)是鞍点。
- H = 0：二阶导数无法判断该临界点的性质，得从更高阶的导数以泰勒公式考虑。

思考：设上图中$z=x^2-y^2$，求其在驻点(0,0)处的Hessian矩阵。

四. 关于向量微分的一些有用等式

表：关于向量微分的一些等式

$f(\mathbf{w})$	$\frac{\partial f}{\partial \mathbf{w}}$
$\mathbf{w}^T\mathbf{x}$	$\mathbf{x}$
$\mathbf{x}^T\mathbf{w}$	$\mathbf{x}$
$\mathbf{w}^T\mathbf{w}$	$2\mathbf{w}$
$\mathbf{w}^T\mathrm{A}\,\mathbf{w}$	$2\mathrm{A}\,\mathbf{w}$

A：方阵

例：$\mathbf{x}=[x_1 \;\; x_2 \;\; 1]^{\mathrm{T}}$, $\mathbf{w}=[w_1 \;\; w_2 \;\; w_3]^{\mathrm{T}}$

对于$f(\mathbf{x}, \mathbf{w})=\mathbf{w}^T\mathbf{x}=w_1x_1+w_2x_2+w_3$

有$$\frac{\partial f(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}=\begin{bmatrix}\dfrac{\partial f}{\partial w_1}\\ \dfrac{\partial f}{\partial w_2}\\ \dfrac{\partial f}{\partial w_3}\end{bmatrix}=\begin{bmatrix}x_1\\ x_2\\ 1\end{bmatrix}=\mathbf{x}$$

五. Lagrange乘数法(Lagrange optimization)

在某些约束条件下，使函数f(x)取到极值的自变量x0的值。根据约束条件的不同，可以分为如下几种情况：(1)一个等式约束方程的情况；(2)多个等式约束方程的情况；(3)多个不等式约束方程的情况。

这里仅介绍第(1)种情况。

若约束条件可以表示为g(x)=0的形式，那么我们可按如下方法求出f(x)的极值。

首先，构造Lagrange函数：

$$L(\mathbf{x}, \lambda)=f(\mathbf{x})+\lambda g(\mathbf{x})$$

其中，λ称为待定的*Lagrange*乘数(undetermined multiplier)

然后，对Lagrange乘数关于x求偏导数，并令其值等于零：

$$\frac{\partial L(x,\lambda)}{\partial x}=\frac{\partial f(x)}{\partial x}+\lambda\frac{\partial g(x)}{\partial x}=0$$

这样就把约束条件下的最优化问题转化为无约束条件的方程求解问题。

通过求解上面方程，就能得到λ的值及相应的极值点x0(通常情况下，$\lambda\partial g/\partial x$不等于0)。最后把x0代入到f(x)函数，就能得到约束条件下f(·)的极值。

五. Fisher准则函数的极值求解

$$\text{Fisher准则函数} J(W) = \frac{W^T S_b W}{W^T S_w W}$$

其中 S_w为类内离散度矩阵, S_b为类间离散度矩阵

采用Langrage乘数法对J(W)求极值，以得到最优的权向量W*。

对$J(W)$求极大值, 得 $W^* = S_w^{-1}(\overline{X}_1 - \overline{X}_2)$

Fisher准则函数的极值求解过程：

由于J(W)与W的函数关系较复杂，极值点不易求出，为此，令$W^T S_w W = C \neq 0$，并构造如下Lagrange函数：

$$L(W, \lambda) = W^T S_b W - \lambda(W^T S_w W - C)$$

其中，λ为待定的*Lagrange*乘数，对上式求关于W的偏导数：

$$\frac{\partial L(W, \lambda)}{\partial W} = 2(S_b W - \lambda S_w W)$$

令$\frac{\partial L(W, \lambda)}{\partial W} = 0$，极值点满足：

$$S_b W^* - \lambda S_w W^* = 0$$

$$S_b W^* - \lambda S_w W^* = 0$$

由于S_w可逆, 因此:

$$S_w^{-1} S_b W^* = \lambda W^*$$

即W^*是$S_w^{-1} S_b$的特征向量, 可以利用一般求特征向量的方法求解。

对于二类问题,Fisher利用S_b的性质实现了W^*求解,方法如下:

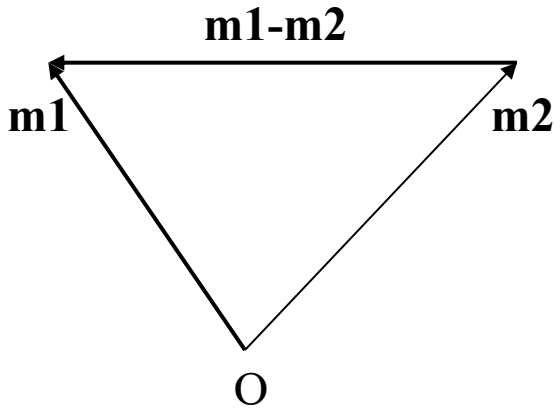
$$S_b W^* = (m1 - m2)(m1 - m2)^T W^*$$

$$= (m1 - m2)[(m1 - m2)^T W^*]$$

$$= (m1 - m2)R$$

其中,$m1, m2$分别为两类样本的均值向量;

$R = (m1 - m2)^T W^*$为一标量,因此$S_b W^*$总在$m1 - m2$方向上,即两类中心点的连线方向,见下图:



从而有:

$$\lambda \mathrm{W}^{*}=\mathrm{S}_{\mathrm{w}}^{-1}\mathrm{S}_{\mathrm{b}}\mathrm{W}^{*}=\mathrm{S}_{\mathrm{w}}^{-1}(m1-m2)R$$

求得:

$$\mathrm{W}^{*}=\frac{\mathrm{S}_{\mathrm{w}}^{-1}(m1-m2)R}{\lambda}$$

忽略常量因子$\frac{R}{\lambda}$，得:

$$\mathrm{W}^{*}=\mathrm{S}_{\mathrm{w}}^{-1}(m1-m2)=\mathrm{S}_{\mathrm{w}}^{-1}(\overline{X}_1-\overline{X}_2)$$

六. 多元函数的Taylor展开式

设有n维模式向量$\mathbf{x}=(x_1,x_2,\mathrm{K}\ ,x_n)^T$, $f(\mathbf{x})$是定义在$D\subset R^n$上的n元判别函数, 且$\mathbf{x}^*\in D$为给定模式样本点, $\mathbf{x}\in D$为任意模式样本点, 那么

1.若$f(\mathbf{x})$连续可微, 则存在$\theta\in(0,1)$, 使

$$f(\mathbf{x})=f(\mathbf{x})+\nabla f(\xi)^T(\mathbf{x}-\mathbf{x}^*)$$

其中$\xi=\mathbf{x}^*+\theta(\mathbf{x}-\mathbf{x}^*)$

2.若$f(\mathbf{x})$连续二阶可微, 则存在$\theta\in(0,1)$, 使

$$f(\mathbf{x})=f(\mathbf{x})+\nabla f(\mathbf{x}^*)^T(\mathbf{x}-\mathbf{x}^*)+\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\nabla^2 f(\xi)^T(\mathbf{x}-\mathbf{x}^*)$$

其中$\xi=\mathbf{x}^*+\theta(\mathbf{x}-\mathbf{x}^*)$

3.若$f(\mathbf{x})$连续二阶可微,则

$$f(\mathbf{x})=f(\mathbf{x}^*)+\nabla f(\mathbf{x}^*)^T(\mathbf{x}-\mathbf{x}^*)+$$

$$+\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\nabla^2 f(\mathbf{x}^*)^T(\mathbf{x}-\mathbf{x}^*)+o(\left\|\mathbf{x}-\mathbf{x}^*\right\|^2)$$