多元函数极值问题（矩阵求导）

1、多元函数导数的定义

2、最速下降方向

3、牛顿法

4、线性回归、岭回归、Logistic回归

1、基本概念

设$f:R^n \to R$为一个n元一阶可微函数$y=f(x_1,\dots,x_n)$，定义其梯度向量为

$$\nabla f(x)(\nabla_x f(\boldsymbol{x}))=\left(\partial_{x_1}f,\partial_{x_2}f,\dots,\partial_{x_n}f\right) \qquad 其中\boldsymbol{x}=(x_1,x_2,\dots,x_n)$$

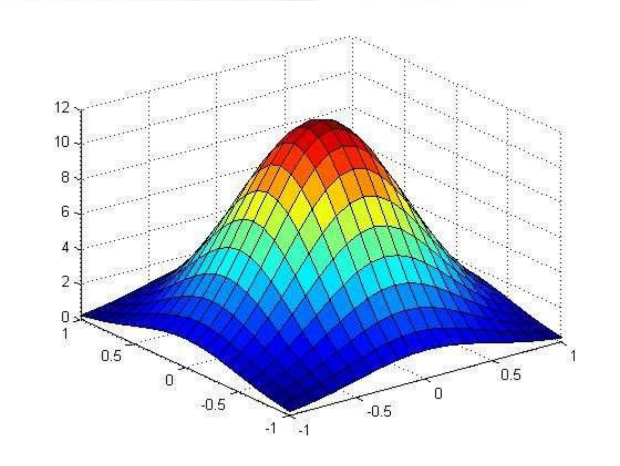
二阶导数（Hessian矩阵）

$$Hf=\left[a_{ij}\right]_{n\times n}=\begin{bmatrix} \frac{\partial^2 f}{\partial_{x_1}^2} & \cdots & \frac{\partial^2 f}{\partial_{x_1}\partial_{x_n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial_{x_n}\partial_{x_1}} & \cdots & \frac{\partial^2 f}{\partial_{x_n}^2} \end{bmatrix}_{n\times n}$$

其中$a_{ij}=\partial_{x_i x_j}f$

注：由于$\partial_{x_i x_j}f=\partial_{x_j x_i}f$，所以$a_{ij}=a_{ji}$，所以$Hf=Hf^T$，即$Hf$是实对称矩阵

最速下降法（梯度下降）



最速下降法（梯度下降）

方向导数： 设$\boldsymbol{u}$(单位向量)是R^n中的一个方向，n元函数$f(x_0)$沿$\boldsymbol{u}$方向的斜率，我们称之为$f(x_0)$在$\boldsymbol{u}$方向的方向导数，计算公式为：

$$\partial_{\boldsymbol{u}} f(\boldsymbol{x}_0)=\boldsymbol{u}\nabla_x f(\boldsymbol{x}_0)^T$$（向量u和梯度向量的内积）

我们希望找到使函数f下降最快的方向u，由内积公式得：

$$\boldsymbol{u}\cdot\nabla_x f(\boldsymbol{x}_0)=|\boldsymbol{u}||\nabla_x f(\boldsymbol{x}_0)|\cos\theta$$

$|\boldsymbol{u}|=1$，$|\nabla_x f(\boldsymbol{x}_0)|$与$u$无关。因此，当$\theta=0$时，方向导数大于0，且取最大值；当$\theta=\pi$时，方向导数小于0，且取最小值。

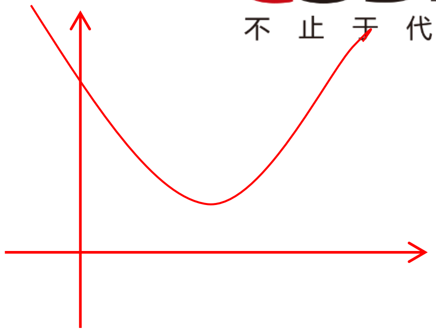
结论：对于函数$f(\boldsymbol{x})$中的任意一点$\boldsymbol{x} \in R^n$，沿着和梯度向量一样的方向时，函数递增最快；沿着和梯度向量相反方向时，函数递减最快。因此最速下降方向为$-\nabla f(\boldsymbol{x})$

最速下降建议点为

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \epsilon \nabla f(\boldsymbol{x}_t)$$

牛顿法

一元函数极值判别法：



导数分析法：若$f'(x_0)=0$，且$f''(x_0)>(<)0$，则$f(x)$在$x=x_0$处取极小（大）值

泰勒公式法：$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(\xi)(x-x_0)^2$

$$=f(x_0)+\frac{1}{2}f''(\xi)(x-x_0)^2 \qquad \xi\in(x,x_0)or(x_0,x)$$

当$f''(x_0)>0$时，由导数的保号性可知，在x_0很小的邻域内有$f''(\xi)>0$，因此

在x_0很小的邻域内有$f(x)>f(x_0)$，即函数$f(x)$在$x=x_0$处取极小值

多元函数的泰勒展开

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)\nabla^T f(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)Hf(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)^T \quad \boldsymbol{x}, \boldsymbol{x_0} \in R^n$$

如果$\boldsymbol{x}_0$点满足$\nabla f(\boldsymbol{x}_0) = \boldsymbol{\theta}$(零向量)，并且 $(\boldsymbol{x}-\boldsymbol{x_0})Hf(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x_0})^T > (<)0$。
那么，函数 $f(\boldsymbol{x})$在$\boldsymbol{x} = \boldsymbol{x}_0$ 处取极小(大)值。

正定（半正定）矩阵

A是一个n阶对称矩阵，即$A = A^T$（$a_{ij} = a_{ji}$）

设n维向量$\boldsymbol{x}$= $(x_1, x_2, \dots, x_n)$，定义二次型（多项式）

$$f(x_1, x_2, \dots, x_n) = \boldsymbol{x}A\boldsymbol{x}^T = \sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{i,j} a_{ij}x_ix_j$$

为A对应的二次型（多项式）。对称方阵A称为二次型对应的矩阵

定义1：若任意的$\boldsymbol{x} \neq \boldsymbol{\theta}$，都有$f(x_1, x_2, \dots, x_n) = \boldsymbol{x}A\boldsymbol{x}^T > (\geq) 0$，则称该二次型为正定（半正定）二次型，对于的矩阵A为正定（半正定）矩阵。

定义2：若任意的$\boldsymbol{x} \neq \boldsymbol{\theta}$，都有$f(x_1, x_2, \dots, x_n) = \boldsymbol{x}A\boldsymbol{x}^T < (\leq) 0$，则称该二次型为负定（半负定）二次型，对于的矩阵A为负定（半负定）矩阵。

结论：如果在$\boldsymbol{x} = \boldsymbol{x}_0$处，有$\nabla f(\boldsymbol{x}_0) = \boldsymbol{\theta}$(零向量)，我们称$\boldsymbol{x}_0$为$f(\boldsymbol{x})$的驻点。

如果$Hf(\boldsymbol{x}_0)$正定矩阵，$f(\boldsymbol{x})$在$\boldsymbol{x} = \boldsymbol{x}_0$处是一个局部极小值

如果$Hf(\boldsymbol{x}_0)$负定矩阵，$f(\boldsymbol{x})$在$\boldsymbol{x} = \boldsymbol{x}_0$处是一个局部极大值

如果$Hf(\boldsymbol{x}_0)$不定矩阵，$f(\boldsymbol{x})$在$\boldsymbol{x} = \boldsymbol{x}_0$处不取极值

二元函数求极值：

step1、 $\begin{cases} f'_x(x,y) = 0 \\ f'_y(x,y) = 0 \end{cases} => \begin{cases} x = x_0 \\ y = y_0 \end{cases}$

Step2、 $A = f''_{xx}(x_0, y_0)$ $B= f''_{xy}(x_0, y_0)$ $C= f''_{yy}(x_0, y_0)$

Step3、 当$B^2 - AC < 0$且$A < (>)0$，$f(x_0, y_0)$是极大(小)值

当$B^2 - AC > 0$时，$f(x_0, y_0)$不取极值

当$B^2 - AC = 0$时，无法判断

正定矩阵的判断法：

引理：对称方阵一定可正交对角对角化，即

任意的对称矩阵A，必然存在一个正交矩阵Q，使得

$$Q^TAQ=\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \dots & 0\\ \vdots & \ddots & \ddots & 0\\ 0 & \dots & 0 & \lambda_n\end{bmatrix} \qquad A=Q\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \dots & 0\\ \vdots & \ddots & \ddots & 0\\ 0 & \dots & 0 & \lambda_n\end{bmatrix}Q^T$$

$$f(x_1,x_2,\dots,x_n)=\boldsymbol{x}A\boldsymbol{x}^T=\boldsymbol{x}\left(Q\,diag(\lambda_1,\lambda_2,\dots,\lambda_n)Q^T\right)\boldsymbol{x}^T$$

$$=(\boldsymbol{x}Q)diag(\lambda_1,\lambda_2,\dots,\lambda_n)(\boldsymbol{x}Q)^T$$

记$\boldsymbol{x}Q=\boldsymbol{y}=(y_1,y_2,\dots,y_n)$

$$f(x_1,x_2,\dots,x_n)=\boldsymbol{y}\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \dots & 0\\ \vdots & \ddots & \ddots & 0\\ 0 & \dots & 0 & \lambda_n\end{bmatrix}\boldsymbol{y}^T=\lambda_1 y_1^2+\lambda_2 y_2^2+\cdots+\lambda_n y_n^2$$

定理1：二次型$f(x_1,x_2,\dots,x_n)=\boldsymbol{x}A\boldsymbol{x}^T$正定（半正定）$\Leftrightarrow A$的每个特征值$\lambda_i>(\geq)0$

定理 2：对称矩阵正定的充分必要条件是的各阶顺序主子式为正数，即

$$a_{11} > 0, \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \cdots \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} > 0.$$

$$A=\begin{pmatrix}6 & 2 & -4\\ 2 & 1 & -2\\ -4 & -2 & 5\end{pmatrix},$$ 各阶顺序主子式为

$$|6|=6>0,\quad \begin{vmatrix}6 & 2\\ 2 & 1\end{vmatrix}=2>0,\quad \begin{vmatrix}6 & 2 & -4\\ 2 & 1 & -2\\ -4 & -2 & 5\end{vmatrix}=2>0,$$

所以 f 是正定二次型.

▶ 最小二乘法（线性回归）

一维模型： 观测到样本点(x_i, y_i) $(i = 1,2,\dots,m)$，试图找到合适的a, b，使得

$$f(x_i) = ax_i + b\text{，并且}f(x_i) \approx y_i$$

$$(a^*, b^*) = \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \quad \left(\underset{(a,b)}{\operatorname{argmin}} \|f(\boldsymbol{x}) - \boldsymbol{y}\|_{L^2} \right)$$

$$= \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (ax_i + b - y_i)^2$$

令$E(a,b)=\sum_{i=1}^{m}(ax_i+b-y_i)^2$

$$\nabla E(a,b)=\left(\frac{\partial E(a,b)}{\partial a}, \frac{\partial E(a,b)}{\partial b}\right)=(0,0)$$

$$a=\frac{\sum_{i=1}^{m} y_i(x_i-\bar{x})}{\sum_{i=1}^{m} x_i^2-\frac{1}{n}\left(\sum_{i=1}^{m} x_i\right)^2} \qquad b=\frac{1}{n}\sum_{i=1}^{m}(y_i-ax_i)$$

其中$\bar{x}=\frac{1}{m}\sum_{i=1}^{m} x_i$

高维模型： 观测到样本点$(\boldsymbol{x}_i, y_i)$ $(i=1,2,\dots,m)$。其中，

$\boldsymbol{x_i} \in R^n, y_i \in R$ 试图找到合适$\boldsymbol{a} \in R^n, b \in R$，使得

$f(x_i) = \boldsymbol{a}x_i + b$，并且$f(x_i) \approx y_i$

注：$f(x_i) = ax_i + b = a_1x_{i1} + a_2x_{i2} + \cdots + a_nx_{in} + b$

令$A=\begin{bmatrix} x_{11} & \dots & x_{1n} & 1 \\ x_{21} & \dots & x_{2n} & 1 \\ \dots & \dots & \dots & \vdots \\ x_{m1} & \dots & x_{mn} & 1 \end{bmatrix}$ $\boldsymbol{\omega} = (\boldsymbol{a}, b)^T \in R^{n+1}$ $\boldsymbol{y} = (y_1, y_2, \dots, y_m)^T$

$\sum_{i=1}^{m}(f(x_i) - y_i)^2 = \|A\boldsymbol{\omega} - \boldsymbol{y}\|_2^2 = (\boldsymbol{y} - A\boldsymbol{\omega})^T(\boldsymbol{y} - A\boldsymbol{\omega})$

$$\boldsymbol{\omega}^* = \underset{\omega}{\operatorname{argmin}}(\boldsymbol{y} - A\boldsymbol{\omega})^T(\boldsymbol{y} - A\boldsymbol{\omega})$$

令$E(\boldsymbol{\omega}) = (\boldsymbol{y} - A\boldsymbol{\omega})^T(\boldsymbol{y} - A\boldsymbol{\omega})$

$$\nabla E(\boldsymbol{\omega}) = 2A^T(A\boldsymbol{\omega} - \boldsymbol{y})=\boldsymbol{\theta}$$

$$A^TA\boldsymbol{\omega} - A^T\boldsymbol{y}=\boldsymbol{\theta}$$

$$\boldsymbol{\omega} = (A^TA)^{-1}A^T\boldsymbol{y}$$

用SVD处理岭回归：

如果矩阵A^TA不可逆，则用岭回归代替线性回归。

线性回归： $\hat{\boldsymbol{y}} = A\boldsymbol{\omega} = A(A^TA)^{-1}A^T\boldsymbol{y}$

岭回归： $\hat{\boldsymbol{y}} = A(A^TA + \lambda I)^{-1}A^T\boldsymbol{y}$

性质：损失无偏性，增加稳定性，从而得到较高的计算精度。

对矩阵A进行SVD，$A = U\Lambda V^T$　　其中U,V是正交矩阵

岭回归：　$\hat{y} = A(A^T A + \lambda I_n)^{-1} A^T \boldsymbol{y}$

$= U\Lambda V^T (V\Lambda^T U^T U\Lambda V^T + \lambda V I_n V^T)^{-1} V\Lambda^T U^T \boldsymbol{y}$

$= U\Lambda(\Lambda^T \Lambda + \lambda I_n)^{-1} \Lambda^T U^T \boldsymbol{y}$

$= \sum_{i=1}^{r} U_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} U_i \boldsymbol{y}$

Logistic回归:

模型：$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$，考虑变量的线性组合函数$g(\boldsymbol{x}) = \omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n$（LR分类器的权重）。概率模型满足sigmoid函数

$$P(y=1|\boldsymbol{x}) = f(x) = \frac{1}{1+e^{-g(x)}}$$

$$P(y=0|\boldsymbol{x}) = f(x) = 1 - \frac{1}{1+e^{-g(x)}}$$

观测到m个样本数据$(\boldsymbol{x}^1, y_1), (\boldsymbol{x}^2, y_2), \dots, (\boldsymbol{x}^m, y_m)$，极大似然函数为：

$$L(\boldsymbol{\omega}) = \prod_{i=1}^{m} \left(f(\boldsymbol{x}^i)\right)^{y_i} \left(1 - f(\boldsymbol{x}^i)\right)^{1-y_i}$$

左右取对数得：

$$lnL(\omega)=\sum_{i=1}^{m}y_i ln\left(f\left(\boldsymbol{x}^i\right)\right)+(1-y_i)ln\left(1-f\left(\boldsymbol{x}^i\right)\right)$$

计算梯度，并令$\nabla lnL(\boldsymbol{\omega})$=$\theta$。得到方程组

$$\frac{\partial lnL(\boldsymbol{\omega})}{\partial \omega_k}=\sum_{i=1}^{m}x_k^i\left[y_i-f\left(\boldsymbol{x}^i\right)\right]=0$$

CSDN
不 止 于 代 码

THANK YOU

AI100