# 最优化：最小二乘、梯度下降、牛顿法及其它

BY JIN

2018-09-11

## 1 最小二乘

欲拟合线性方程$y=f_\beta(x)=\beta^Tx$，现已获得N组采样值$X=[x_1,x_2,...,x_n],Y=[y_1,y_2,...,y_n]$，以最小化误差方差为目标：

$$\beta=\text{Argmin}_\beta\sum\|f_\beta(x)-y\|^2=\text{Argmin}_\beta(\beta^TX-Y)(\beta^TX-Y)^T=\text{Argmin}\,D(\beta)$$

对$D(\beta)$求极值：

$$\frac{\partial D}{\partial\beta}=2(\beta^TX-Y)X^T=0$$

$$\beta=(XX^T)^{-1}XY^T$$

## 2 梯度下降

若$f(x)$非线性，则$D(\beta)$非$\beta$的二次形式，无法用最小二乘方法求出$\text{Argmin}_\beta D(\beta)$的解析解，考虑用迭代方法求数值解。由于某点的梯度方向$\nabla D(\beta)=\frac{\partial D}{\partial\beta}$[1]代表该点上变化最剧烈的方向，沿其反方向前进一小段步长$\alpha$应最能接近极小值：

$$\beta\leftarrow\beta_0-\alpha\nabla D^T(\beta_0)$$

## 3 牛顿法

用牛顿法求$D'(\beta)=0$的根，即为$D(\beta)$的极值点：

$$0=D'(\beta)=D'(\beta_0)+D''(\beta_0)(\beta-\beta_0)+O((\beta-\beta_0)^2)$$

$$\beta\leftarrow\beta_0-(HD(\beta_0))^{-1}\nabla D^T(\beta_0)$$

其中梯度$\nabla D^T(\beta)=\frac{\partial D}{\partial\beta^T}$即为$D$的一阶导，Hessian矩阵$HD(\beta)=\frac{\partial}{\partial\beta^T}\left(\frac{\partial D}{\partial\beta}\right)=J_{\nabla D}(\beta^T)$即为$D$的二阶导。

应用中也可为迭代公式增加一松弛系数$\alpha\in(0,1]$，变为：

$$\beta\leftarrow\beta_0-\alpha\,(HD(\beta_0))^{-1}\nabla D^T(\beta_0)$$

## 4 高斯-牛顿法及其它

一般而言牛顿法收敛快于梯度下降，但当待优化系数$\beta_{m\times1}$项数较多时，每步均需计算一个$m\times m$大小的海森矩阵（二阶导），导致更大的计算开销。因此一个优化方向便是寻找较简单的函数$B(\beta)\approx HD(\beta)$作为近似，以减小每步计算量。

1. 文中向量、矩阵导数，全部采用Numerator Layout，因此$\nabla D(\beta)=\frac{\partial D}{\partial\beta}$是一个行向量。

当待优化函数$D=\sum r^2(\beta)=R(\beta)R^T(\beta)$ 具有二次求和形式时，$\nabla D(\beta)=2J_R(\beta^T)R^T(\beta)$，以$B(\beta)=2\,J_R(\beta^T)J_R^T(\beta^T)$作为$HD(\beta)$的近似，有

$$\beta\leftarrow\beta_0-(J_R(\beta^T)J_R^T(\beta^T))^{-1}J_R(\beta^T)R^T(\beta)$$

此即为Gauss-Newton法($J_R(\beta^T)$为$R$对$\beta^T$的导数即雅可比矩阵)。

此外还有Quasi-Newton等多种近似算法。