线性回归：梯度下降法的数学原理

上期和大家分享了线性回归的基本概念，其要点如下：

- 线性回归假设输出变量是若干输入变量的线性组合，并根据这一关系求解线性组合中的最优系数；
- 多元线性回归问题也可以用最小二乘法求解，但极易出现过拟合现象；
- 岭回归和 LASSO回归分别通过引入$L2$范数惩罚项和$L2$范数惩罚项抑制过拟合。

上期补充

范数是对单个向量大小的度量，描述的是向量自身的性质，其作用是将向量映射为一个非负的数值。通用的L^p范数定义如下：

$$|x|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$$

对于一个给定的向量，L^1范数计算的是向量所有元素绝对值的和，L^2范数计算的是通常意义上的向量长度，L^∞范数计算的则是向量中最大元素的取值。

证明：**如果实值函数F(x)在点a处可微且有定义，那么函数F(x)在点a沿着梯度相反的-∇F(a)下降最快。**

提到梯度，就必须从导数、偏导数和方向导数讲起，弄清楚这些概念，才能够正确理解为什么在优化问题中能够使用梯度下降法来优化目标函数。

在这里先简要介绍一下导数和偏导数。

在微积分中，导数反映的是函数$y = f(x)$在某一点处沿x轴正方向的变化率。而偏导数与导数在本质上是一致的，都是当自变量的变化量趋于0时，函数值的变化量与自变量变化量比值的极限。直观地说，偏导数也就是函数在某一点上沿坐标轴正方向的的变化率。

二者区别主要在于：

- **导数**，指的是一元函数中，函数$y = f(x)$在某一点处沿x轴正方向的变化率；
- **偏导数**，指的是多元函数中，函数$y = f(x_1, x_2, \ldots, x_n)$在某一点处沿某一坐标轴$(x_1, x_2, \ldots, x_n)$正方向的变化率。

简要介绍一下导数和偏导数之后，我们主要介绍一下方向导数和梯度，包含完整的推导公式。

现在我们先来讨论函数$z = f(x, y)$在一点P沿某一方向的变化率问题。

为了解决这个问题，我们得引入如下**定义**：

设函数$z = f(x, y)$在点$P(x, y)$在某一领域 $U(p)$内有定义,从点P引一条射线ll ,设x轴正向到射线l的转角为φ，并设$P'(x + \Delta x, y + \Delta y)$为$l$上的另一点且$P' \in U(p)$ 。我们考虑函数的增量$f(x + \Delta x, y + \Delta y) - f(x, y)$与$)P, P'$两点间距$\rho = \sqrt{(\Delta x)^2 + (\Delta y)^2}$的比值，当$P'$沿着$l$趋于$P$时，如果这个比的极限存在，则称这极限为函数$f(x, y)$在点$P$沿方向$l$的**方向导数**，记做$\frac{\partial f}{\partial l}$，即：

$$\frac{\partial f}{\partial l}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}$$

从定义可知，当函数$f(x,y)$在点$P(x,y)$的偏导数f_x、f_y存在时，函数在点P沿着x轴正向$e_1=(1,0)$，y轴正向$e_2=(0,1)$的方向导数存在且其值依次为f_x、f_y，函数在点沿x轴负向$e_1'=(-1,0)$，y轴负向$e_2'=(0,-1)$的方向导数也存在且其值依次为$-f_x$、$-f_y$。

关于方向导数$\frac{\partial f}{\partial l}$的存在及计算，我们有如下**定理**：

如果$z=f(x,y)$在点$P(x,y)$是可微的，那么函数在该点沿任一反向的方向导数都存在，且有

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}cos\varphi+\frac{\partial f}{\partial y}sin\varphi$$

其中φ为x轴到方向l的转角。

证：根据函数$z=f(x,y)$在点$P(x,y)$可微分的假定，函数的增量可以表达为：

$$f(x+\Delta x,y+\Delta y)-f(x,y)=\frac{\partial f}{\partial x}\Delta x+\frac{\partial f}{\partial y}\Delta y+o(\rho)$$

两边各除以ρ，得到

$$\begin{aligned}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}&=\frac{\partial f}{\partial x}\frac{\Delta x}{\rho}+\frac{\partial f}{\partial y}\frac{\Delta y}{\rho}+\frac{o(\rho)}{\rho}\\&=\frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi+\frac{o(\rho)}{\rho}\end{aligned}$$

根据

$$\frac{\partial f}{\partial l}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}$$

我们就可以证明方向导数存在且其值为

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi$$

对于三元函数$u=f(x,y,z)$来说，它在空间一点$P(x,y,z)$沿着方向l(设方向的方向角为(α,β,γ)的方向导数，同样可以定义为

$$\frac{\partial f}{\partial l}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y,z+\Delta z)-f(x,y,z)}{\rho}$$

其中$\rho=\sqrt{(\Delta x)^2+(\Delta y)^2+(\Delta z)^2}$，$\Delta x=\rho\,cos\alpha,\Delta y=\rho\,cos\beta,\Delta z=\rho\,cos\gamma$。

同样可以证明，如果函数在所考虑的点处可微分，那么函数在该点沿着l方向的方向导数为:

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}\cos\alpha+\frac{\partial f}{\partial y}\sin\beta+\frac{\partial f}{\partial z}\cos\gamma$$

同样可以扩展到n元函数$u=f(x_1,x_2,\dots,x_n)$中，这里就不一一陈诉了。

与方向导数有关联的一个概念是函数的**梯度**。其定义为：

设函数$z=f(x,y)$在平面区域D内具有一阶连续偏导数，则对于每一点$(x,y)\in D$，都可定义出一个向量$\frac{\partial f}{\partial x}i+\frac{\partial f}{\partial y}j$，这向量称为函数$z=f(x,y)$在点$P(x,y)$的梯度，记作$gradf(x,y)$,即

$$gradf(x,y)=\frac{\partial f}{\partial x}i+\frac{\partial f}{\partial y}j$$

如果设$e=cos\varphi\,i+sin\varphi\,j$是与方向$l$同方向的单位向量，则由方向导数的计算公式可知:

$$\begin{aligned}\frac{\partial f}{\partial l} &= \frac{\partial f}{\partial x}cos\varphi + \frac{\partial f}{\partial y}sin\varphi \\ &= \left\{\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right\} * \{cos\varphi, sin\varphi\} \\ &= grad\ f(x,y) * e \\ &= |grad\ f(x,y)| * cos(grad\ f(x,y), e)\end{aligned}$$

其中$(grad f(x,y), e)$表示向量$grad\ f(x,y)$与e的夹角。

由此可以看出，方向导数就是梯度在射线上的投影，当方向l与梯度的方向一致时，有

$$cos\ (grad\ f(x,y), e) = 1$$

从而有$\frac{\partial f}{\partial l}$最大值。沿梯度方向的方向导数达到最大值，也就是说梯度的方向是函数$f(x,y)$在这点增长最快的方向。因此，我们可以得到如下结论：

函数在某点的梯度是这样一个向量，它的方向与取得最大方向导数的方向一致，而它的模为方向导数的最大值。