前言

       欢迎来到“每天进步一点点 2015”微信公众号，从本期开始，我们将给各位网友分享有关数据挖掘的理论与实战知识，实战部分将结合 Python 和 R 语言完成模型的落地。在这一期，我们将从统计模型中的回归开始入手，回归堪称是经典中的经典，很多现实问题都可以通过回归思想来解决。

       从有监督、无监督和半监督的角度来看，回归其实是有监督的算法模型之一，反映的是根据某些已知变量(解释变量或自变量)去预测某个未知变量(被解释变量或因变量)。例如根据国民生产总值，预测人口的失业率；根据房屋的面积，朝向，交通状况等信息，预测房价；根据田地的面积，施肥状况，稻谷的品种，预测粮食产量等。你会发现，对于类似数值型的因变量预测，我们就可以借助于回归来完成。关于回归有很多种类，如多元线性回归、岭回归、Lasso 回归等，本期开始，我们就介绍多元线性回归，后面也会分享到岭回归和 Lasso 回归。

多元线性回归

       对多元线性回归模型有所了解的朋友，都知道因变量 y 与自变量 X 之间存在某种线性组合。 例如，现在手上有 n 个观测，p+1 个变量，其中 p 个变量是自变量，1 个变量是因变量，即如下方所示：

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

       如上面所说的某种线性组合，指的是因变量 y 应该可以用自变量 X 来表示，并且它们之间是存在线性关系，即：

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_p x_p+\varepsilon$$

其中，$\beta=\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$，它们分别代表的是多元线性回归模型的偏回归系数和误差项。

为了书写的方便，可以将回归模型的方程式写成 $y=X\beta+\varepsilon$ 其中，$\varepsilon \sim N(0,\sigma^2)$ 。

极大似然估计

       既然我们知道了多元线性回归模型中 y 与 X 的组合关系，那接下来关心的就是如何求出模型的偏回归系数。由于误差项是服从正态分布的，而误差项又是关于偏回归系数的表达式，即 $\varepsilon=y-X\beta$ 。

       首先来看一下正态分布的概率密度函数：

$$\mathrm{f(x)}=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

其中，μ为x的均值，σ为x的标准差。

其次，根据该密度函数，可以将误差项的概率密度函数表示为：

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(\varepsilon)^2}{2\sigma^2}}$$

最后，我们可以这样理解，如果已知了 X 的观测和偏回归系数的值，那么就可以求得 y 值的概率值，即：

$$f(y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y-X\beta)^2}{2\sigma^2}}$$

       上面的理解只是由结果往前推断，但现在的问题是，不知道偏回归系数。上式反应的是计算 y 的条件概率，如果概率值越大，则说明预测出来的 y 会越接近于真实的 y，所以，现在的问题就变成了计算概率的最大值。根据，观测之间的 y 是独立的假设，我们可以对其构造极大似然函数，即：

$$L(\varepsilon) = \prod \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y-X\beta)^2}{2\sigma^2}}$$

为了求解的方便，我们在等式两边取对数：

$$\begin{aligned}
l(\varepsilon) &= \log(L(\varepsilon)) \\
&= \log\left(\prod \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y-X\beta)^2}{2\sigma^2}}\right) \\
&= \sum \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y-X\beta)^2}{2\sigma^2}}\right) \\
&= \sum \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) + \log\left(e^{\frac{-(y-X\beta)^2}{2\sigma^2}}\right) \\
&= n\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) + \sum \frac{-(y-X\beta)^2}{2\sigma^2} \\
&= n\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum \frac{(y-X\beta)^2}{2\sigma^2}
\end{aligned}$$

由于等式右边的前半部分 $n\log(\frac{1}{\sqrt{2\pi\sigma}})$ 是一个常数，而后半部分是一个负值。所以求解似然函数的极大值问题就转换成了求 $\sum \frac{(y-X\beta)^2}{2\sigma^2}$ 的最小值，即

$$\underset{\arg\min}{J(\beta)} = \frac{1}{2}\sum (y-X\beta)^2 = \frac{1}{2}\sum \varepsilon^2$$

最小二乘法

       根据上面的极大似然函数的推导可知，要实现最优

化问题的解决，就是求解误差平方和最小。这也很容易理解，即要想求得合理的偏回归系数，得到回归模型，就要保证该模型尽可能的拟合好真实的数据，而是否很好的拟合，不就是用误差来度量吗？误差越小，则预测的越接近于现实，否则就越偏离现实。接下来，我们就借助于最小二乘法的思想再来推导如何求得偏回归系数。

         在推导之前，需要了解一些基本的线性代数知识，具体在下面给出：

- 向量的平方和

$$x=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$\begin{aligned}\sum_{i=1}^{m} x_i^2 &= x_1^2 + x_2^2 + \cdots + x_m^2 \\ &= x'x\end{aligned}$$

- 矩阵乘法的转置

$$(AB)' = B'A'$$

- 矩阵的偏导数

$$\begin{cases} \dfrac{\partial A\theta}{\partial \theta} = A' \\ \dfrac{\partial A\theta'}{\partial \theta} = A \end{cases}$$

求解偏回归系数的推导

$$\begin{aligned} J(\beta) &= \frac{1}{2}\sum (y - X\beta)^2 = \frac{1}{2}\sum \varepsilon^2 \\ &= \frac{1}{2}(y - X\beta)'(y - X\beta) \\ &= \frac{1}{2}(y' - \beta' X')(y - X\beta) \\ &= \frac{1}{2}(y'y - y'X\beta - \beta' X'y + \beta' X'X\beta) \end{aligned}$$

要想求得上面目标函数的最小值，可以通过求偏导数，然后使偏导数为 0 即可：

$$\begin{aligned} \underset{\arg\min}{J(\beta)} &= \frac{\partial J(\beta)}{\partial \beta} \\ &= \frac{1}{2}(0 - X'y - X'y + 2X'X\beta) = 0 \end{aligned}$$

$$\begin{aligned} 2X'X\beta &= 2X'y \\ \beta &= (X'X)^{-1}X'y \end{aligned}$$

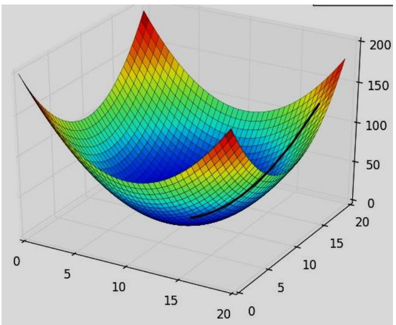
梯度下降法

       根据上面的推导就可以得到多元线性回归模型的偏回归系数了。如果你也一步步的推导一遍，我相信对你理解多元线性回归模型是有一定的帮助的。但是，上面的普通最小二乘有一个小小的瑕疵（这个瑕疵发生的概率还是非常小的），并不能确保方阵 X′X 是可逆的，即 X′X 的行列式一定不为 0，如果自变量之间存在高度共线性的话，那就会导致 X′X 是不可逆。这里，我们再分享一种利用“梯度下降”的方法实现偏回归参数的求解，该方法就可以很好的避免上面的瑕疵。

       我们知道，目标函数 $\underset{\arg\min}{J(\beta)}=\frac{1}{2}\sum(y-\beta X)^2$ 是关于偏回归系数的二次函数，且开口向上，即凸函数，那这样的目标函数就会存在极小值。所以，我们就可以对每个偏回归系数求偏导数，而偏导数据梯度的概念：

$$\begin{aligned}\frac{\partial J(\beta)}{\partial\beta_i}&=\frac{\partial}{\partial\beta_i}\frac{1}{2}\sum(y-\beta X)^2\\&=2\times\frac{1}{2}\sum(y-\beta X)\times\frac{\partial}{\partial\beta_i}(y-\beta X)\\&=\sum(y-\beta X)\times X_i\end{aligned}$$

       那梯度下降中的“下降”是什么意思呢？其实就是指迭代，每迭代一次，就是一次下降的过程，这个过程，就是为了找到目标函数的极小值，如下面的形象图示：



这种下降的迭代，可以用下面的公式表示：

$$\begin{aligned}\beta_i&:=\beta_i-\alpha\frac{\partial J(\beta)}{\partial\beta_i}\\&:=\beta_i-\alpha(\sum(y-\beta X)\times X_i)\end{aligned}$$

其中，α为学习率，即迭代的步长。

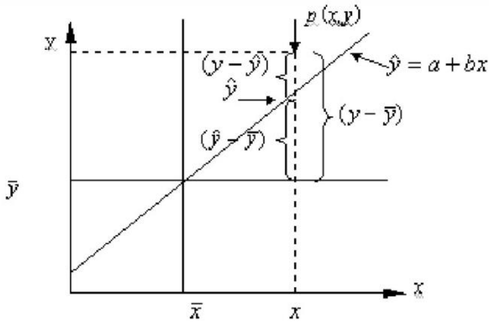
注意，这里的步长既不能太小，也不能太大，如果太小的话，会导致迭代次数暴增，降低算法的运行效率，加大运行的时间成本和运行空间；反之容易跨过极小值，无法达到全局最优。

模型的显著性检验

       关于“梯度下降法”的介绍属于线性回归中的知识的扩展，我们还是把重点回归到最小二乘法。通过最小二乘法我们可以得到模型的偏回归系数，但计算得到系数就一定能够保证模型是 OK 的吗？这里还需要对模型的显著性进行必要的检验，而模型显著性检验的假设条件为：

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_p = 0$$
$$H_1: \beta_0, \beta_1, \cdots, \beta_p \text{不全为}0$$ 。

我们先从下面这张图来理解一下几个离差平方和的概念：



$$\begin{cases} \sum_{i=1}^{n}(y_i-\overline{y})^2 = TSS \ -> \text{总的离差平方和} \\ \sum_{i=1}^{n}(\hat{y}-\overline{y})^2 = RSS -> \text{回归离差平方和} \\ \sum_{i=1}^{n}(y_i-\hat{y})^2 = ESS -> \text{误差平方和} \end{cases}$$

上面的三种离差平方和，存在这样的等式关系：TSS=RSS+ESS。实际上，TSS 是固定，而 ESS 和 RSS 是跟模型的预测值有关的，如果模型拟合的越好，则误差平方和(ESS)应该越小，对应的 RSS 越大。所以，根据这两个离差平方和就可以构造模型检验的统计量 F：

$$F = \frac{RSS / p}{ESS / (n-p-1)} \sim F(p, n-p-1)$$

其中，p 和 n-p-1 为 RSS 和 ESS 的自由度。当我需要检验模型是否 OK 的时候，只需要将计算出来的统计量 F 与理论的 $F(p, n-p-1)$ 值作对比，如果统计量 F 值大于理论的临界值，则认为可以拒绝原假设 H_0，即接受备注假设 H_1。

参数的显著性检验

        上面是针对模型的显著性检验作了相关的理论说明，但模型 OK(所有的偏回归系数不全为 0)，并不代表每一个自变量对因变量都是重要的，即每一个偏回归系数都是 OK(所有的偏回归系数都不为 0)的，所以，我们还需要对模型的每个偏回归系数进行显著性检验。在检验之前，我们需要先了解一下偏回归系数的期望和方差：

$$E(\hat{\beta}) = \beta$$
$$D(\hat{\beta}) = \alpha^2 (X'X)^{-1}$$

公式推导如下：

$$\hat{\beta}=(X'X)^{-1}X'y$$

$$\begin{aligned}E[\hat{\beta}]&=E[(X'X)^{-1}X'y]\\&=E[(X'X)^{-1}X'(X\beta-\varepsilon)]\\&=E[(X'X)^{-1}X'X\beta-\varepsilon(X'X)^{-1}X']\\&=E[\beta]-E[\varepsilon](X'X)^{-1}X'\\&=E[\beta]=\beta\end{aligned}$$

$$\begin{aligned}Var(\hat{\beta})&=E\hat{\beta}^2-(E\hat{\beta})^2\\&=E[((X'X)^{-1}X'y)^2]-b^2\\&=E[((X'X)^{-1}X'(X\beta-\varepsilon))^2]-b^2\\&=E[((X'X)^{-1}X'X\beta-\varepsilon(X'X)^{-1}X')^2]-b^2\\&=E[(\beta-\varepsilon(X'X)^{-1}X')^2]-b^2\\&=E[\beta^2+\varepsilon^2(X'X)^{-1}X'X(X'X)^{-1}-2\beta\varepsilon(X'X)^{-1}X']-b^2\\&=E[\beta^2]+E[\varepsilon^2](X'X)^{-1}-2\beta E[\varepsilon](X'X)^{-1}X'-b^2\\&=b^2+\sigma^2(X'X)^{-1}-0-b^2\\&=\sigma^2(X'X)^{-1}\end{aligned}$$

       既然有了偏回归系数的期望和方差，我们就可以根据标准正态分布来构造 t 分布了(之所以是 t 分布，是因为总体方差未知)。如果变量 x 服从正态分布，则可以通过下面的方式将其转换为标准正态分布：

$$x\sim N(\mu,\sigma^2)$$

$$\frac{x-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

当总体方差未知的时候，则使用样本方差来代替，但要付出一些代价，不再是标准正太分布，而是自由度为 n-1 的 t 分布：

$$\frac{x-\mu}{s/\sqrt{n}}\sim t(n-1)$$

       接下来就是要进行参数的显著性检验了，其检验的假设条件为：

$$H_0:\beta_j=0,\quad j=0,1,2,\cdots p$$

$$H_1:\beta_j\neq 0$$

构造检验偏回归系数的 t 统计量

$$t=\frac{\hat{\beta}_j-\beta_j}{se(\hat{\beta}_j)}\sim t(n-p-1)$$

其中，$\hat{\beta}_j$是偏回归系数β_j的估计

$se(\hat{\beta}_j)$是偏回归系数β_j标准误

$$se(\hat{\beta}_j)=\sqrt{\frac{\varepsilon^2/(n-p-1)}{(n-1)Cov(X,X)}}$$

       最终，通过计算统计量 t 的值与理论的 t(n-p-1)值作对比，如果统计量 t 值大于理论的临界值，则认为可以拒绝原假设 H_0，否则就得接受原假设。