# 线性回归 Linear regression

## 概述

线性回归（Linear regression）是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，大于一个自变量情况的叫做多元回归。线性回归把焦点放在给定X值的y的条件概率分布，而不是X和Y的联合概率分布。

## 理论模型

给一个随机样本 $Y_i, X_{i1}, ..., X_{ip}$，一个线性回归模型假设响应变量$Y_i$和自变量$X_{i1}, ..., X_{ip}$ 之间的关系是除了X的影响以外，还有其他的变数存在。我们加入一个误差项 $\varepsilon_i$（也是一个随机变量）来捕获除了自变量以外的影响，所以一个多变量线性回归模型表示为以下的形式：

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i$$

如果向X添加一维$X_{i0}$，使其恒为1，令$\theta = (\beta_0, \beta^T)^T$，则模型可以简述为：

$$Y_i = X_i^T \theta + \varepsilon_i$$

其中 $X_i = (1, X_{i1}, ..., X_{ip})^T, \theta = (\beta_0, \beta_1, ..., \beta_p)^T$
将$Y_i, X_i^T, \varepsilon_i$ 堆叠写成矩阵的形式：

$$Y = X\theta + \varepsilon$$

其中

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}, X = \begin{bmatrix} X_1^T \\ X_2^T \\ ... \\ X_n^T \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1p} \\ 1 & X_{21} & X_{22} & ... & X_{2p} \\ ... & ... & ... & ... & ... \\ 1 & X_{n1} & X_{n2} & ... & X_{np} \end{bmatrix}, \theta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

线性模型的线性性体现在$Y$与参数$\theta$，所以即使令$X_{i2} = X_{i1}^2$，模型仍然是线性的。

## 模型假设

1.样本是从群体中随机抽取的
2.因变量Y在是实值连续的
3.残差项是独立同分布的，这里假设服从高斯分布
在这些假设下，建立一个条件预期模型的简单线性回归：

$$E(Y|X=x)=x^T\theta$$

# 模型求解

定义cost function：

$$J(\theta;X=x,Y=y)=\sum_{i=1}^{n}(x_i^T\theta-y_i)^2$$

写成矩阵的形式：

$$J(\theta;X,Y)=||X\theta-Y||_2^2$$

求函数的最小值，只要令

$$\begin{aligned}0 &= \nabla_\theta J(\theta;x,y)=\nabla_\theta(X\theta-Y)^T(X\theta-Y)\\ &= \nabla_\theta(\theta^TX^TX\theta-\theta^TX^TY-Y^TX\theta+Y^TY)\\ &= \nabla_\theta tr(\theta^TX^TX\theta-\theta^TX^TY-Y^TX\theta+Y^TY)\\ &= \nabla_\theta(tr\theta^TX^TX\theta-2tr\theta^TX^TY)\\ &= (2X^TX\theta-2X^TY)\end{aligned}$$

即有 $\theta=(X^TX)^{-1}X^TY$

# 从极大似然的角度解释Cost funtion

由前面的假设，$\varepsilon_i$服从高斯分布，那么

$$p(\varepsilon_i)\propto exp(-\varepsilon_i^2/2\sigma^2)$$

也就是说 $Y\sim N(x^T\theta,\sigma^2)$

$$p(y_i|x_i,\theta)\propto exp[-(y_i-x_i^T\theta)^2/2\sigma^2]$$

似然函数为

$$L(\theta)=\prod_{i=1}^{n}p(y_i|x_i,\theta)\propto\prod_{i=1}^{n}exp[-(y_i-x_i^T\theta)^2/2\sigma^2]$$

对数似然函数

$$l(\theta) = c_1 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2$$

$c_1$ 是正则化常数的对数，最大化$l(\theta)$就相当于最小化 $y - x^T\theta$，所以用前面的最小二乘就可以求解了，同样的，由于对数似然函数是可导的，基于梯度的方法如SGD也同样可以用于求解线性回归模型，$l(\theta)$梯度为

$$\frac{\partial l(\theta)}{\theta_j} = (y - x^T\theta)x_{ij}$$

求$-l(\theta)$的极小值点的梯度下降更新方程为

$$\theta_j := \theta_j + \alpha\sum_{i=1}^{n}\left[y_i - g(x_i^T\theta)\right]x_{ij}$$

# 模型推广（非参）

原模型

$$E(Y \mid X = x) = x^T\theta$$

$$J(\beta; X, Y) = \sum_{i=1}^{n}(x_i^T\theta - y_i)^2$$

局部加权线性回归(Locally weighted linear regression)模型

$$E(Y \mid X = x) = x^T\theta$$

$$J(\beta; X, Y) = \sum_{i=1}^{n}w_i(x_i^T\theta - y_i)^2$$

写成矩阵的形式

$$J(\beta; X, Y) = (X\theta - Y)^T W(X\theta - Y)$$

其中W为对角阵

$$W = diag(w_1, ..., w_n)$$

$w_i$的一个标准的选择是

$$w_i = exp\left[\frac{(x_i - x)^2}{2\tau^2}\right]$$

$\tau$是模型的一个参数，又称带宽(bandwidth parameter)，求解方法同上,解为

$$\theta = (X^T W X)^{-1} X^T W Y$$

显然这是一个非参的方法，对于每一个新输入x，要用训练集的数据计算$W$才能计算$\theta$，若$W = I_n$则退化为原线性回归模型。

# 用numpy和scikit-learn分别实现

用numpy直接计算公式得到参数，与sklearn.linear_model.LinearRegression的结果对比

## 数据说明

使用scikit-learn内建的boston数据集