# Linear Regression

## 定义

回归是监督学习的一个重要问题，回归用于预测输入变量和输出变量之间的线性关系，特别是当输入变量的值发生变化时，输出变量的值也随之发生变化。回归模型正是表示从输入变量到输出变量之间映射的函数

## 算法推导

**1.1 本文符号规定**

$x_j^{(i)}$ 表示数据集第 $m$ 个数据的第 $j$ 个属性取值，数据集一共有 $m$ 个数据， $j$ 个属性（特征）。

**1.2 线性回归模型**

模型定义为： $f(x) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$ 。

使用矩阵来表示就是 $f(x) = XW$ ，其中： $W = \begin{bmatrix} w_0 \\ w_1 \\ \ldots \\ w_n \end{bmatrix}$ 是所要求得一系列参数，

$X = \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_m^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_m^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_1^{(m)} & \ldots & x_n^{(m)} \end{bmatrix}$ 是输入的数据矩阵，因为考虑 $w_0$ 常数项，所以在 $X$ 第一列加上了一列1。 $X$ 的一行可以看做一个完整的输入数据， $n$ 代表一个数据有 $n$ 个属性（特征）， $m$ 行代表一共是 $m$ 个数据。数据集标签为 $y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \ldots \\ y^{(m)} \end{bmatrix}$ 。

线性回归模型的目标就是找到一系列参数 $w$ 来使得 $f(x) = XW$ 尽可能地贴近 $y$ 。

具体目标如图找到一条直线使得尽可能符合数据的分布，从而有一个新的样本点时，可利用学习得到的这条直线进行预测。

$$y^{(1)} = W_0 + x_1^{(1)}W_1 + x_2^{(2)}W_2 + \ldots\ldots + x_n^{(n)}W_n$$

根据每一行数据的得出以上数据

**1.3 损失函数**

使用均方误差作为损失函数，使用均方误差最小化目标函数的方法称为最小二乘法。

> 使用均方误差的原因：有十分好的几何意义，对应了常用的欧式距离。在线性回归中，就是找到一个直线，使得所有样本到直线的欧式距离最小。

损失代价函数定义为： $J(w) = \frac{1}{m}\sum_{i=1}^{m}(f(x^{(i)}) - y^{(i)})^2 = \frac{1}{m}(XW - y)^T(XW - y)$ ，

展开后得到：

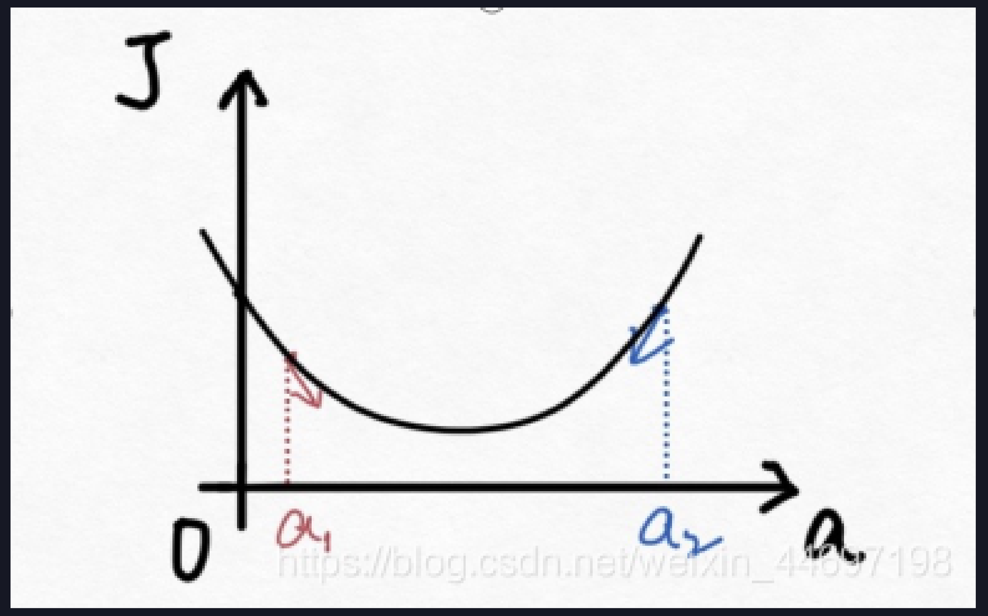
$$J(w) = \frac{1}{m}(W^TX^TXW - W^TX^Ty - y^TXW + y^Ty) = \frac{1}{m}(W^TX^TXW - 2W^TX^Ty + y^Ty)$$

那么线性回归的目标就是如何让f(x)和y之间的差异最小，换句话说就是w取什么值的时候f(x)和y最接近。也称为最小二乘法

当 $X^TX$ 为满秩矩阵或者正定矩阵时，可使用正规方程法，直接求得闭式解。

令 $\frac{\partial J(w)}{\partial w} = 0$ ，即： $\frac{\partial J(w)}{\partial w} = \frac{2X^T(XW - y)}{m} = 0$ ，可得： $W^* = (X^TX)^{-1}X^Ty$ 。

normal equation（矩阵方程法）：既然损失函数J(w)是凸函数，那么关于w对J(w)求偏导，并令其为零解出w。
前提，数据特征不能有严重的共线性，不然矩阵 (XTX) 不可逆
优缺点：
No need to choose α (不需要α)
No need to iterate (不需要迭代计算)
Needs to calculate (XTX)^-1 (需要计算(XTX)^2-1，其时间复杂度为O(n3))
Slow if n is very large (当n非常大时，速度非常慢)




梯度下降法：意味着W向右/左移一点。然后重复这个动作，直到J(W)到达最小值。

优缺点：
Needs to choose α (需要选取合适的α)
Needs many iterations (需要很多次的迭代计算)
Works well even when n is very large (在n很大时，能工作得很好)

这里再举个生活中的栗子，梯度下降法中随机给a赋一个预设值就好比你随机出现在一个山坡上，然后这时候你想以最快的方式走到山谷的最低点，那么你就得判断你的下一步该往那边走，走完一步之后同样再次判断下一步的方向，以此类推就能走到山谷的最低点了。而公式中的α我们称它为学习率，在栗子中可以理解为你每一步跨出去的步伐有多大，α越大，步伐就越大。（实际中α的取值不能太大也不能太小，太大会造成损失函数J接近最小值时，下一步就越过去了。好比在你接近山谷的最低点时，你步伐太大一步跨过去了，下一步往回走的时候又是如此跨过去，永远到达不了最低点；α太小又会造成移动速度太慢，因为我们当然希望在能确保走到最低点的前提下越快越好。）

## 线性回归的过拟合和欠拟合

解决线性回归过拟合的方法：

分析数据，重新做数据清洗，将征工程。
扩充数据集，收集更多数据。
减少特征数量 。
采用正则化方法：
L1正则化（Lasso回归）：稀疏化模型参数。
L2正则化（Rideg/岭回归）：缩小模型参数。

Lasso regression

$argmin(y - wx)^2 + \alpha|w|$

它是这样做惩罚的，在OLS拟合的基础上，对其系数的绝对值进行惩罚，目标函数长这样：

这样写目标函数就是想达到一个平衡，第一拟合的误差要小，第二系数的绝对值不能太大。

Ridge regression

$argmin(y - wx)^2 + \lambda w^2$

再说RIDGE，目标函数长这样

也是想达到一个类似的平衡，第一拟合的误差要小，第二系数的平方不能太大。

接下来的一个问题是，既然惩罚力度alpha太大了容易拟合不足，太低了容易过度拟合。究竟多大的惩罚力度是合适的？

这个问题对于OLS，LASSO，和RIDGE，有一个相对标准的做法，用赤池信息准则(AIC)或贝叶斯信息准则(BIC)进行判断。

解决线性回归欠拟合的方法：

分析数据，增加特征淮度。
增加多项式特征阶数。
减小正则项的超参系数值。
局部加权回归（详情见第7节）

## 共线性问题

多重共线性（Multicollinearity）是指线性回归模型中的自变量之间由于存在高度相关关系而使模型的权重参数估计失真或难以估计准确的一种特性，多重是指一个自变量可能与多个其他自变量之间存在相关关系。

我们进行回归分析需要了解每个自变量对因变量的单纯效应，多重共线性意味着自变量间之间存在某种函数关系，例如，你的两个自变量间（X1和X2）存在函数关系，那么X1改变一个单位时，X2也会相应地改变，此时你无法做到固定其他条件，单独考查X1对因变量Y的作用，你所观察到的X1的效应总是混杂了X2的作用，这就造成了分析误差，使得对自变量效应的分析不准确，所以做回归分析时需要排除多重共线性的影响。

回归模型缺乏稳定性。样本的微小扰动都可能带来参数很大的变化；
难以区分每个解释变量的单独影响；
参数的方差增大；
变量的显著性检验失去意义；
影响模型的泛化误差；

解决方法：
共线性问题并不是模型的设定错误，它是一种数据缺陷，可以通过增加样本量来解决

在特征比较多的时候，先变量聚类，每类中选择单特征比较强的，也可以根据1-r^2小的选择有代表性的特征（r^2表示的是其他变量能否线性解释的部分，1-r^2表示的是容忍度，也就是其他变量不能解释的部分；变量聚类是多选一，因此需要选择一个具有代表性的变量，选择容忍度小的变量；另vif就是容忍度的倒数）

在变量聚类的步骤中也可以结合 方差膨胀因子、相关系数以及业务理解来筛选特征

## 线性回归的优缺点

优点
直接。
快速。
可解释性好。

缺点
需要严格的假设。
需处理异常值，对异常值很敏感，对输入数据差异也很敏感。
线性回归存在共线性，自相关，异方差等问题。