回归算法总结

常用的回归算法有线性回归、多项式回归、岭回归、Lasso 回归、ElasticNet 回归。

一、 线性回归

线性回归是指完全由线性变量组成的回归模型，线性回归模型它是体现单个或多个输入变量（自变量）与输出变量（因变量）之间的关系。输出变量是输入变量的线性组合。线性回归的建模如下：

$$y = a_1 * X_1 + a_2 * X_2 + a_3 * X_3 \ldots + a_n * X_n + b$$

或者：

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

其中a_i是系数，X_i是变量，b是偏置。它只适用于建模线性可分的数据，使用随机梯度下降(SGD)算法来确定这些权重。

它的损失函数是:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^i\right) - y^i\right)^2$$

目标则是$\min\ J(\theta_0, \theta_1 \ldots \theta_n)$

线性回归的几个关键点：

1） 建模快速简单，比较适用于来建模关系不是很复杂并且数据量不大的情况。

2） 有直观的理解和解释。

3） 线性回归对异常值比较敏感。

线性回归中可能会遇到的问题：

1） 求解损失函数的最小值有两种方法：梯度下降以及正规方程。

2） 特征缩放：即对特征数据进行归一化操作，进行特征缩放的好处有两个，一个是能够提升模型的收敛速度，因为如果特征间的数据相差级别较大的话

二、 多项式回归

当我们要创建适合处理非线性可分数据的模型时，我们需要使用多项式回归。在这种回归模型中，最佳拟合的不是一条直线，而是一条符合数据点的曲线。对于一个多项式回归，一些自变量的指数是大

于 1 的。例如，下面是他们的数学表达式：

$$Y = a_1 * X_1 + a_2 * X_2^2 + \cdots + a_n * X_n^m + b$$

其中的一些变量有指数，而一些变量没有指数。

多项式回归的几个特点：

1） 能够模拟非线性可分的数据。

2） 要设置变量的指数

3） 需要仔细的设计变量的指数，所以需要一些先验知识才能选择最佳指数。

4） 如果指数选择不当，容易过拟合。

三、 岭回归

岭回归和后续的 Lasso 回归可以视为在线性回归的基础上加了正则项，只是区别是一个加的是 L2 正则(岭回归)，一个加的是 L1 正则(Lasso 回归)。

标准的线性或者多项式回归，在特征变量之间有很高的共线性时会失败，共线性指的是自变量之间存在近似线性关系，这会对回归分析带来很大的影响。

我们进行回归分析的目的是需要了解每个自变量对因变量的单纯效应，高共线性就是说自变量间存在某种函数关系，如果你的两个自变量（x1 和 x2）存在函数关系，那么 x1 改变一个单位时，x2 也会相应的改变，此时你无法做到固定其他条件，单独考察 x1 对应变量 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，使得对自变量的分析不准确，所以做回归分析时需要排除高共线性的影响。

高共线性的存在可以通过以下几种方式来确定：

--1）尽管从理论上讲，该变量应该与 y 高度的相关，但是回归系数并不明显，这就表明有可能其他变量可能与该变量相关。

--2）添加或删除 x 特征变量时，回归系数会发生显著变化。

--3）x 特征变量具有较高的成对相关性（检查相关矩阵）

我们可以再看下标准线性回归的优化函数，然后分析岭回归如何解决上述问题：

$$\min \ ||Xw - y||^2$$

其中 X 表示特征变量，w 表示权重，y 表示真实情况。

由于共线性，所以多元回归模型中的一个特征变量可以有其他变量进行线性预测。为了解决这个问题，岭回归为变量增加了一个小的平方偏差因子（其实就是正则项）。

$$\min||Xw - y||^2 + z||w||^2$$

这种平方偏差因子向模型中映入少量偏差，但是大大减小了方差

岭回归的要点：

- 这种回归的假设和最小平方回归相同，不同点在于最小平方回归的时候，我们假设数据的误差服从高斯分布使用的是极大似然估计（MLE），在岭回归的时候，**由于添加了偏差因子，即 w 的先验信息，使用的是极大后验估计（MAP）来得到最终的参数。**
- **它缩小了系数的值，但没有达到零，这表明没有特征选择功能。**

四、 Lasso 回归

Lasso 回归与岭回归非常相似，因为两种技术都有相同的前提：它们都是在回归优化函数中增加一个偏置项，以减少共线性的影响，从而减少模型的方差。然而不像岭回归那样使用平方偏差，Lasso 回归使用绝对值偏差最为正则化项：

$$\min\left|\left|Xw - y\right|\right|^2 + z||w||$$

岭回归和 Lasso 回归之间存在一些差异，基本上可以归结为 L2 和 L1 正则化的性质差异。

Lasso 回归的特性：

- **内置的特征选择：**这是 L1 范数的一个非常有用的属性，而 L2 范数不具备这种特性。这实际上是因为 L1 范数倾向于产生稀疏系数，例如，假设模型有 100 个系数，但其中只有 10 个系数是非零系数，这实际上是说“其它 90 个变量对预测目标值没有用处”。而 L2 范数产生非稀疏系数，所以没有这个属性。因此可以说 Lasso 回归做了一种“参数选择”形式，未被选中的特征变量对整体的权重为 0.
- **稀疏性：**指矩阵中只有极少数条目非零。L1 范数具有产生

有零值或有很少大系数的非常小值的许多系数的属性。

- **计算效率：**L1 范数没有解析值，但 L2 范数有。这使得 L2 范数的解可以通过计算得到。然而，L1 范数的解具有稀疏性，这使得它可以与稀疏算法一起使用，这使得在计算上更有效率（暂时不懂其具体计算过程）。

五、 弹性网络回归（ElasticNet Regression）

ElasticNet 是 Lasso 回归和岭回归技术的混合体，它使用了 L1 和 L2 正则化，也达到了两种技术的效果：

$$\min ||Xw - y||^2 + Z_1\big||w\big| + Z_2||w||^2$$

在 Lasso 回归和岭回归之间进行权衡的一个实际优势就是，它允许 ElasticNet 继承岭回归的一些稳定性。

ElasticNet 的特点：

- 它鼓励在高度相关的变量的情况下的群体效应，而不是像 Lasso 那样将其中一些置零。**当多个特征和另一个特征相关的时候弹性网络非常有用。Lasso 倾向于随机选择其中一个，而弹性网络倾向于选择两个。**
- 对所选变量的数量没有限制。

小结

上述所有的回归正则化方法(Lasso 回归、岭回归和 ElasticNet)在数据集中的变量之间具有高维度和多重共线性的情况下也能有良好的效果。