ZHANG RONG

数据挖掘与机器学习

线性回归（Linear Regression）

OCTOBER 28, 2016 | ZR9558 | LEAVE A COMMENT

回归方法是为了对连续型的数据做出预测，其中最简单的回归方法当然就是线性回归。顾名思义，线性回归就是使用线性方程来对已知的数据集合进行拟合，达到预测未来的目的。线性回归的优点就是结果十分容易理解，计算公式简单；缺点则是对非线性的数据拟合程度不够好。例如，用一个线性函数 $y = kx + b$ 去拟合二次函数 $f(x) = x^2$，结果总是不尽人意。为了解决这类问题，有人提出了局部加权线性回归（**locally weighted linear regression**），岭回归（**ridge regression**），**LASSO** 和 前向逐步线性回归（**forward stagewise linear regression**）。本文中将会一一介绍这些回归算法。

（一）线性回归（**Linear Regression**）

假设矩阵 X 的每一行表示一个样本，每一列表示相应的特征，列向量 Y 表示矩阵 X 所对应的取值，那么我们需要找到一个列向量 Θ 使得 $Y = X\Theta$。当然，这样的 Θ 在现实的数据集中几乎不可能存在。不过，我们可以寻找一个 Θ 使得列向量 $Y - X\Theta$ 的 Eulidean 范数足够小。换言之，我们需要找到一个向量 Θ 使得

$$\sum_{i=1}^{m}(y_i - x_i\Theta)^2 = (Y - X\Theta)^T(Y - X\Theta)$$

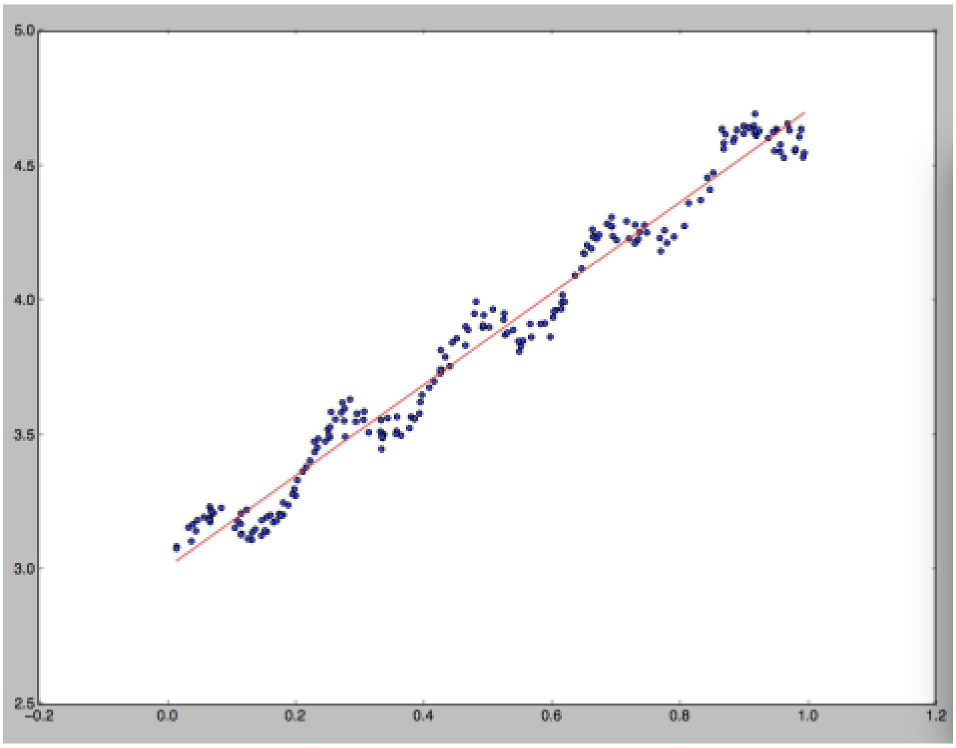
的取值足够小，其中 m 是矩阵 X 的行数，x_i 表示矩阵 X 的第 i 个行向量。通过数学计算可以得到：

$$(Y - X\Theta)^T(Y - X\Theta) = Y^TY - 2Y^TX\Theta + \Theta^TX^TX\Theta$$

对 Θ 求导之后得到：$-2X^TY + 2X^TX\Theta = 0$，求解 Θ 之后得到 $\Theta = (X^TX)^{-1}X^TY$。因此，对于矩阵 X 和列向量 Y 而言，最佳的线性回归系数是

$$\Theta = (X^TX)^{-1}X^TY.$$

举例说明：蓝色的是数据集，使用线性回归计算的话会得到一条直线。



（二）局部加权线性回归（Locally Weighted Linear Regression）

线性回归的一个问题就是会出现欠拟合的情况，因为线性方程确实很难精确地描述现实生活的大量数据集。因此有人提出了局部加权线性回归（Locally Weighted Linear Regression），在该算法中，给每一个点都赋予一定的权重，也就是

$$\sum_{i=1}^{m} w_i(y_i - x_i\Theta)^2 = (Y - X\Theta)^T W(Y - X\Theta),$$

其中 W 表示以 $\{w_1, ..., w_m\}$ 为对角线的对角矩阵，其中 m 是矩阵 X 的行数，x_i 表示矩阵 X 的第 i 个行向量。通过计算可以得到：

$$(Y - X\Theta)^T W(Y - X\Theta) = Y^T WY - 2Y^T WX\Theta + \Theta^T X^T WX\Theta,$$

对 Θ 求导之后得到：

$$-2(Y^T WX)^T + 2X^T WX\Theta = -2X^T WY + 2X^T WX\Theta.$$

令导数等于零之后得到：$\Theta = (X^T WX)^{-1} X^T WY$。因此，如果使用局部加权线性回归的话，最佳的系数就是

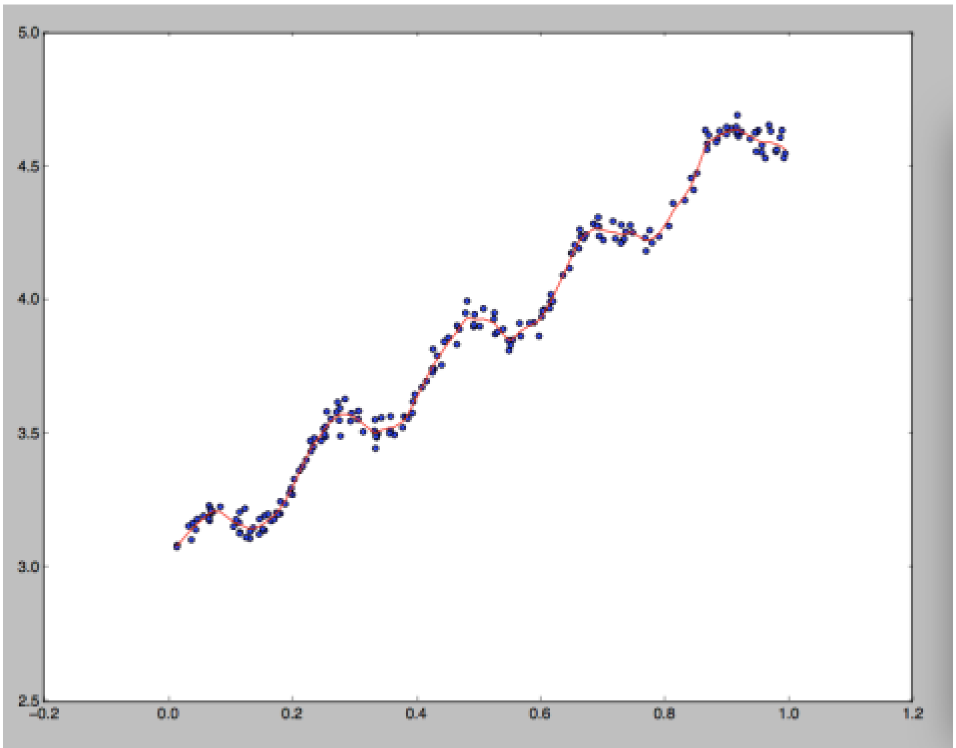
$$\Theta = (X^T WX)^{-1} X^T WY.$$

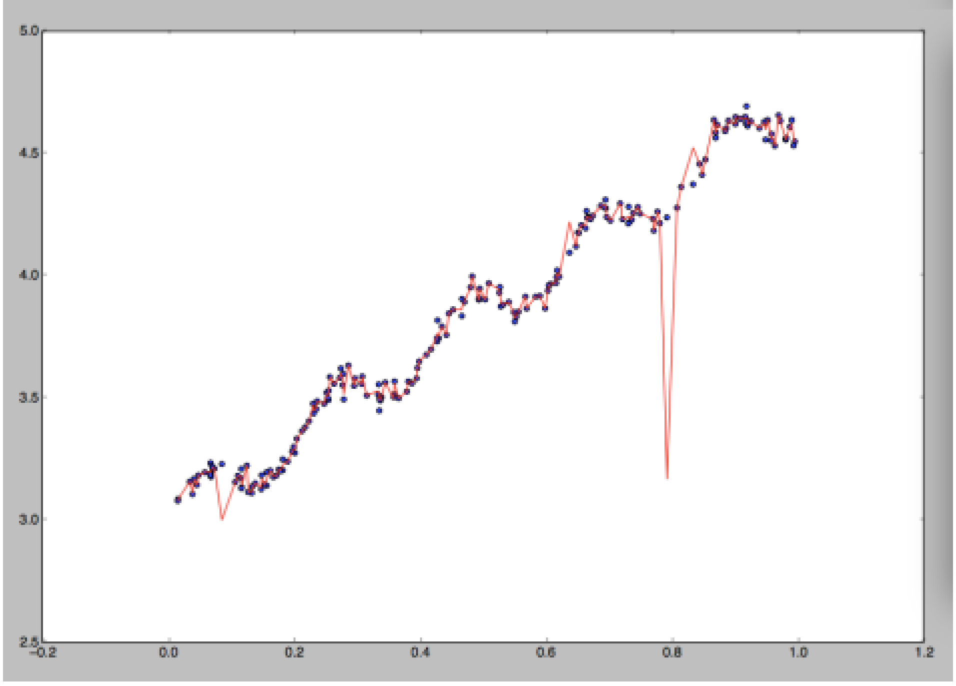
局部加权线性回归需要确定权重矩阵 W 的值，那么就需要定义对角线的取值，通常情况下我们会使用高斯核。

$$w_i = \exp\{-\frac{(x_i - x)^2}{2k^2}\}.$$

其中 k 是参数。从高斯核的定义可以看出，如果 x 与 x_i 隔得很近，那么 w_i 就会较大；如果隔得较远，那么 w_i 就会趋向于零。意思就是说：在局部形成了线性回归的算法，在整体并不一定是线性回归。在局部线性回归中，k 就是唯一的参数值。

如果选择了合适的 k，可以得到一条看上去还不错的曲线；如果选择了不合适的 k，就有可能出现过拟合的情况。





（三）岭回归（**Ridge Regression**）和 **LASSO**

如果在某种特殊的情况下，特征的个数 n 大于样本的个数 m，i.e. 矩阵 X 的列数多于行数，那么 X 不是一个满秩矩阵，因此在计算 $(X^TX)^{-1}$ 的时候会出现问题。为了解决这个问题，有人引入了岭回归（ridge regression）的概念。也就是说在计算矩阵的逆的时候，增加了一个对角矩阵，目的是使得可

以对矩阵进行求逆。用数学语言来描述就是矩阵 X^TX 加上 λI，这里的 I 是一个 $n\times n$ 的对角矩阵，使得矩阵 $X^TX+\lambda I$ 是一个可逆矩阵。在这种情况下，回归系数的计算公式变成了

$$\Theta=(X^TX+\lambda I)^{-1}X^TY.$$

岭回归最初只是为了解决特征数目大于样本数目的情况，现在也可以用于在估计中加入偏差，从而得到更好的估计。

从另一个角度来讲，当样本的特征很多，而样本的数量相对少的时候，$\sum_{i=1}^m(y_i-x_i\Theta)^2$ 很容易过拟合。为了缓解过拟合的问题，可以引入正则化项。如果使用 L^2 正则化，那么目标函数则是

$$\sum_{i=1}^m(y_i-x_i\Theta)^2+\lambda||\Theta||_2^2=(Y-X\Theta)^T(Y-X\Theta)+\lambda\Theta^T\Theta,$$

其中 $\lambda>0$。通过数学推导可以得到：

$$(Y-X\Theta)^T(Y-X\Theta)+\lambda\Theta^T\Theta=Y^TY-2\Theta^TX^TY+\Theta^TX^TX\Theta+\lambda\Theta^TI\Theta.$$

对 Θ 求导之后得到：

$$-2X^TY+2(X^TX+\lambda I)\Theta,$$

令导数等于零可以得到：$\Theta=(X^TX+\lambda I)^{-1}X^TY.$ 因此，从另一个角度来说，岭回归（Ridge Regression）是在线性规划的基础上添加了一个 L^2 范数的正则化，可以用来降低过拟合的风险。

需要注意的是：在进行岭回归的时候，需要在一开始就对特征进行标准化处理，使得每一维度的特征具有相同的重要性。具体来说就是 (特征-特征的均值)/特征的方差，让每一维度的特征都满足零均值和单位方差。

另外，如果把岭回归中的 L^2 范数正则化替换成 L^1 范数，那么目标函数就变成了

$$\sum_{i=1}^m(y_i-x_i\Theta)^2+\lambda||\Theta||_1$$

其中的参数 $\lambda>0$。L^1 和 L^2 范数都有助于降低过拟合的风险，使用 L^1 范数的方法被称为 LASSO（Least Absolute Shrinkage and Selection Operation）。使用 L^1 范数比使用 L^2 范数更加容易获得稀疏解（sparse solution），即它求得的参数 Θ 会有更少的非零分量。Θ 获得稀疏解意味着初始的 n 个特征中仅有对应着 Θ 的非零分量的特征才会出现在最终的模型中。于是，求解 L^1 范数正则化的结果是得到了仅采用一部分原始特征的模型；从另一个角度来说，基于 L^1 正则化的学习方法就是一种嵌入式的特征选择方法，其特征选择的过程和训练的过程融为一体，同时完成。

（四）前向逐步线性回归（**Forward Stagewise Linear Regression**）

前向逐步线性回归算法是一种贪心算法，目的是在每一步都尽可能的减少误差。初始化的时候，所有的权重都设置为1，然后每一步所做的据测就是对某个权重增加或者减少一个很小的值 ϵ。

该算法的伪代码如下所示：

```
数据标准化，使其分布满足零均值和单位方差
在每一轮的迭代中：
   设置当前最小误差为正无穷
   对每个特征：
      增大或者缩小：
         改变一个系数得到一个新的权重W
         计算新W下的误差
         如果误差Error小于当前误差：设置Wbest等于当前的W
      将W设置为新的Wbest
```

（五）总结

与分类一样，回归也是预测目标值的过程。但是分类预测的是离散型变量，回归预测的是连续型变量。但是在大多数情况下，数据之间会很复杂，这种情况下使用线性模型确实不是特别合适，需要采用其余的方法，例如非线性模型等。