线性回归

理论

回归

- 将实例点用一条线来表达它们的关系

线性

- 实例点看起来是一条直线
- 这个直线不一定是二维的，可以是更高维的直线

线性回归

- 用一条直线来表达实例点的关系
- 主要可分为：简单一元线性回归，多元线性回归

简单一元线性回归

- 由一元线性方程演变而来

$$y = wx + b$$

- 使用机器学习概念参数w和b

$$\hat{y} = w_0 + w_1x + \epsilon$$

多元线性回归

- 拥有更多的特征

$$\hat{y}_i = w_0 + w_1x_1 + w_2x_2 + \ldots + w_ix_i$$

- 如果我们利用矩阵形式的表达式

$$\hat{y} = X^Tw$$

损失函数

- 在我们寻找最合适的那条直线时，我们该如何判断那一条才是最合适的？
  - 最简单的方法就是使用平方差公式
  - 更为准确的是残差平方和

最小二乘法

- 由残差平方和演变而来
- 对于简单一元线性回归，仍保持最基础的残差平方和公式

- 而对于多元线性回归，我们以矩阵形式的表达式便是

$$SSE = (y - Xw)^T (y - Xw)$$

这个形式是寻找一条直线残差平方和

- 如果想寻找一条最合适的直线，通过求导求出极限值，我们得出

$$w = (X^T X)^{-1} X^T y$$

但是当我们的实例矩阵不是满秩矩阵时，我们会遇上最小二乘求解结果并非唯一解的情况，因此对此算法进行优化

变化

岭回归（Ridge Regression）

- 为了使实例矩阵为可逆矩阵，我们在其中增加了一个单位矩阵I, 从而得到下列新的公式

$$w = (X^T X + \lambda I)^{-1} X^T y$$

- 可逆矩阵的必要条件是不存在多重共线性（同时有多个解）
- 这个方法也被称为L2正则化

Lasso回归（Lasso Regression）

- 不同于L2正则化，Lasso回归里增加了一个带约束条件的损失函数，从此得到

$$w = (X^T X)^{-1} X^T y + \lambda \sum |w_i|$$

- 这个方法被称为L1正则化

参考文献

- 机器学习| 算法笔记-线性回归（Linear Regression）
- 用人话讲明白线性回归LinearRegression