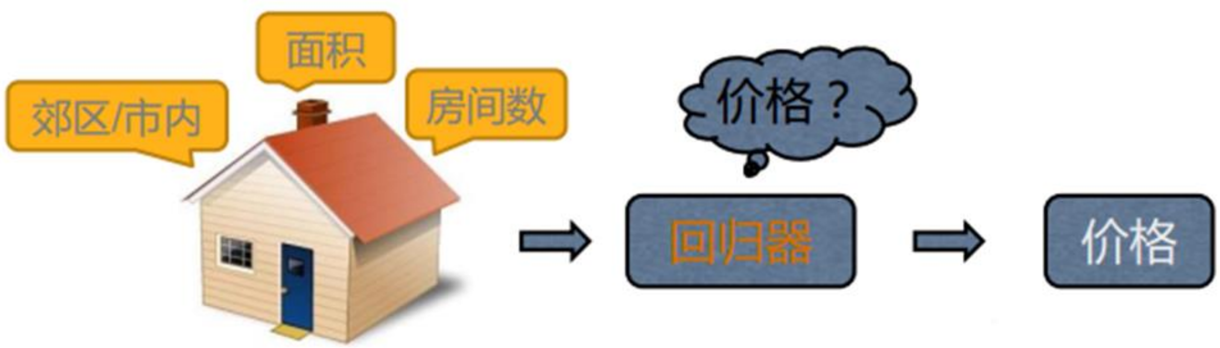
精英班系列课程

线性回归

线性回归

- 回归分析：
  - ✓ 回归：统计学分析数据的方法，目的在于了解两个或多个变量间是否相关、研究其相关方向与强度，并建立数学模型以便观察特定变量来预测研究者感兴趣的变量
  - ✓ 回归分析可以帮助人们了解在自变量变化时因变量的变化量。一般来说，通过回归分析我们可以由给出的自变量估计因变量的条件期望



线性回归

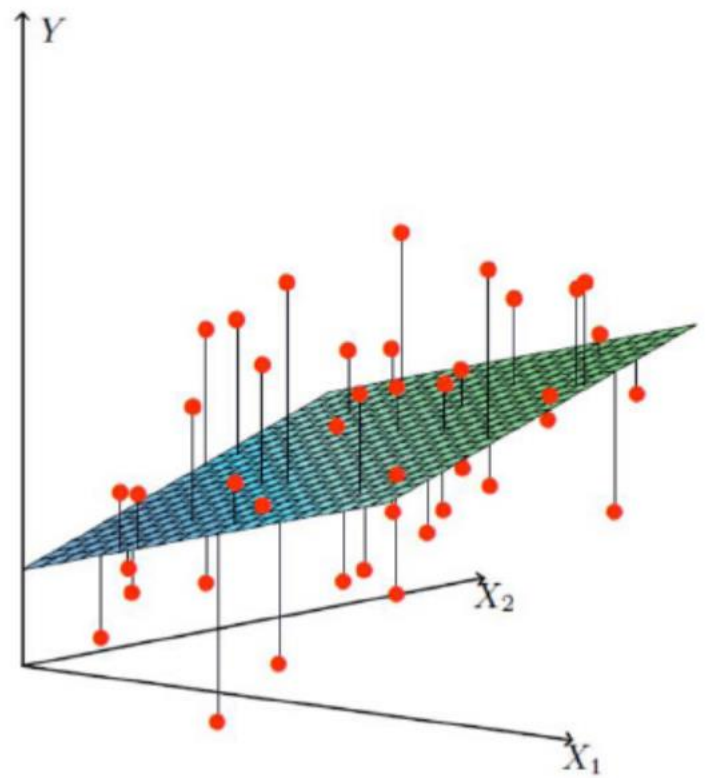
- 例子：
  - ✓ 特征：工资和年龄
  - ✓ 标签：预测银行会贷款多少钱
  - ✓ 考虑：工资和年龄都会影响最终银行贷款的结果，各自影响有多大呢？（要求解的参数）

工资	年龄	额度
4000	25	20000
8000	30	70000
5000	28	35000
7500	33	50000
12000	40	85000

线性回归

CDA数据分析师 CERTIFIED DATA ANALYST

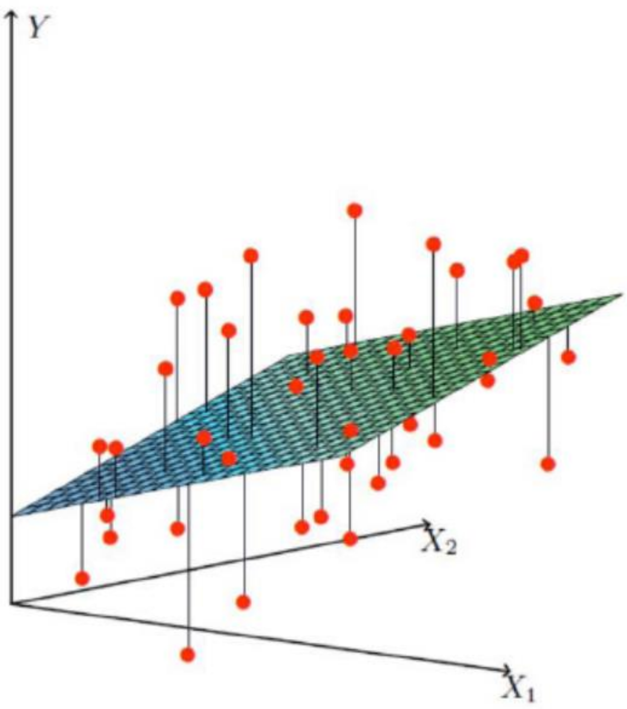
- 通俗解释：
  - ✓ x1，x2是两个特征（年龄，工资），Y是银行最终会借给我们多少钱
  - ✓ 找到最合适的一个超平面来拟合我们的数据点



线性回归

- 数学推导：
  - ✓ 假设θ_1是年龄的参数，θ_2是工资的参数
  - ✓ 拟合的超平面（假设函数）：

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = X\theta$$

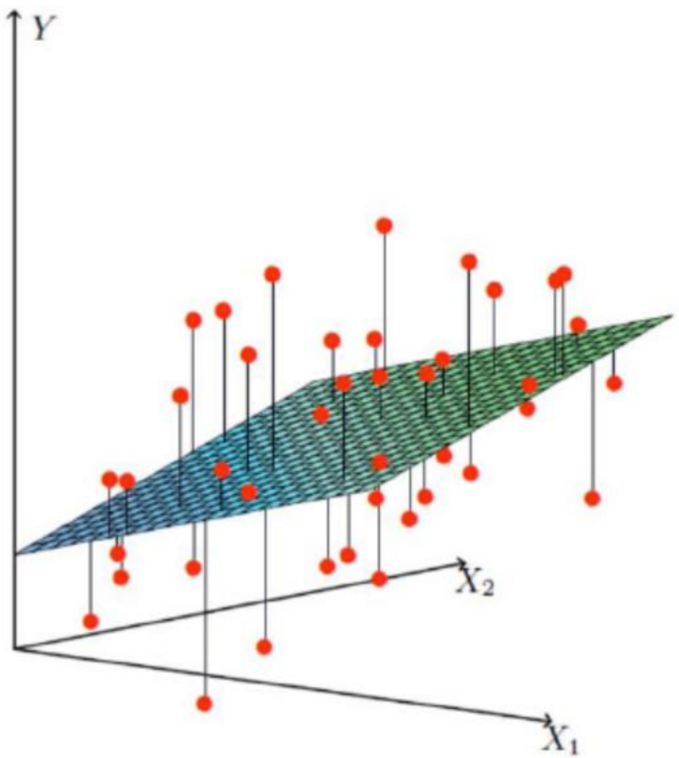


线性回归

CDA数据分析师
CERTIFIED DATA ANALYST

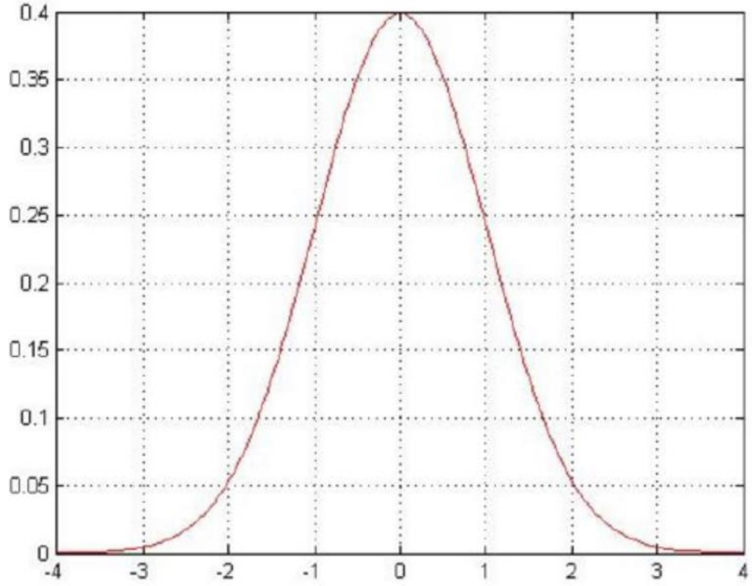
- 误差分布：
  - ✓ 真实值和预测值之间的差异：
  - ✓ 对于每个样本：

$$y^{(i)} = x^{(i)}\theta + \varepsilon^{(i)}$$



线性回归

- 误差分布：
  - ✓ 误差 $\varepsilon^{(i)}$是独立并且具有相同的分布，服从均值为0，方差为σ^2的高斯分布



线性回归

- 误差分布：
  - ✓ 独立：张三和李四同时贷款，互相没关系
  - ✓ 同分布：都是来自同一家银行贷款
  - ✓ 高斯分布：银行可能会多给，也可能少给，但是绝大多数情况下，这个浮动不会太大，极小情况下浮动会比较大

线性回归

- 线性回归的最小二乘法推导：

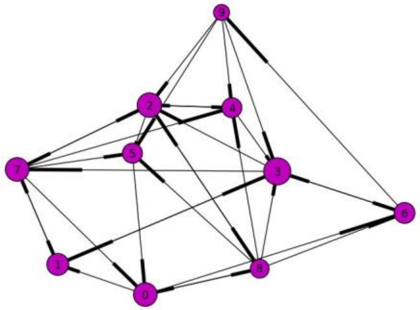
代价函数： $J(\theta)=\frac{1}{2}\sum_{i=0}^{m}(y^{(i)}-x^{(i)}\theta)^2=\frac{1}{2}\sum_{i=0}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$

参数求解： $\theta=(X^TX)^{-1}X^TY$

线性回归

- 利用线性回归的最小二乘法实现拟合

梯度下降法

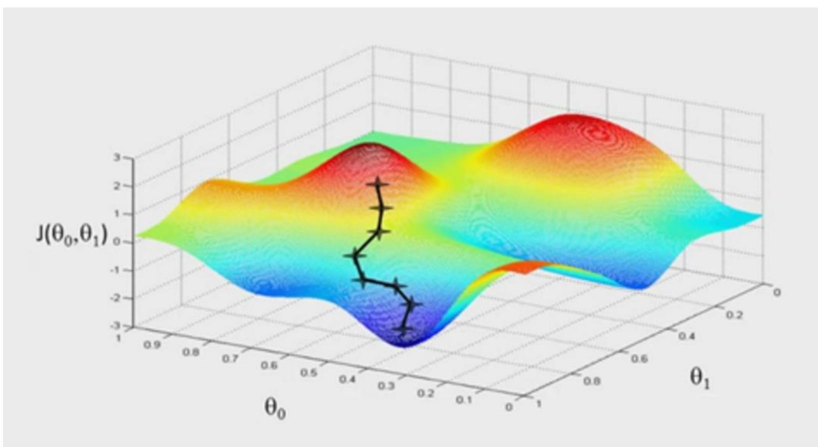


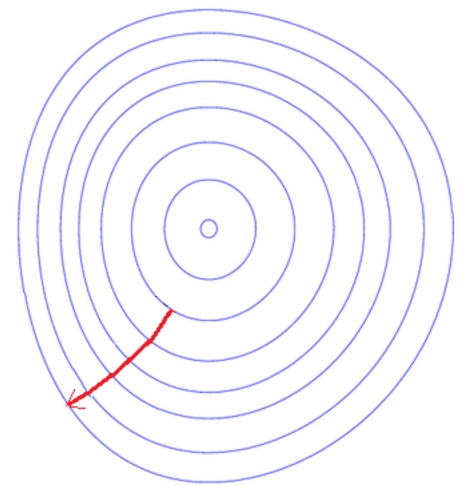
线性回归

- 最小二乘法回归存在问题：
  - ✓ X^TX 矩阵必须可逆
  - ✓ 当特征数较多时，求逆运算时间开销较大
  - ✓ 引入梯度下降法

线性回归

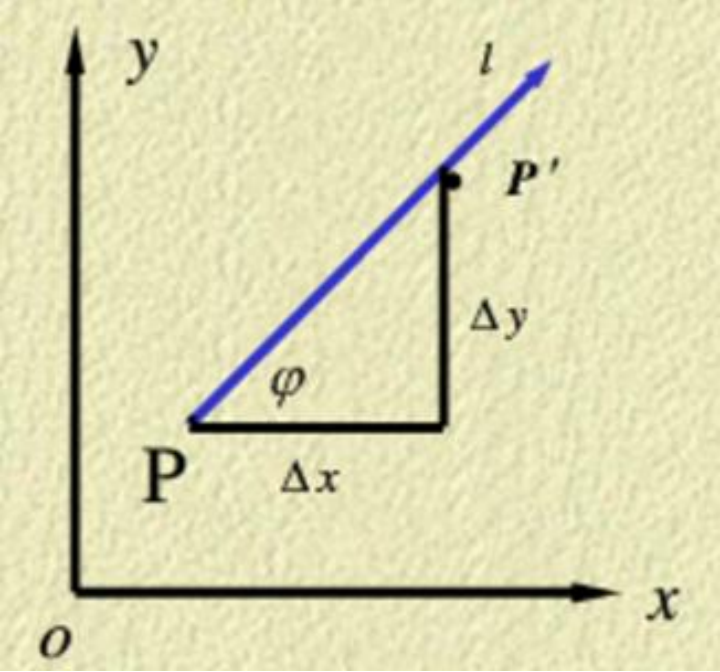
- 梯度下降：
  - ✓ 梯度下降法（Gradient descent）是使得代价函数达到最小的经典方法之一





线性回归

- 梯度的概念：
  - ✓ 设函数z=f(x, y)在点P(x, y)的某一邻域U(P)内有定义，自点P引射线l到点P'(x+Δx, y+Δy)且P'∈U(P)，如下图所示



线性回归

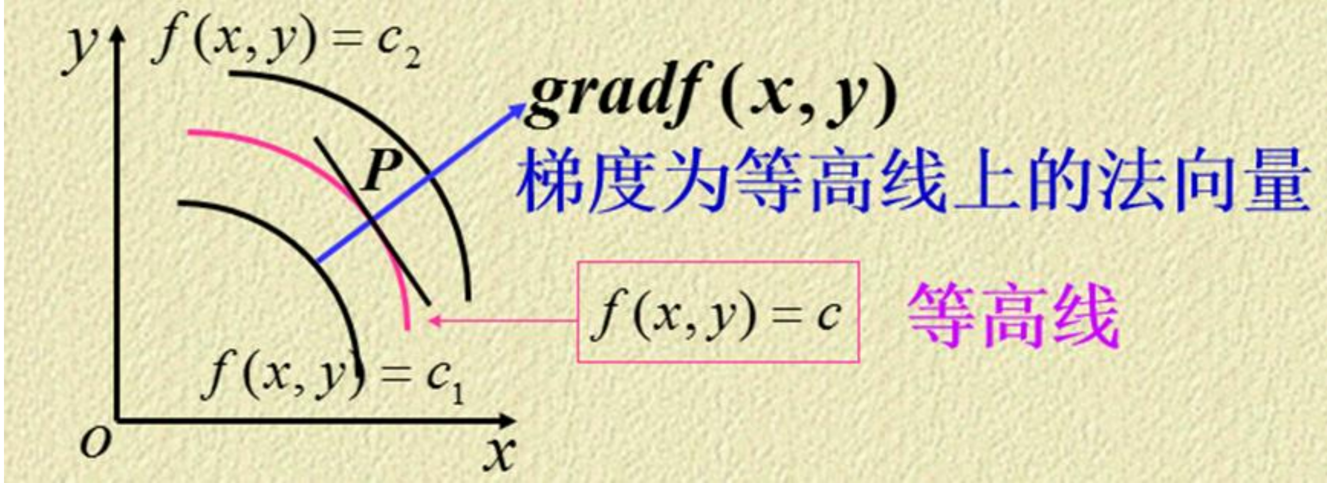
- 梯度的概念：
  - ✓ 定义函数z=f(x, y)在点P沿方向l的方向导数为：

$$\frac{\partial f}{\partial l}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}，\text{其中}\rho=\sqrt{(\Delta x)^2+(\Delta y)^2}$$

  - ✓ 方向导数可以理解为，函数z=f(x, y)沿某个方向变化的速率。函数z=f(x, y)在点P沿哪个方向增加的速度最快？这个方向就是梯度的方向

$$gradf(x,y)=\frac{\partial f}{\partial x}\vec{i}+\frac{\partial f}{\partial y}\vec{j}$$

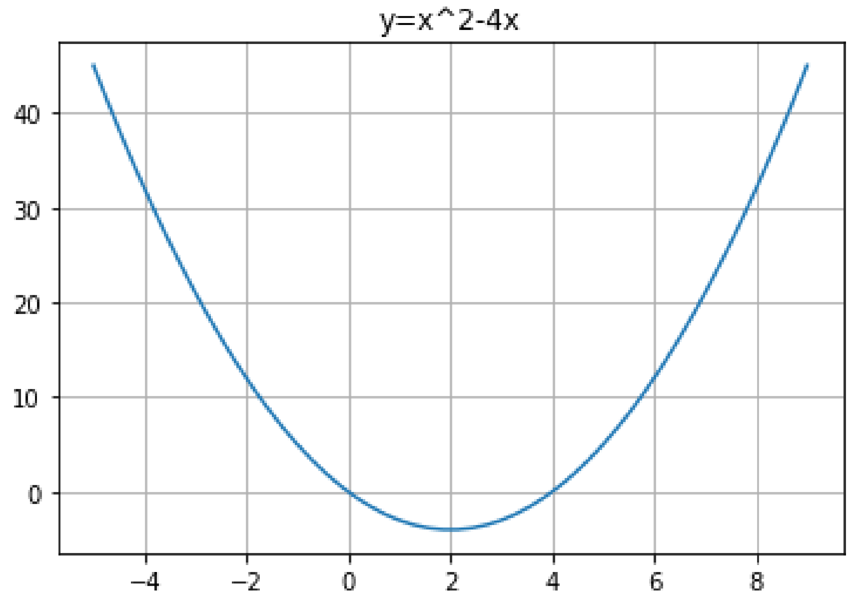
- 梯度的概念：



✓ z=f(x, y) 在点P(x, y) 处的梯度方向与点P的等高线f(x, y)=c在这点的法向量的方向相同，且从数值较低的等高线指向数值较高的等高线

梯度下降法

- 梯度下降法：
  - ✓ 求解使得f(x)最小值时候的x值，使用梯度下降算法



迭代公式： $$x_{i+1} = x_i - \alpha \frac{\partial f(x_i)}{x_i}$$

线性回归

- 梯度方向的计算：
  - ✓ 对于每一个向量θ的每一维分量θj，我们都可以求出梯度的方向，也就是误差函数J(θ)下降最快的方向

$$\frac{\partial}{\partial\theta_j}J(\theta)=\frac{\partial}{\partial\theta_j}\frac{1}{2}\sum_{i=0}^{m}(h_\theta(x^{(i)})-y^{(i)})^2=\sum_{i=0}^{m}(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$$

  - ✓ 参数的迭代：

$$\theta'_j=\theta_j-\alpha\frac{\partial J(\theta)}{\partial\theta_j}=\theta_j-\alpha\sum_{i=0}^{m}(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$$

线性回归

- 梯度下降法的Python实现

线性回归

- 梯度下降的算法调优：
  - ✓ 算法的步长选择
  - ✓ 算法参数的初始值选择
  - ✓ 归一化

线性回归

- 梯度下降法分类：
  - ✓ 批量梯度下降算法（BGD，Batch gradient descent algorithm）
  - ✓ 随机梯度下降算法（SGD，Stochastic gradient descent algorithm）
  - ✓ 小批量梯度下降算法（MBGD，Mini-batch gradient descent algorithm）

线性回归

- 批量梯度下降算法（BGD）：每一次迭代使用全部的样本

$$J(\theta_0,\theta_1,\ldots,\theta_n)=\sum_{i=0}^{m}(h_\theta(x_0,x_1,\ldots,x_n)-y_i)^2$$

$$\theta_i=\theta_i-\alpha\frac{\partial J(\theta_1,\theta_2,\ldots,\theta_n)}{\partial\theta_i}$$

- 特点：
  - ✓ 能达到全局最优解（凸函数情况下）
  - ✓ 当样本数目很多时，训练过程缓慢

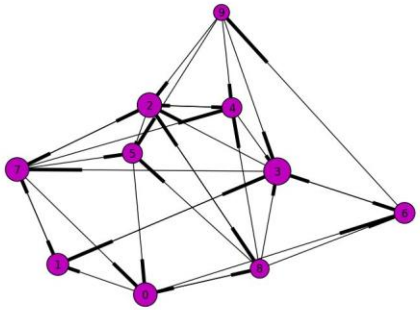
线性回归

- 随机梯度下降算法（SGD）：每一次更新参数只使用一个样本，进行多次更新
- 特点：
  - ✓ 迭代速度快
  - ✓ 准确度下降，每次不一定朝着收敛的方向，容易陷入局部最优
  - ✓ 非凸函数情况下有可能跳出局部最优

线性回归

- 小批量梯度下降算法（MBGD）：更新每一参数时都使用一部分样本来进行更新
- 相对于随机梯度下降，mini-batch梯度下降降低了收敛波动性，即降低了参数更新的方差，使得更新更加稳定。相对于批量梯度下降，其提高了每次学习的速度
- 更新随机选择样本数量需要根据具体问题而选择，实践中可以进行多次试验，选择一个更新速度与次数都较适合的样本数。mini-batch梯度下降可以保证收敛性，常用于神经网络中

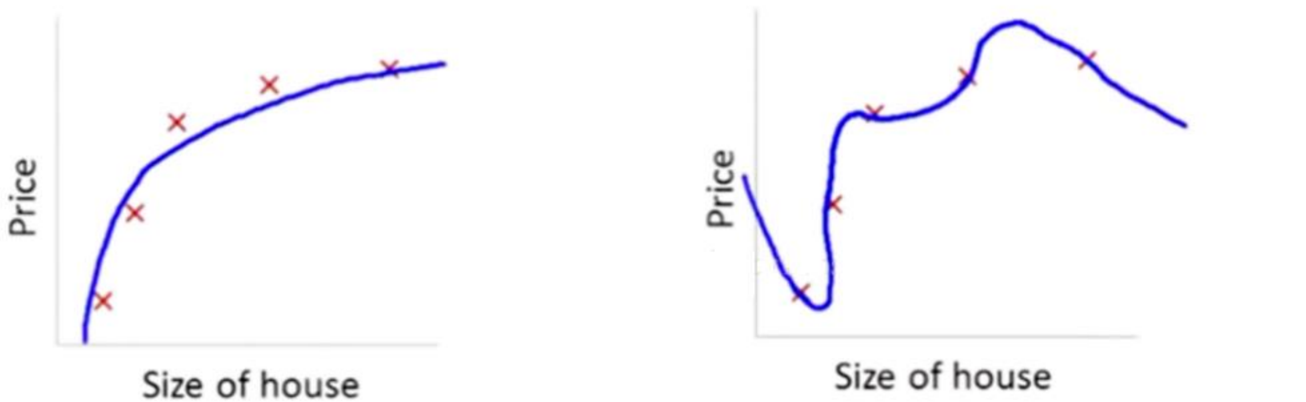
岭回归(Ridge Regression)



岭回归

CDA数据分析师 CERTIFIED DATA ANALYST

- 线性回归存在的问题？
  - ✓ 如果数据的特征的数目比样本的数目还多，那么输入数据的矩阵X将不是满秩矩阵(特征值之间具有相关性)，非满秩矩阵不存在逆矩阵
  - ✓ 为防止模型过拟合，建立线性模型的时候经常加入正则化项。一般有L1正则化和L2正则化



$\theta_0 + \theta_1 x + \theta_2 x^2$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

岭回归

- 岭回归模型

✓ 岭回归是在平方误差的基础上增加正则项（L2正则）：

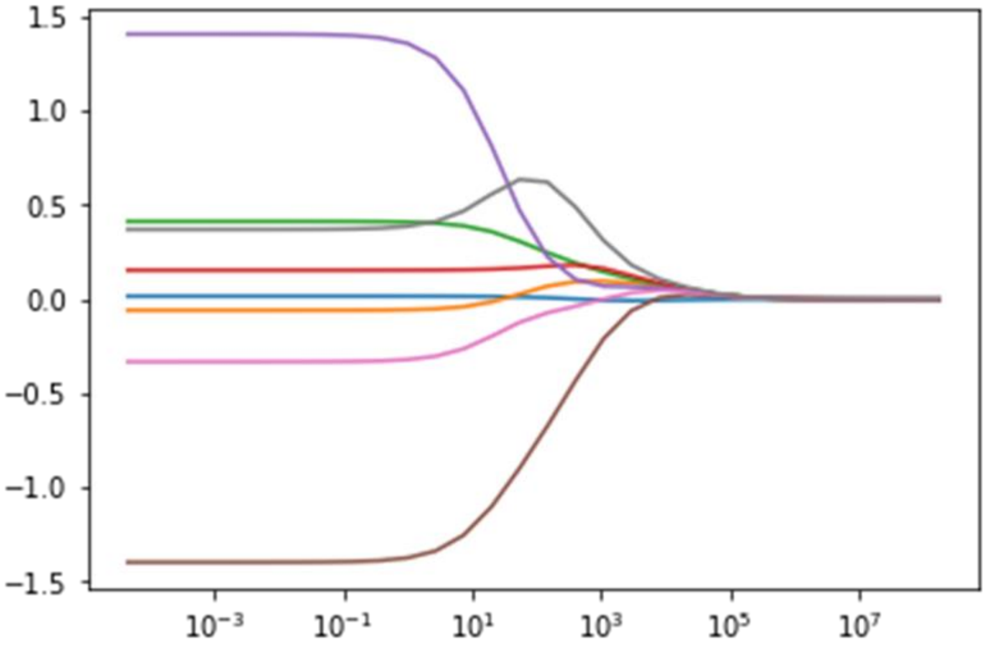
$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h_{\theta}(x^{(i)})-y^{(i)})^{2}+\lambda\sum_{i=1}^{\mathrm{n}}\theta_{j}^{\ 2}$$

✓ 其中，$\lambda>0$ 。通过确定 λ 的值可以在模型方差（回归系数的方差）和偏差之间达到平衡，随着 λ 的增大，模型方差减小而偏差增大

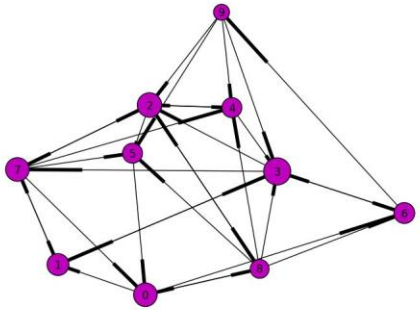
✓ 与线性回归一样，利用最小二乘法求解岭回归模型参数时，可得：$\theta=(X^{T}X+\lambda I)^{-1}X^{T}Y$

岭回归

- 岭回归模型的系数选择



Lasso Regression



Lasso Regression

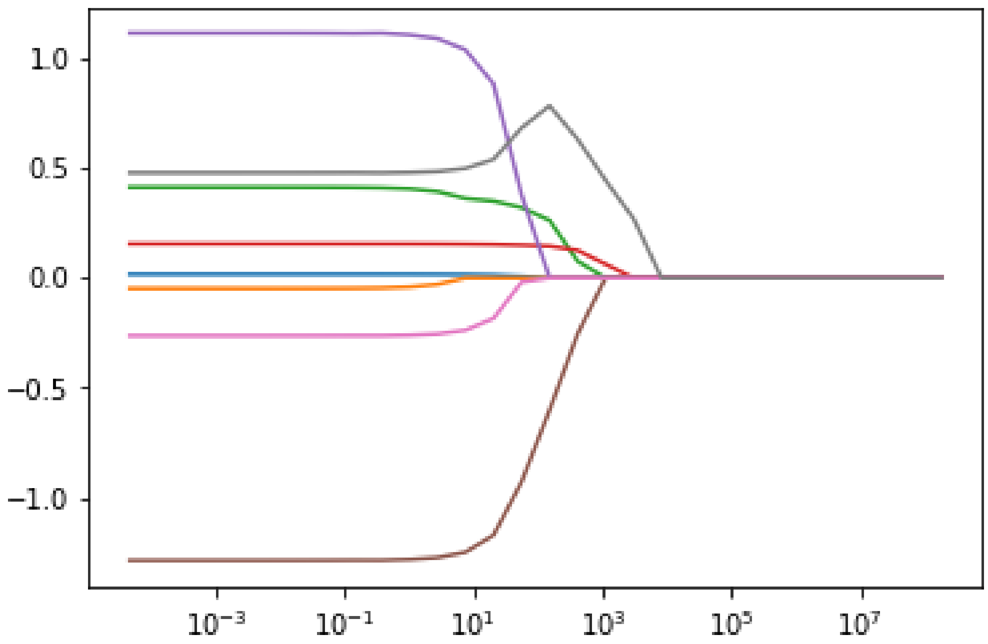
- Lasso回归模型（The Least Absolute Shrinkage and Selection Operator），在平方误差的基础上增加L1正则：

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{i=1}^{n}|\theta_j|$$

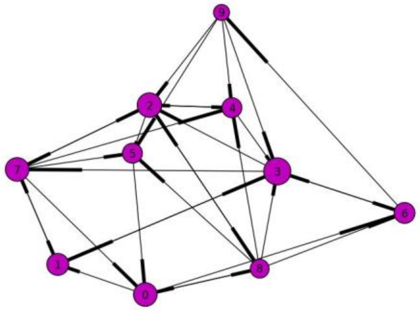
其中，$\lambda>0$。通过确定λ值可以使得在方差和偏差之间达到平衡，随着λ值的增大，模型方差减小而偏差增大

Lasso Regression

- Lasso回归模型的系数选择



弹性网（Elastic Net）



Elastic Net

CDA数据分析师 CERTIFIED DATA ANALYST

- ElasticNet是一种使用L1和L2作为正则化矩阵的线性回归模型，是岭回归和Lasso回归的组合

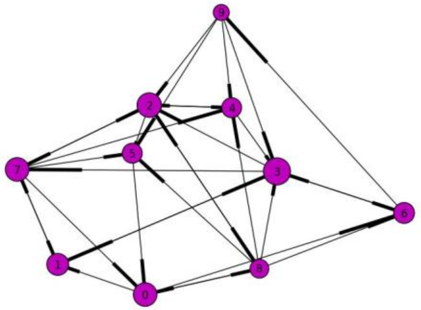
$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\lambda_1\sum_{i=1}^{n}|\theta_j|+\lambda_2\sum_{i=1}^{n}\theta_j^2$$

- 解决模型训练过程中的过拟合问题，弹性网结合了岭回归的计算精准的优点，同时又结合了LASSO回归特征选择的优势

Elastic Net

- 三种方法的选择：
  - ✓ 岭回归优缺点：计算精准（没有丢失特征），不具备特征选择功能
  - ✓ Lasso回归优缺点： 具有特征选择功能，但是将某些θ化为0，可能导致特征丢失，使得最终的模型的偏差比较大
  - ✓ 在进行正则化的过程中，通常要先使用岭回归
  - ✓ 当特征非常多时，尤其当多个特征和另一个特征相关的时弹性网络非常有用

scikit-learn中的线性回归



scikit-learn中的线性回归

- 标准线性回归
- 带交叉验证的线性回归
- 利用网格搜索进行高效参数调优

Talk is cheap Show me the CODE