

正则化项

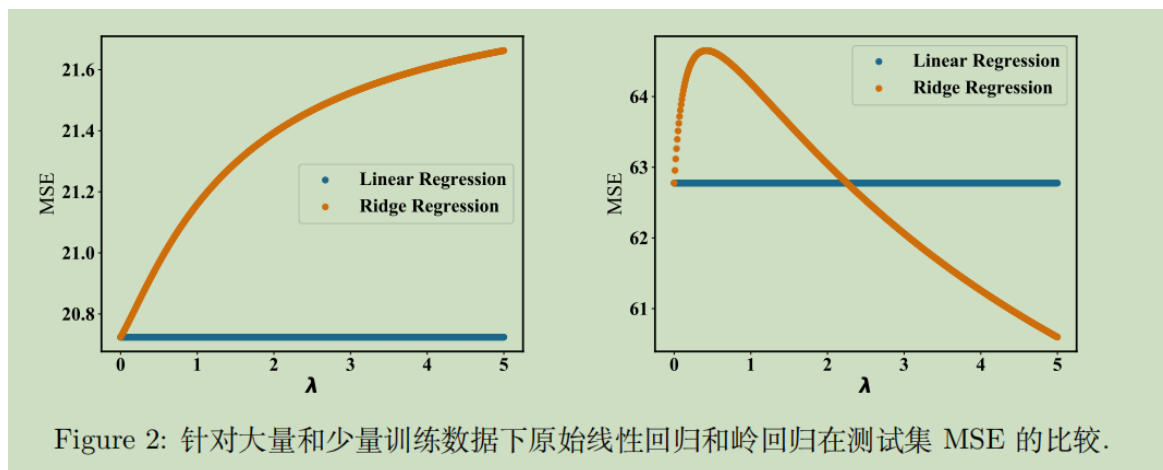
- 实际问题中常常会遇到示例较少，而特征很多的场景。
- 此时如果直接求解线性回归模型，较少的示例无法获得唯一的模型参数，会具有多个模型能够“完美”拟合训练集中的所有样例
- 此外，模型很容易过拟合。

为缓解这些问题，常在线性回归的闭式解中引入正则化项 $\Omega(w)$

- 引入归纳偏好：对模型的一种偏好，例如 $\Omega(w)$
- 增大正则项参数 λ 缓解过拟合
- 一般对模型的复杂度进行约束，相当于从多个训练集上表现类似的模型中选出复杂度最低的一个
- 使用 L1 范数正则化能够使权重 w 的元素稀疏（大量元素0）
- L2使 w 分量取值均衡，即非零分量个数尽量稠密
- L0和L1使 w 分量尽量稀疏，即非零分量个数尽量少

岭回归

训练样例少时，将 λ 设置较大可以使MSE小于原始方法



PCA降维第一步必须标准化

- 防止投影后数值大小对特征重要性的影响
- 有利于梯度下降法的收敛。

为什么要比较检验

- 希望比较泛化性能，但评估只能得到测试性能，所以评估结果不能用于判断优劣
- 测试性能与测试集的选择有很大关系

- 很多ML算法有一定随机性，相同测试集结果也不一样

为什么核函数

- 只要原始空间维数有限，一定存在高维空间使线性可分
- 高维不好算，核函数换元