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 Ω(w)

- 引入归纳偏好：对模型的一种偏好, 例如 Ω(w)
- 增大正则项参数 lambda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

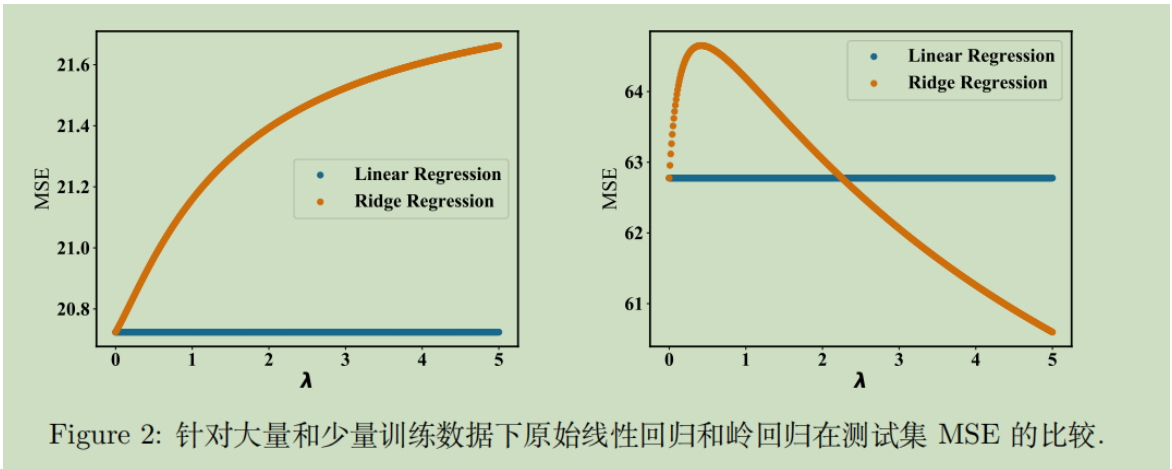

Figure 2: 针对大量和少量训练数据下原始线性回归和岭回归在测试集 MSE 的比较.

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

为什么选择交叉熵作为损失函数

> 交叉熵主要用于度量两个概率分布间的差异性信息

真实的数据分布是不可知的，只能用所获得的部分训练数据近似的代替真实数据的数据分布
衡量两种概率分布之间的差异性非常可行的方法就是用计算它们的相对熵
等价于当交叉熵最低时，我们学到了“最好的模型”。

对相对熵KL散度的理解

> 相对熵可以理解为两个随机变量之间的距离
> 常用p(x)表示label，q(x)表示predict值
> p(x)=[1，0，0]表示该样本属于第一类，q(x)=[0.9，0.3，0.2]表示预测该样本有0.9的概率属于第一类

$$D_{KL}(p||q) = \sum p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

其中熵的定义为：$H(x) = \sum p(x_i) log(p(x_i))$

显然$D_{KL}(p||q) = -H(x) + (-\sum p(x_i) log(q(x_i)))$

我们将公式的后半部分称之为交叉熵定义为：$H(p,q) = -\sum p(x_i) log(q(x_i))$