机器学习在化学中的应用：上机作业 01

王崇斌 1800011716

2021 年 10 月 11 日

1 背景介绍：岭回归

对于给定的一组数据 $(\mathbf{X}, \mathbf{t})$，我们希望用一个有着准确定义的模型 $f(\mathbf{w}, \mathbf{x})$ 去描述数据的规律和预言将要产生的结果，同时需要一个标准 $E(\mathbf{w}, \mathbf{x}, \mathbf{y})$ 去衡量模型的好坏。（与机器学习的定义还差一个提升精确度的手段）那么需要定义清楚两个概念：1. 模型如何设计；2. 衡量误差的标准如何制定。下面我们简要讨论一下线性模型与损失函数。

线性基函数模型 假设我们的模型可以用如下所线性组合形式的函数来描述：

$$f(\mathbf{w}, \mathbf{x}) = \sum_{k=0}^{M} w_k \phi_k(\mathbf{x}) = \mathbf{w}^t\ (\mathbf{x}) \tag{1}$$

最小二乘法 对于给定数据集 $(\mathbf{X}, \mathbf{t})$ 与模型 $f(\mathbf{w}, \mathbf{x})$，可以定义最小二乘的误差函数：

$$E_D(\mathbf{w}, \mathbf{X}, \mathbf{t}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - f(\mathbf{w}, \mathbf{x}_n))^2 \tag{2}$$

对于线性模型，上述损失函数可以写为：

$$E_D(\mathbf{w}, \mathbf{X}, \mathbf{t}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^t\ (\mathbf{x}_n))^2 \tag{3}$$

可以从多种角度来理解上述误差函数，最直观的是它描述了数据点距离模型曲线的距离（有点类似与函数空间的 L_2 范数）。更重要的是可以从概率论的角度来理解。假设数据是通过实验获得的，实验本身会给数据引入随机误差，假设对于给定的 $\mathbf{x}$，测量值 t 作为一个随机变量其概率密度为均值 $f(\mathbf{w}, \mathbf{x})$，方差为 β^{-1} 的正态分布，同时，假设对于不同的 $\mathbf{x}$，随机变量 t 是独立的，那么在给定 $(\mathbf{x}_1, ..., \mathbf{x}_N)$ 的情况下，$(t_1, ..., t_n)$ 的联合分布的密度函数可以表示为：

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | f(\mathbf{w}, \mathbf{x}_n), \beta^{-1}) \tag{4}$$

那么在 $\mathbf{t}$ 给定的情况下，通过最大化上述概率密度函数，可以给出一种确定参数 $\mathbf{w}$ 的方法，容易证明这将等价于最小二乘函数的最小化。对于线性模型，可以进一步化简，将上述最小化的问题转化为求解线性方程组的问题。

正则化的最小二乘法 有时候我们不希望模型中的参数 $\mathbf{w}$ 太大，参数太大往往是过拟合的一个特征（同时也会引入更大的数值不稳定性），可以将损失函数定义为（其中 λ 是一个给定的参数）：

$$E_r(\mathbf{w}, \mathbf{X}, \mathbf{t}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^t\ (\mathbf{x}_n))^2 + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w} \tag{5}$$

如果最小化这个损失函数，同样也可以将最小化问题转化为求解线性方程组的问题（这也是线性模型的一个特点），进一步得到参数 λ。同样的，这个损失函数也可以通过最大化后验概率（假设 $\mathbf{w}$ 的先验概率是高斯分布）的方法来获取，这里不详细展示具体过程。

2 作业部分

(1) ：以多项式为基函数时，在训练集和测试集上的均方误差随 λ 的关系、拟合参数随 λ 的关系如表 1 所示。

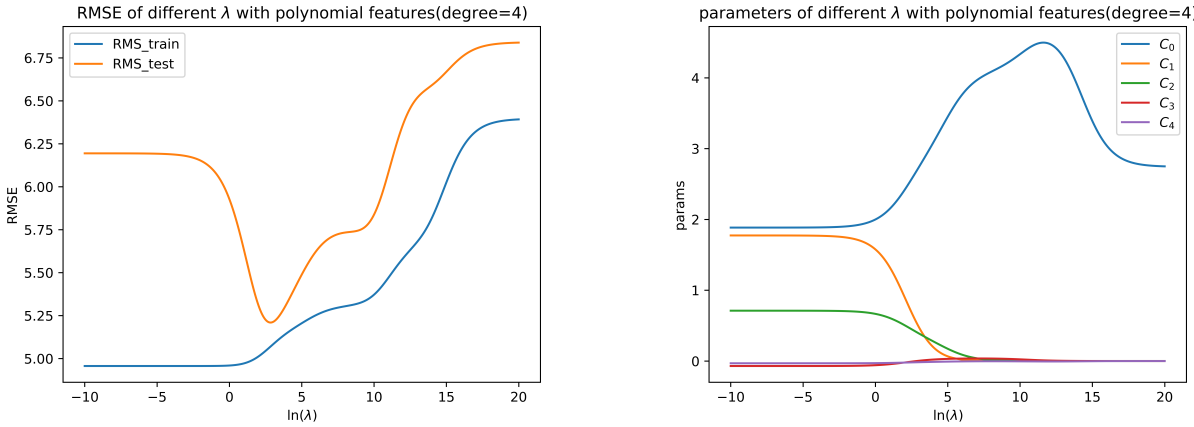

表 1: 多项式为基函数时均方误差与拟合参数随超参数的关系

(2) ：从表 1 可以看出，当 $\ln(\lambda) = 2.5$ 左右时，可以在训练集和测试集上都得到一个比较小的均方误差，选定这个参数。得到表 2 的数据 同时给出拟合图像图 1。

表 2: 最佳超参数下的多项式拟合数据

RMSE-train	RMSE-test	C_0	C_1	C_2	C_3	C_4
5.0445	5.2261	2.5690	0.6870	0.4498	-0.0081	-0.0192

(3) ：以余弦多项式为基函数时，在训练集和测试集上的均方误差随 λ 的关系、拟合参数随 λ 的关系如表 3 所示。

(4) ：从表 3 可以看出，当 $\ln(\lambda) = -0.5$ 左右时，可以在训练集和测试集上都得到一个比较小的均方误差，选定这个参数。得到表 4 的数据: 同时给出拟合图像图 2。

(5) ：将给定的拟合函数的 x 移项到等式左边，然后可以同样用余弦多项式拟合。在训练集和测试集上的均方误差随 λ 的关系、拟合参数随 λ 的关系如表 5 所示。从表 5 可以看出，当 $\ln(\lambda) = -0.5$ 左右时，可以在训练集和测试集上都得到一个比较小的均方误差，选定这个参数。得到表 6 的数据: 同时给出拟合图像图 3。

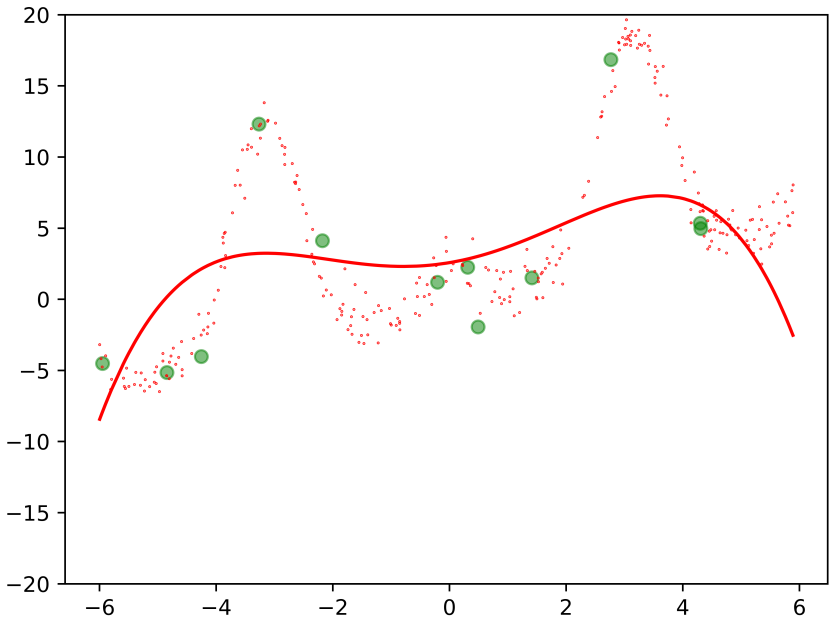

图 1: 多项式拟合曲线

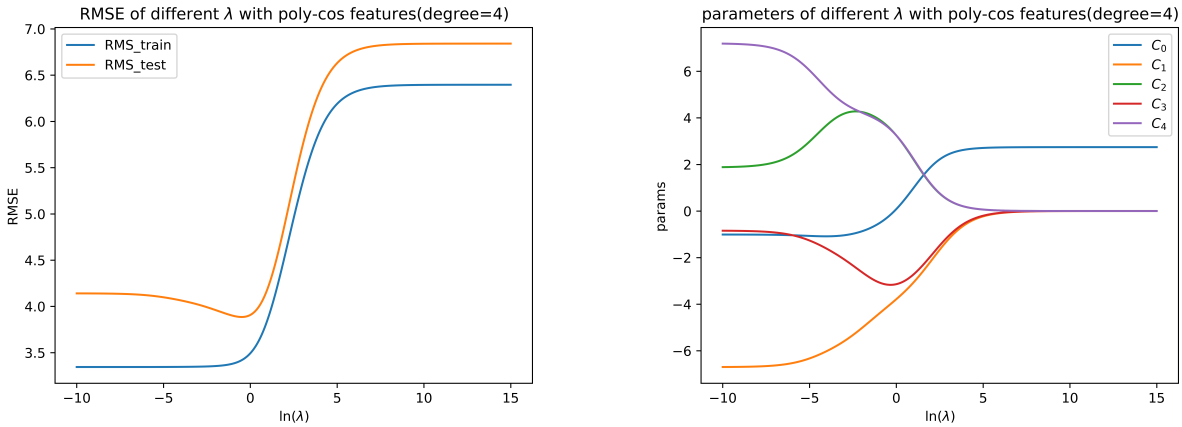

表 3: 余弦多项式为基函数时均方误差与拟合参数随超参数的关系

表 4: 最佳超参数下的余弦多项式拟合数据

RMSE-train	RMSE-test	C_0	C_1	C_2	C_3	C_4
3.4184	3.8854	-0.2650	-4.1013	3.6405	-3.1622	3.6111

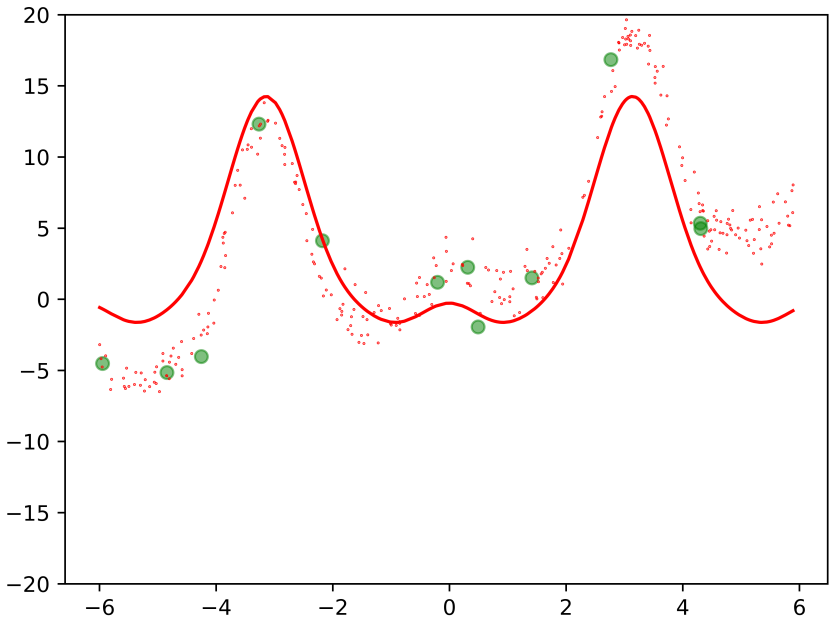

图 2: 余弦多项式拟合曲线

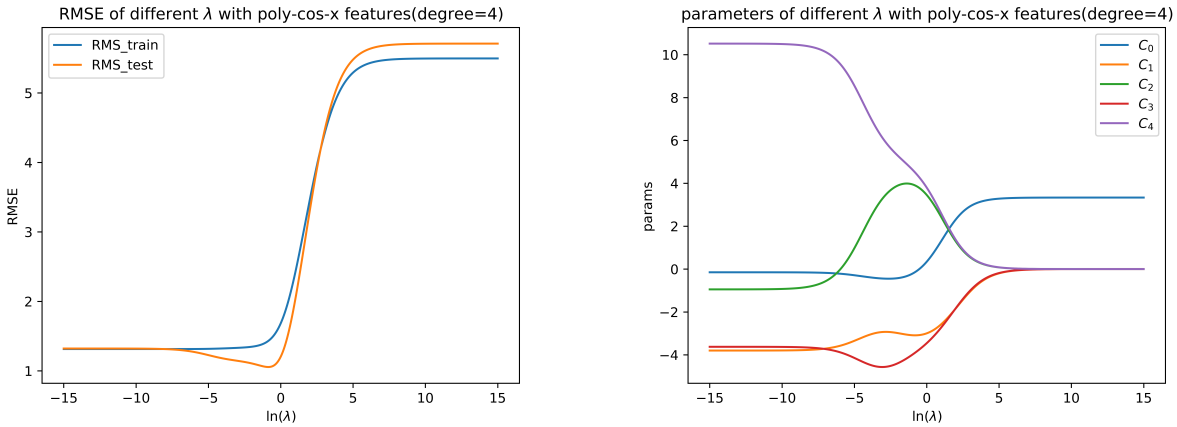

表 5: 余弦多项式 + 线性函数为基函数时均方误差与拟合参数随超参数的关系

表 6: 最佳超参数下的余弦多项式拟合数据

RMSE-train	RMSE-test	C_0	C_1	C_2	C_3	C_4
1.5036	1.0738	0.0075	-3.0780	3.7790	-3.7457	4.2670

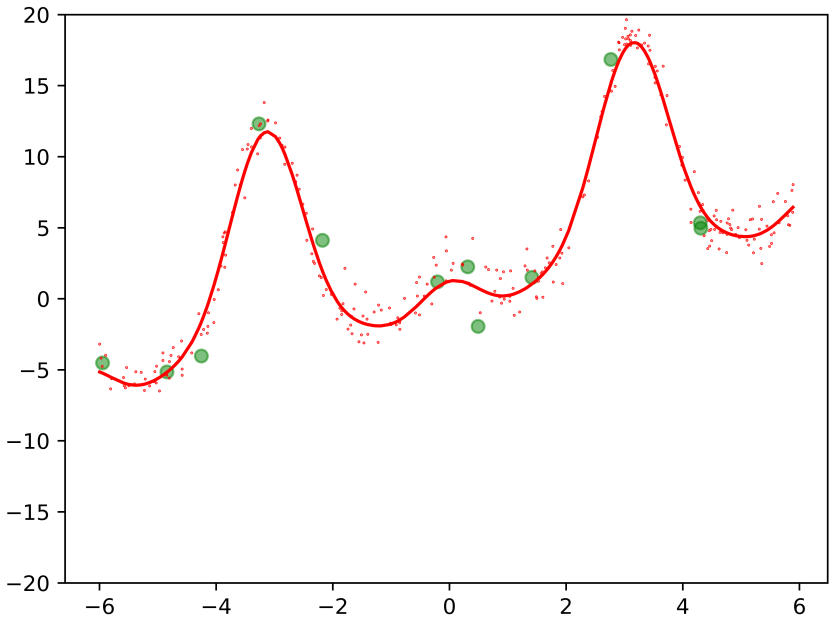

图 3: 余弦多项式 + 线性函数拟合曲线