泛化误差分析

关于泛化误差也就是在总体样本上的测试误差，但是我们无法得到全部样本，所以只能通过有限的样本进行估计；

假设存在训练数据集 $D_{tr}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_i,y_i),\ldots,(\mathbf{x}_N,y_N)\}$，并假设该数据集 $\mathbf{x}$ 和 y 存在如下潜在的真实模型($\mathbf{x}$ 和 y 的对应关系)：

$$y_i = f_{true}(\boldsymbol{x}_i) + \varepsilon_i \quad (1)$$

其中ε_i是独立同分布均值为 0，方差为 σ^2 的噪声。例如：$f_{true}(x) = \sin(x)$，那么我们可以得到如图 2 所示的训练数据 D_{tr}。那么根据这些数据我们采用多项式回归来估计数据间的模型（函数关系）$f_{ls}(\mathbf{x})$, 脚标 *ls* 表示采用最小二乘（least square method）优化求解，即通过求解公式(2)的目标函数学习得到回归函数 $f_{ls}(\mathbf{x})$：

$$\min_{\boldsymbol{w}} trErr(\boldsymbol{w}, D_{tr}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_{ls}(\mathbf{x}_i))^2 \quad (2)$$

其中多项式回归函数可以表示为：$f_{ls}(\mathbf{x})=\boldsymbol{w}^T\boldsymbol{\phi}(\mathbf{x})$。

假设测试集为测试集 $D_{te}=\{(\mathbf{x}_{*1},y_{*1}),(\mathbf{x}_{*2},y_{*2}),\ldots,(\mathbf{x}_{*i},y_{*i}),\ldots,(\mathbf{x}_{*M},y_{*M})\}$，那么在该测试集上的测试误差为：

$$teErr(f_{ls}, D_{tr}, D_{te}) = \frac{1}{M}\sum_{i=1}^{M}(y_{*i} - f_{ls}(\mathbf{x}_{*i}))^2 \quad (3)$$

公式(3)可以认为是在训练集 D_{tr} 上训练得到的模型 f_{ls} 在测试集 D_{te} 上的误差。但是该测试误差依赖于训练集 D_{tr}，我们可以通过平均在所有样本数为 N 的训练集上学习得到的函数 f_{ls}来降低训练集的依赖，即计算学习得到的函数 f_{ls}的期望泛化误差，在这里 E_D 表示在不同训练数据集 D 上求期望，E_T表示测试集 T 上的期望。

$$\begin{aligned}
Err(f_{ls}, D) &= E_D\left\{E_T\{[f_{ls}(\mathbf{x}; D) - y]^2\}\right\} \\
&= E_D\{E_T[f_{ls}(\mathbf{x}; D) - f_{true}(\mathbf{x}) - \varepsilon]^2\} \\
&= E_D\left\{E_T\{[f_{ls}(\mathbf{x}; D) - f_{true}(\mathbf{x})]^2\}\right\} + \sigma^2 \text{ (为了简化表示后边不在写 } E_T) \\
&= E_D\{\{f_{ls}(\mathbf{x}; D) - E_D[f_{ls}(\mathbf{x}; D)] + E_D[f_{ls}(\mathbf{x}; D)] - f_{true}(\mathbf{x})\}^2\} + \sigma^2 \\
&= E_D\{\{f_{ls}(\mathbf{x}; D) - E_D[f_{ls}(\mathbf{x}; D)]\}^2\} + E_D\{\{E_D[f_{ls}(\mathbf{x}; D)] - f_{true}(\mathbf{x})\}^2\} \\
&\quad + 2E_D\{f_{ls}(\mathbf{x}; D) - E_D[f_{ls}(\mathbf{x}; D)]\}E_D\{E_D[f_{ls}(\mathbf{x}; D)] - f_{true}(\mathbf{x})\} + \sigma^2 \\
&= E_D\{\{f_{ls}(\mathbf{x}; D) - E_D[f_{ls}(\mathbf{x}; D)]\}^2\} + E_D\{\{E_D[f_{ls}(\mathbf{x}; D)] - f_{true}(\mathbf{x})\}^2\} + \sigma^2 \\
&= \mathrm{var}(f_{ls}(\mathbf{x}; D)) + bias^2\left(f_{ls}(\mathbf{x}; D)\right) + \sigma^2
\end{aligned}$$

其中$f_{ls}(\mathbf{x};D)$表示在训练集 D 上学习得到的函数，$\mathrm{E_D}\{[f_{ls}(\mathbf{x};D)-y]^2\}$表示对不同训练集上得到的$f_{ls}(\mathbf{x};D)$的期望的泛化误差。通过上式，我们将估计的回归模型 f_{ls} 期望的泛化误差分解为:模型的期望输出与真实标记的差别（称作偏置或偏差）$\mathrm{bias}^2(f_{ls}(\mathbf{x};D))$, 和使用样本数相同的不同训练集产生的不同回归模型 f_{ls} 的方差。

为了展示上述泛化误差分析中的偏置和方差，我们继续以潜在的真实模型为 $f_{true}(\mathrm{x})=\sin(\mathrm{x})$为例，假设我们有 50 个训练数据集，每个训练集包含 20 个样本和一个包含 20 个样本的测试集；其中训练集和测试集采用相同的生成方式 $\mathrm{y}=\sin(\mathrm{x})+\varepsilon$，$\varepsilon$服从均值为 0，标准差 0.2 的高斯分布，$f_{true}=\sin(\mathrm{x})$；采用 n 阶多项式回归，$f_{ls}(\mathbf{x},\boldsymbol{w})$:

$$f_{ls}(\mathbf{x},\boldsymbol{w})=w_0+\sum_{k=0}^{n} w_k x^k \tag{4}$$

这样，在每个训练集上都可以得到一个回归函数，我们可以尝试将该多项式回归函数从 0 阶到 N 阶进行计算，从而观察训练误差、测试误差，偏置和方差，具体计算过程相见 MATLAB 代码，运行结果如下图 1 所示，这里纵坐标采用 $\log_2(\mathrm{MSE}+1)$是为了便于展示。通过结果可以看到，随着多项式阶数的增加（对应模型复杂度增加），训练误差在减小，但是测试误差增加，同时 Bias^2 减少，模型预测的方差增加；通过该图，我们可以看到当多项式阶数等于 3，回归函数具有较好的估计性能和稳定性。图 2，展示了当多项式阶数为 1,2,3,..., 8 时，得到的多项式回归函数和训练数据的关系，我们可以看到，高阶多项式尽可能的逼近每个训练数据（训练数据本身是有噪声的），因此，可以等效的看成随着多项式回归阶数的增加，实际该回归模型开始越来越逼近噪声数据。

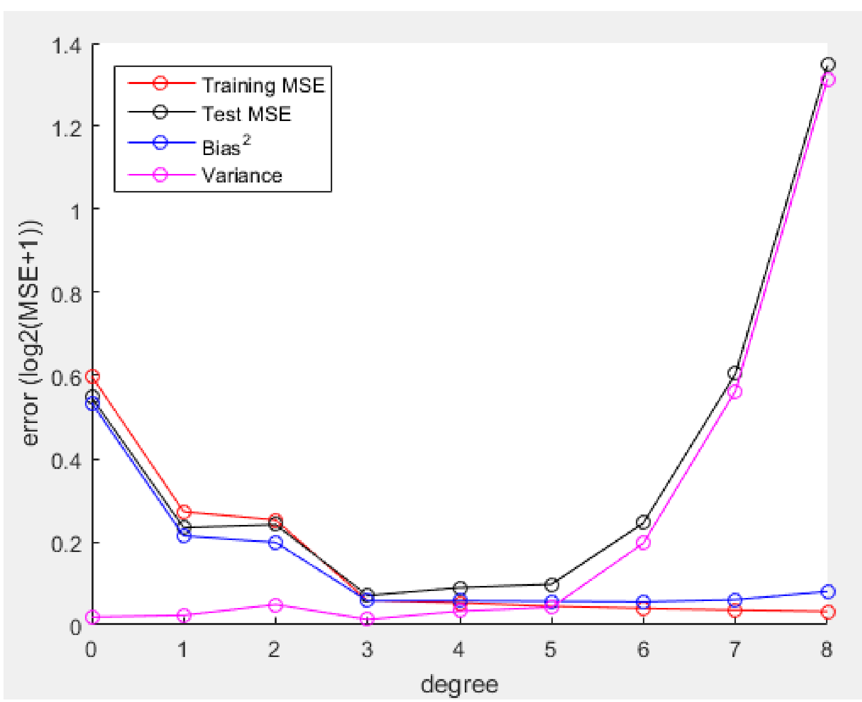

图 1：训练误差，测试误差，偏置和方差

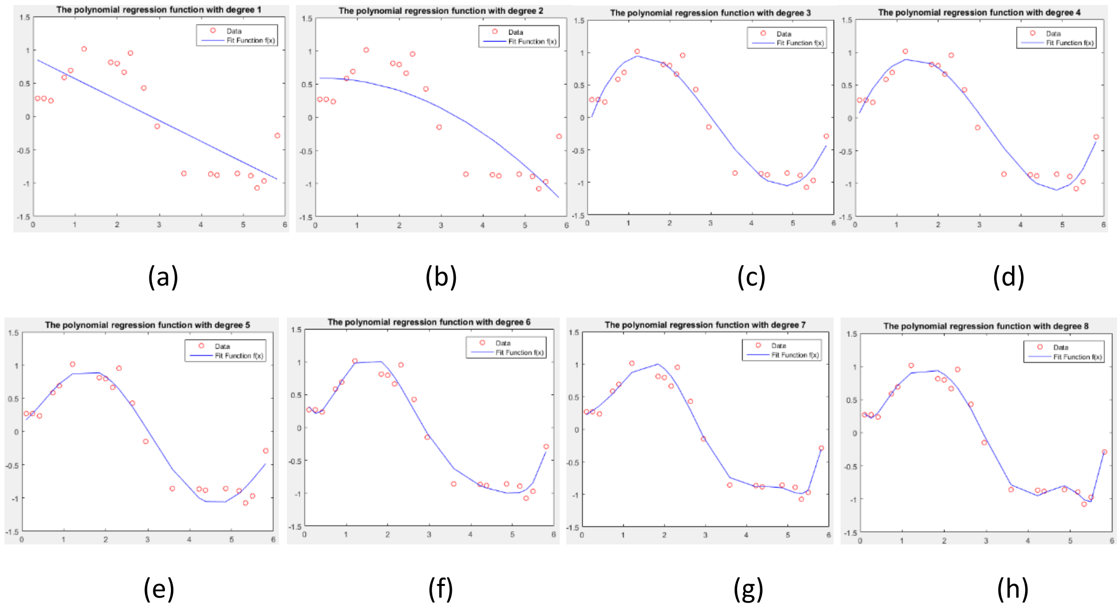

图 2：不同阶次的多项式回归函数在和训练数据，从(a)到(h)对应了多项式阶次从 1 到 8 的训练数据和估计的函数 f_{ls}(x)