模式识别作业1

张蔚桐 2015011493 自55

2017 年 3 月 6 日

1

根据线性回归方程可以得到

$$\hat{y} = \hat{\theta_0} + \hat{\theta_1} x$$
$$\hat{\theta_0} = \bar{y} - \hat{\theta_1}\bar{x}$$
$$\hat{\theta_1} = \frac{E((y - E(y))(x - E(x)))}{D(x)}$$

其中$D(x), E(x)$分别为样本的方差和期望, 因此可以得到

$$R^2 = \frac{E((\bar{y} - \hat{y})^2)}{D(y)}$$
$$r^2 = \frac{E^2((y - E(y))(x - E(x)))}{D(x)D(y)}$$

欲证$R^2 = r^2$即证明

$$E((\bar{y} - \hat{y})^2) = \frac{E^2((y - E(y))(x - E(x)))}{D(x)}$$

左侧$= E(\hat{\theta_1}^2(x - \bar{x})^2) = \hat{\theta_1}^2 D(x) = \frac{E^2((y-E(y))(x-E(x)))}{D(x)}$ =右侧

进而得证$R^2 = r^2$

2

2.1 前三题

如图1所示，为训练集为10个样本点时的线性拟合和各阶过拟合的情况。图中r为拟合的相关系数，R为采用新的100组数据得到的方均根。第一行为$\sigma = 0.5$时的情况，可以看出，这个时候因为数据集的线性性比较好，尽管出现了过拟合但是在局部差别不大。尽管如此，仍可以看出随着阶数

的增加，对样本的拟合系数不断提高。但是对于测试集的误差方均根同样不断上升。

这种情况在$\sigma = 2$的第二行中更加明显。此时高阶拟合出现很大的误差。尽管拟合系数相对于较低阶的拟合（如线性拟合）很高，但是在测试集中表现了很大的误差方均根。

2.2 后两题

如图2所示，为训练集为100个样本点的各阶拟合的情况。r, R同上节所示。经过测试总结，可以看出如下关系。

随着σ增大，训练集的拟合程度r^2减小，测试集的误差系数R^2增大。

随着模型复杂程度的增大。训练集的拟合程度r^2增大，测试集的误差系数R^2增大，甚至出现过拟合。

随着训练样本的增大，训练集的拟合程度r^2增大并趋近1，测试集的误差系数R^2减小。

2 3

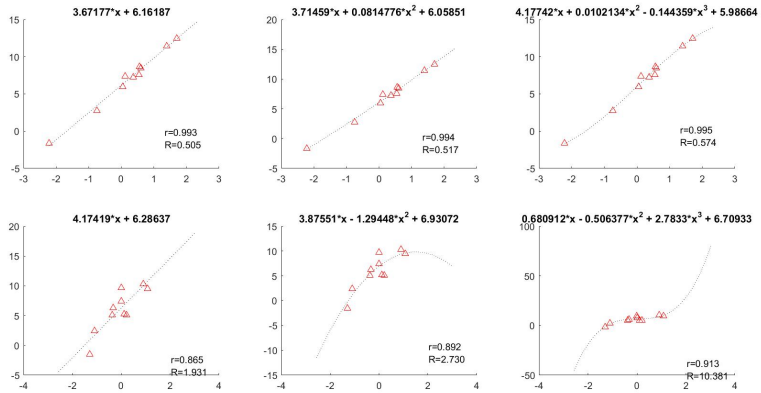

图 1: 10个样本点的情况

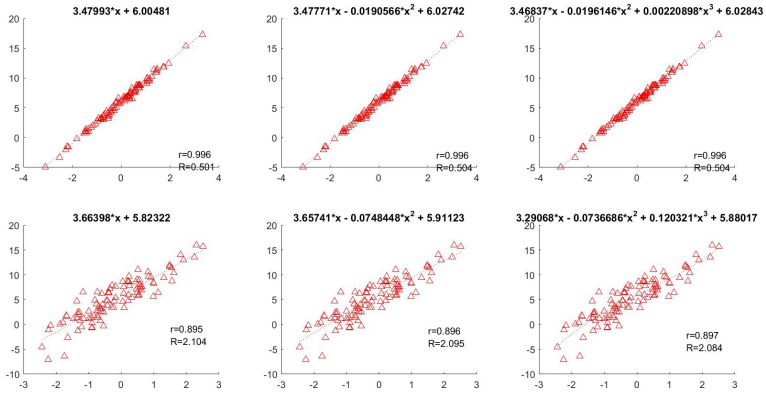

图 2: 100个样本点的情况