模式识别作业：线性回归

贾成君 2014011552

1. 证明：

代入可得R^2和r^2,

即：

1. 解：

整理测试结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 训练数 | 训练结果 | 训练 | 测试 |
| 0.5 | 10 |  | 0.980845 | 1.0222 |
|  | 0.981191 | 1.0448 |
|  | 0.981192 | 1.0380 |
| 训练图形 | 测试图形 | |
|  |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 训练数 | 训练结果 | 训练 | 测试 |
| 0.5 | 100 |  | 0.981888 | 0.923517 |
|  | 0.981907 | 0.926000 |
|  | 0.981911 | 0.925682 |
| 训练图形 | 测试图形 | |
|  |  | |
|  | | | | |
|  | 训练数 | 训练结果 | 训练 | 测试 |
| 2 | 10 |  | 0.4636 | 0.9821 |
|  | 0.4739 | 1.1043 |
|  | 0.5264 | 18.6198 |
| 训练图形 | 测试图形 | |
|  |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 训练数 | 训练结果 | 训练 | 测试 |
| 2 | 100 |  | 0.72253 | 0.7335 |
|  | 0.72396 | 0.7598 |
|  | 0.72397 | 0.7538 |
| 训练图形 | 测试图形 | |
|  |  | |

**计算测试使用的方法为按照**，因为使用测试样本计算过程中不满足估计的y与实际的y无偏，所以计算出的可能大于1，一般来说，此时计算出的结果距离1越大说明其估计效果越差。

通过数据的对比，我们可以看出，**误差的离散程度越小()、模型复杂度越高（阶次越高、参数越多）、训练样本越多，训练得到的越大，对于训练样本的拟合程度越好**。而当误**差的离散程度越小、训练样本越多，测试得到的越大，对于测试样本的拟合程度越好**；**但是模型复杂度与测试数据的拟合效果并不是单调的关系**，上述四种数据都是使用一元二次模型拟合得到的对于测试数据的符合度最好。当模型不太复杂时增加模型的复杂度可以增加拟合效果，而当模型国语复杂时，拟合效果反而变差（过拟合）。

代码如下：

%指定各个参数

theta1 **=** 3.5**;** theta0 **=** 6**;** sigma0 **=** 2**;** num **=** 100**;**

%生成训练样本

x **=** **(**randn**(**1**,**num**))';** sigma **=** **(**randn**(**1**,**num**)\***sigma0**)';**

y **=** x**\***theta1**+**theta0**+**sigma**;**

%线性回归

y0 **=** y**;**

x0 **=** **[**x ones**(**num**,**1**)];**

theta **=** regress**(**y0**,**x0**);**

t11 **=** theta**(**1**);**t10 **=** theta**(**2**);**

%一元二次回归

x0 **=** **[**x **.^**2 x0**];**

theta **=** regress**(**y0**,**x0**);**

t22 **=** theta**(**1**);** t21 **=** theta**(**2**);** t20 **=** theta**(**3**);**

%一元三次回归

x0 **=** **[**x **.^**3 x0**];**

theta **=** regress**(**y0**,**x0**);**

t33 **=** theta**(**1**);** t32 **=** theta**(**2**);** t31 **=** theta**(**3**);** t30 **=** theta**(**4**);**

%显示训练后的参数

disp**(**'参数[从高次到低]：'**),**

disp**([**t11 t10**]),**disp**([**t22 t21 t20**]),**disp**([**t33 t32 t31 t30**])**

%显示训练结果

xin **=** x**;**

yin **=** y**;**

yout1 **=** t11**\***xin**+**t10**;**

yout2 **=** t22**\*** xin **.^**2**+**t21**\***xin**+**t20**;**

yout3 **=** t33**\***xin **.^**3**+**t32**\***xin **.^**2**+**t31**\***xin**+**t30**;**

disp**(**'训练R^2'**),**

ybar **=** mean**(**yin**);** yall **=** sum**(** **(**ybar**-**yin**)** **.^**2**);**

disp**(**sum**((**ybar**-**yout1**)** **.^**2**)/**yall**),**

disp**(**sum**((**ybar**-**yout2**)** **.^**2**)/**yall**),**

disp**(**sum**((**ybar**-**yout3**)** **.^**2**)/**yall**)**

%显示训练图形

figure**,**scatter**(**xin**,**yin**),**hold on**,**

xshow **=** min**(**x**):**0.01**:**max**(**x**);**

xin **=** xshow**;**

yout1 **=** t11**\***xin**+**t10**;**

yout2 **=** t22**\*** xin **.^**2**+**t21**\***xin**+**t20**;**

yout3 **=** t33**\***xin **.^**3**+**t32**\***xin **.^**2**+**t31**\***xin**+**t30**;**

plot**(**xin**,**yout1**),**plot**(**xin**,**yout2**),**plot**(**xin**,**yout3**),**

title**(**'训练'**),**legend**(**'训练数据'**,**'一次拟合'**,**'二次拟合'**,**'三次拟合'**,**'Location'**,**'southeast'**),**grid on**;**

%生成测试数据

testx **=** randn**(**100**,**1**);**

testsigma **=** randn**(**100**,**1**)\***sigma0**;**

testy **=** testx**\***theta1**+**theta0**+**testsigma**;**

%比较预测结果和实际结果

xin **=** testx**;**

yin **=** testy**;**

yout1 **=** t11**\***xin**+**t10**;**

yout2 **=** t22**\*** xin **.^**2**+**t21**\***xin**+**t20**;**

yout3 **=** t33**\***xin **.^**3**+**t32**\***xin **.^**2**+**t31**\***xin**+**t30**;**

disp**(**'实际R^2'**),**

ybar **=** mean**(**yin**);** yall **=** sum**(** **(**ybar**-**yin**)** **.^**2**);**

disp**(**sum**((**ybar**-**yout1**)** **.^**2**)/**yall**),**

disp**(**sum**((**ybar**-**yout2**)** **.^**2**)/**yall**),**

disp**(**sum**((**ybar**-**yout3**)** **.^**2**)/**yall**)**

figure**,**scatter**(**xin**,**yin**),**hold on**,**

xshow **=** min**(**testx**):**0.01**:**max**(**testx**);**

xin **=** xshow**;**

yout1 **=** t11**\***xin**+**t10**;**

yout2 **=** t22**\*** xin **.^**2**+**t21**\***xin**+**t20**;**

yout3 **=** t33**\***xin **.^**3**+**t32**\***xin **.^**2**+**t31**\***xin**+**t30**;**

plot**(**xin**,**yout1**),**plot**(**xin**,**yout2**),**plot**(**xin**,**yout3**),**

title**(**'测试'**),**legend**(**'测试数据'**,**'一次拟合'**,**'二次拟合'**,**'三次拟合'**,**'Location'**,**'southeast'**),**grid on**;**

1. 解：
2. 使用regress进行多元线性回归，可得7. 代入计算要求的女性生存时间，可得约为220天
3. 如果考虑变量的交叉项，可以对已有的样本的拟合程度改善。如：考虑x1\*x2的交叉项，可得各组基1,x1,x2,x3, x1\*x2的系数为2060.82，-4.79，-973.51，52.72， 2.45，根据拟合结果来看，（1）的决定系数为0.23，此式的决定系数为0.58，明显改善。但是使用这些参数预测数据可得生存天数为-525<0，显然不合常理，预测效果变差了。同理，考虑其他交叉项发现，增加交叉项后的结果要么没有显性增加决定系数，要么是对（1）的预测结果出现负数，从这种意义上，增加交叉项并没有达到更好的预测效果。尝试去掉一些交叉项，也没有明显改善预测效果。

代码如下

%读取文件

FileX **=** 'C:\Users\贾成君\Desktop\线性回归\X.txt'**;**

FileY **=** 'C:\Users\贾成君\Desktop\线性回归\Y.txt'**;**

load**(**FileX**),**load**(**FileY**);**

x0 **=** **[**ones**(**size**(**X**,**1**),**1**)** X**];**

y0 **=** Y**;**

%进行回归和预测

variable **=** regress**(**y0**,**x0**);**

variable

**[**1 110 3 1**]\***variable

%计算决定系数

yp **=** x0 **\*** variable**;**

ybar **=** mean**(**y0**);**

corr1 **=** sum**(** **(**yp**-**ybar**)** **.^**2**)/**sum**(** **(**y0**-**ybar**)** **.^**2**);**

%考虑交叉项

x3 **=** **[**ones**(**size**(**X**,**1**),**1**)** X X**(:,**1**).\***X**(:,**2**)];**

var3 **=** regress**(**y0**,**x3**);**

yp3 **=** x3 **\*** var3**;**

corr3 **=** sum**(** **(**yp3**-**ybar**)** **.^**2**)/**sum**(** **(**y0**-**ybar**)** **.^**2**);**

**[**1 110 3 1 110**\***3**]** **\*** var3

%去掉一些项

x4 **=** **[**ones**(**size**(**X**,**1**),**1**)** X**(:,**2**)** X**(:,**3**)** X**(:,**1**).\***X**(:,**2**)** X**(:,**1**).\***X**(:,**3**)];**

var4 **=** regress**(**y0**,**x4**);**

yp4 **=** x4 **\*** var4**;**

corr4 **=** sum**(** **(**yp4**-**ybar**)** **.^**2**)/**sum**(** **(**y0**-**ybar**)** **.^**2**);**

**[**1 3 1 110**\***3 110**\***1**]** **\*** var4