多种模型评价波士顿房价

摘要

我们经过查阅资料，选定X14（户主拥有住房价值的中位数）这一变量为房价，即因变量。

针对问题一：我们选择使用了皮尔逊相关性分析的方法，我们对十四个变量每两个的相关系数进行了逐一求解，以矩阵图形的形式得出了变量之间的相关系数，并且主要考虑的是选定户主拥有住房价值的中位数这一变量与其他13个变量的相关性，然后根据得出的相关性结果进行有选择的回归拟合度的检测与分析，通过多次对不同变量的拟合度的分析，得出最终的结果，既变量对于波士顿房价的影响程度，以及用多少变量进行拟合波士顿房价的回归方程，以便于预测未来房价以及提高结果的准确度。最终我们得出，无论是使用六个相关较强变量还是八个相关较强变量抑或是十三个变量来进行预测，我们可以发现，利用十三个变量拥有最好的效果。

针对问题二：使用BP神经网络建模进行预测。将所给的506个地区的观察值划分为训练样本（前496行）和预测样本（后10行）两部分。通过不断调参得到相对较优秀的神经网络模型。在训练样本中，模拟值和真实值的相关性R高达0.98717，均方误差 MSE低至0.004315；在预测样本部分，预测值与真实值的均方误差MSE = 18.4096，由于量纲的影响，这个值是可以接受的。

针对问题三：主要解决按照宜居性对波士顿的房子进行分类。文章首先选择所需数据，其次通过秩和比（RSR）法对数据进行处理，绘制出3个递进的表格，最后根据秩和比（RSR）分档排序表按照非常宜居、宜居、较宜居、一般、较不宜居、不宜居、非常不宜居七个档次进行分类排序，划分出七种房子类型。

针对问题四：问题四是要针对不同人群给出买房建议，我们从多个角度出发对人群进行分类，对于人群的分类有很多种分类标准，我们在分析中不一一列出，综合分析人群对于房子的建议，对于购房者而言，对所购买的房子的宜居性要求就是购房者对房子的需求。根据我们之前的数据分析，购房者倾向于选择人均犯罪率低、不靠近河流、非商业占地比例低、空气中氮氧化物浓度低等环境条件较好的房子。周围环境是否安全影和空气等客观环境因素都会影响购房者对于房子的选择。

关键词：多重共线性； BP神经网络； Vissim仿真；模糊数学；Ridit分析。

一、问题重述

问题一重述：问题一想要表达的内容主要是对于这14个变量我们要怎么选择，选择什么样的变量进行拟合会有最好的效果，为什么要这样选择以及这些变量会对波士顿房价的预测产生怎样的影响。

问题二重述：根据题设所给的解释变量构造价预测模型，并通过相关的指标评价预测好坏。

问题三重述：通过将波士顿房子的宜居性进行量化分析或者按某种方法进行分类，并作出相应分析。

问题四重述：问题四是要针对不同人群给出买房建议，根据我们之前的数据分析，购房者倾向于选择人均犯罪率低、不靠近河流、非商业占地比例低、空气中氮氧化物浓度低等环境条件较好的房子。周围环境是否安全影和空气等客观环境因素都会影响购房者对于房子的选择。

二、问题分析

本文主要解决的是根据所给数据，对曼哈顿房价做出各类评价的问题。题设所给的数据是由Harrison and Rubinfeld [1]在1978年收集整理的。我们通过对该论文的分析，确定将X14（户主拥有住房价值的中位数）作为房价的参考值。

2.1问题一的分析

我们经过查阅资料，选定户主拥有住房价值的中位数这一变量为房价，既因变量，首先，我们选择使用了皮尔逊相关性分析的方法，我们对十四个变量每两个的相关系数进行了逐一求解，以矩阵图形的形式得出了变量之间的相关系数，并且主要考虑的是选定户主拥有住房价值的中位数这一变量与其他13个变量的相关性，然后根据得出的相关性结果进行有选择的回归拟合度的检测与分析，通过多次对不同变量的拟合度的分析，得出最终的结果，既变量对于波士顿房价的影响程度，以及用多少变量进行拟合波士顿房价的回归方程，以便于预测未来房价以及提高结果的准确度。

2.2问题二的分析

根据问题二的结论，确定使用全部的13个解释变量来进行房价预测模型的建立。根据这13个解释变量，将前496个数据作为训练样本，将后20个数据作为待预测样本，使用 MATLAB建立神经网络模型。

对预测结果的评价包括两个方面。第一个是综合用均方误和相关性指标观察496个训练样本的拟合效果好坏；第二个是用均方误差指标将20个预测出的值与真实值作比较。

2.3问题三的分析

第三问主要是分析评价、综合评价类问题，看到宜居性首先会想到模糊数学，但是考虑到模糊数学评价法受到主观权重的限制，为了避免结果主观性，故选择了秩和比（RSR）综合评价法。选择方法后考虑到数据较为庞大，故选择七个与宜居性关联度较高的数据类别，将这些数据进行分类整理并汇总，按照秩和比法的评价逻辑顺序进行依次绘制表格，最终根据秩和比法分档标准进行分类。

2.4问题四的分析

问题四是要针对不同人群给出买房建议，我们从多个角度出发对人群进行分类，对于人群的分类有很多种分类标准，我们在分析中不一一列出，综合分析人群对于房子的建议，对于购房者而言，对所购买的房子的宜居性要求就是购房者对房子的需求。根据我们之前的数据分析，购房者倾向于选择人均犯罪率低、不靠近河流、非商业占地比例低、空气中氮氧化物浓度低等环境条件较好的房子。周围环境是否安全影和空气等客观环境因素都会影响购房者对于房子的选择。

三、基本假设

1. 假设第14列数据，即X14（户主拥有住房价值的中位数）作为房价

2. 假设除题中给的14个变量外，其余影响因素皆不考虑

四、符号说明

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X1 | X2 | X3 | X4 | X5 | X6 | X14（Y） |
| 城镇的人均犯罪率 | 大块占地住宅区比例 | 非零售商业占地比例 | 查尔斯河虚拟变量 | 氮氧化物浓度 | 每户平均房间数 | 户主拥有住房价值的中位数（单位：千美元） |
| X7 | X8 | X9 | X10 | X11 | X12 | X13 |
| 1940年前建造的户主所有房比例 | 与五个波士顿劳动力聚集区的加权距离 | 与辐射式公路接近指数 | 每1万美元的全值财产税 | 学生/教师比例 | 1000（B-0.63）（B-0.63）I，B为非洲裔美国人比例 | 低社会地位人口的比例（%） |

1. 模型建立与求解

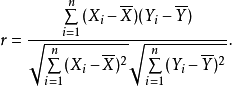
5.1 对问题一的分析

5.1.1数据处理

原始数据较多，将其制成一个506×14 的一个矩阵，其中自变量有13维，然后，通过建立模型来拟合回归出最终的因变量波士顿房价，即户主拥有住房价值的中位数。

5.1.2模型建立与求解

首先，利用spss软件对各个自变量与因变量进行相关系数的检测与分析，并绘制出散点图。所用到的模型为：

 (公式5.1)

得到结果如表5-1所示，表中显示因变量与X13（低社会地位人口的比例（%））的相关系数最高，与X4（查尔斯河虚拟变量）的相关系数最低，其相应的散点图如图5-1所示，但是只有这些数据远远不能够说明各个变量对于波士顿房价的影响程度，为此，我们进行了进一步的数据分析。

表5-1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 房价与各个变量间的相关系数 | | | | | | | | | | | | | |
|  | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 |
| r | -0.388 | 0.36 | -0.484 | 0.175 | -0.427 | 0.695 | -0.377 | 0.25 | -0.382 | -0.469 | -0.508 | 0.333 | -0.738 |

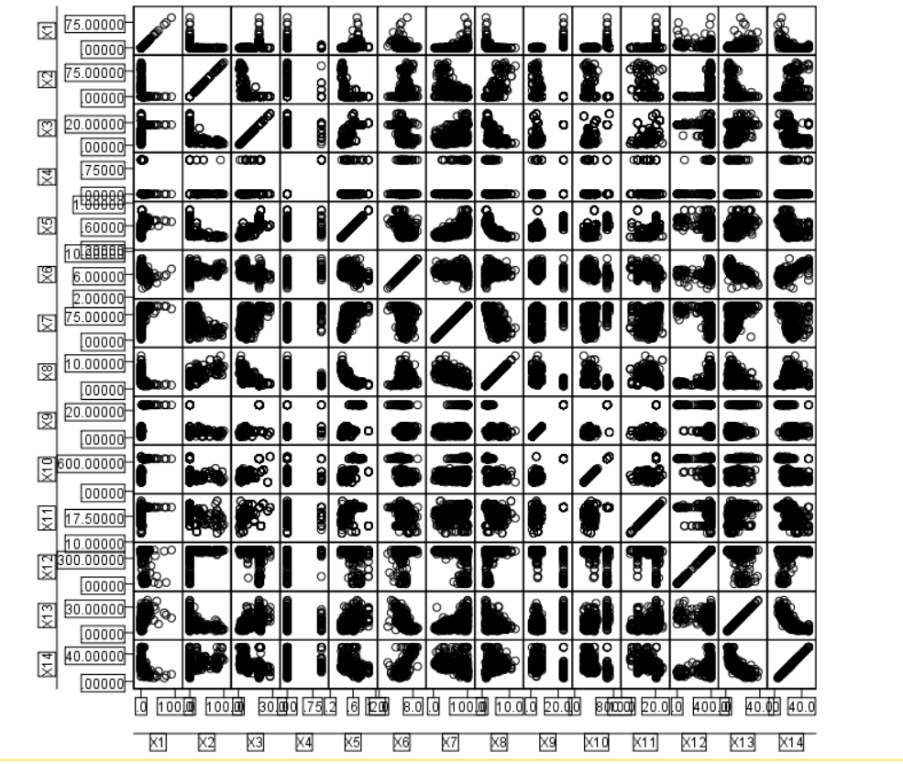
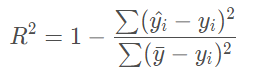


图5-1

为寻找到最优的回归模型，我们分别使用六个相关性较强变量、八个相关性较强和十三个变量分别来对波士顿房价进行预测，并通过计算其拟合度对他们进行对比。如表5-2、5-3、5-4分别为对应的拟合度。相关模型图图5-2。



(公式5.2)

表5-2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **模型摘要** | | | | | | | |
| 模型 | R | R 方 | 调整后 R 方 | 标准估算的误差 | 更改统计 | | |
| R 方变化量 | F 变化量 | 自由度 1 |
| 1 | .826a | .683 | .678 | 5.2171 | .683 | 153.062 | 7 |

|  |
| --- |
| a. 预测变量：(常量), X13, X11, X1, X5, X6, X10, X3 |

表5-3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **模型摘要** | | | | | | | |
| 模型 | R | R 方 | 调整后 R 方 | 标准估算的误差 | 更改统计 | | |
| R 方变化量 | F 变化量 | 自由度 1 |
| 1 | .831a | .690 | .684 | 5.1664 | .690 | 122.705 | 9 |

|  |
| --- |
| a. 预测变量：(常量), X13, X12, X11, X2, X1, X6, X3, X5, X10 |

表5-4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型摘要** | | | | |
| 模型 | R | R 方 | 调整后 R 方 | 标准估算的误差 |
| 1 | 0.861a | 0.741 | 0.734 | 4.7453 |

|  |
| --- |
| a. 预测变量：(常量), X13, X4, X12, X11, X2, X1, X6, X3, X7, X9, X8, X5, X10 |

最后，我们又对其进行了多重共线性的检验，以保证其可以更好的运用最小二乘法进行拟合。如表5-5.

表5-5

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **共线性诊断** | | | | | | | | | | | | | | | | | |
| **模型模** | **维度维** | **特特征值** | **条条件指标** | **方差比例** | | | | | | | | | | | | | |
| **(常量** | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **X8** | **X9** | **X10** | **X11** | **X12** | **X13** |
| **11** | **11** | **10.093** | **1.000** | **..00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** |
| **22** | **1.594** | **2.516** | **..00** | **..07** | **..08** | **.00** | **.00** | **.00** | **.00** | **.00** | **..01** | **.00** | **.00** | **.00** | **.00** | **.00** |
| **33** | **.960** | **3.243** | **..00** | **..02** | **..02** | **.00** | **..80** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** |
| **44** | **.662** | **3.905** | **..00** | **..31** | **..17** | **.00** | **..12** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** | **.00** |
| **55** | **.241** | **6.470** | **..00** | **..51** | **..18** | **..02** | **.00** | **.00** | **.00** | **.00** | **..02** | **..08** | **..01** | **.00** | **..01** | **.00** |
| **66** | **.167** | **7.786** | **..00** | **.00** | **..25** | **..03** | **.00** | **.00** | **.00** | **..02** | **..03** | **..09** | **.00** | **.00** | **.00** | **..16** |
| **77** | **.108** | **9.651** | **.00** | **..05** | **..07** | **..06** | **..03** | **.00** | **.00** | **..02** | **..11** | **..02** | **.00** | **.00** | **..02** | **..36** |
| **88** | **.074** | **11.640** | **.00** | **..03** | **..02** | **..48** | **.00** | **.00** | **.00** | **..16** | **..09** | **..03** | **.00** | **.00** | **.00** | **..01** |
| **99** | **.042** | **15.581** | **.00** | **.00** | **..01** | **.00** | **.00** | **.00** | **.00** | **..19** | **..16** | **..01** | **.00** | **.00** | **..69** | **..08** |
| **110** | **.026** | **19.840** | **..01** | **.00** | **.00** | **..07** | **.00** | **..04** | **..04** | **..56** | **..31** | **..03** | **.00** | **..01** | **..18** | **..06** |
| **111** | **.013** | **27.663** | **.00** | **.00** | **..01** | **..23** | **..02** | **..02** | **..02** | **..00** | **..01** | **..64** | **..90** | **..01** | **.00** | **..01** |
| **112** | **.012** | **28.912** | **.00** | **.00** | **..08** | **..04** | **..01** | **..37** | **..01** | **..02** | **..16** | **..01** | **..03** | **..27** | **..03** | **.00** |
| **113** | **.007** | **37.418** | **.00** | **.00** | **..11** | **..05** | **.00** | **..26** | **..39** | **..00** | **..00** | **.00** | **..05** | **..32** | **..01** | **..23** |
| **114** | **.001** | **87.318** | **..99** | **.00** | **.00** | **.00** | **.00** | **..31** | **..53** | **..02** | **..10** | **..08** | **..01** | **..39** | **..06** | **..10** |
| **a 因变量：X14** | | | | | | | | | | | | | | | | | |

5.1.2结论

通过以上的相关系数分析、拟合度分析以及多重共线性分析可得出，无论是使用六个相关较强变量还是八个相关较强变量抑或是十三个变量来进行预测，我们可以发现，利用十三个变量拥有最好的效果。

5.2 对问题二的分析

5.2.1 模型建立与求解部分

根据问题一的结论，将13个解释变量均用于BP神经网络[2]的模型预测上。

首先，作数据处理。将原始的506行数据拆分成训练样本（前496行），预测样本（后10行）两部分；使用MATLAB内置的premnmx函数将两组数据都做相同的归一化处理，使其取值范围在 [-1, 1] 之间。

其次，建立网络。使用newff函数建立一个三层神经网络（其中第一个隐藏层包括5个维度的神经元，第二个隐藏层包括10个维度的神经元，第三层即输出层包括1个维度的神经元）；合理设置网络参数（如将训练步数设置为50000等，具体参数见附录代码）。

再次，训练网络。将训练样本交由train函数，得到训练后的网络。同时，得到以下图表。其中，图5-2中包含了神经网络图的拓扑结构，使用的算法，训练过程等。

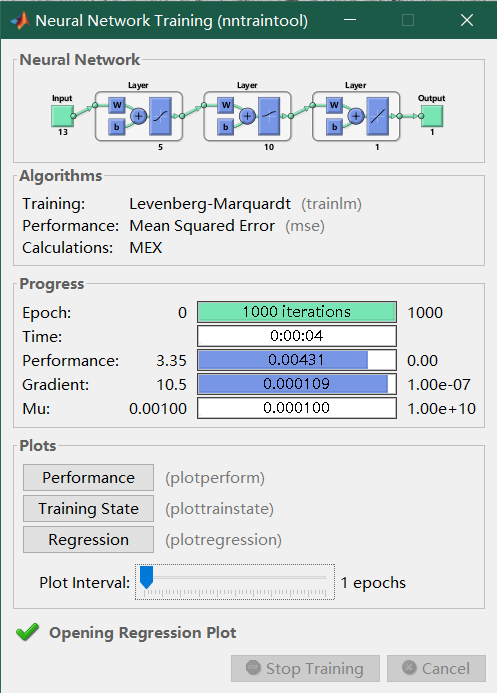


图5-2 训练终止截图

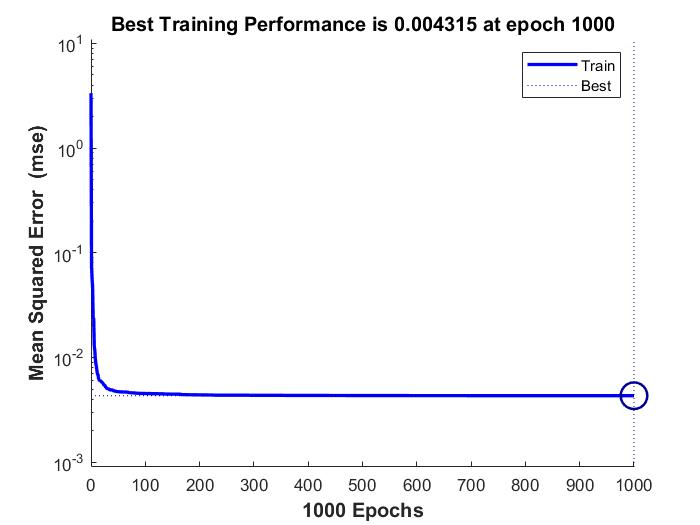


图5-3 训练表现

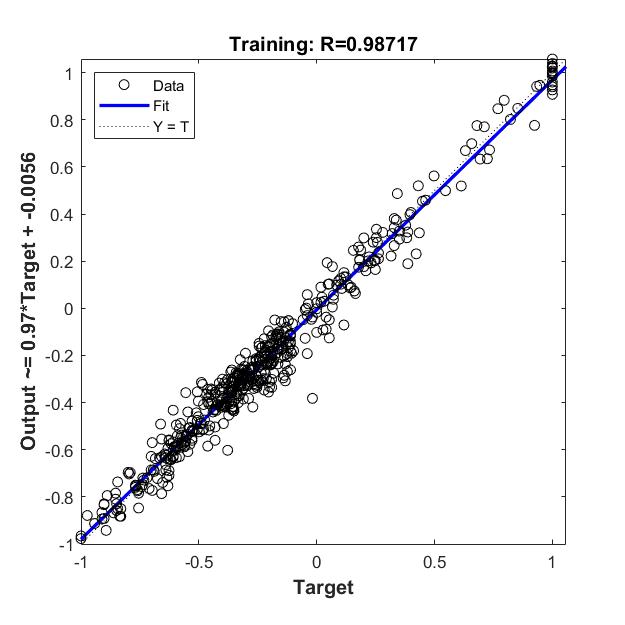
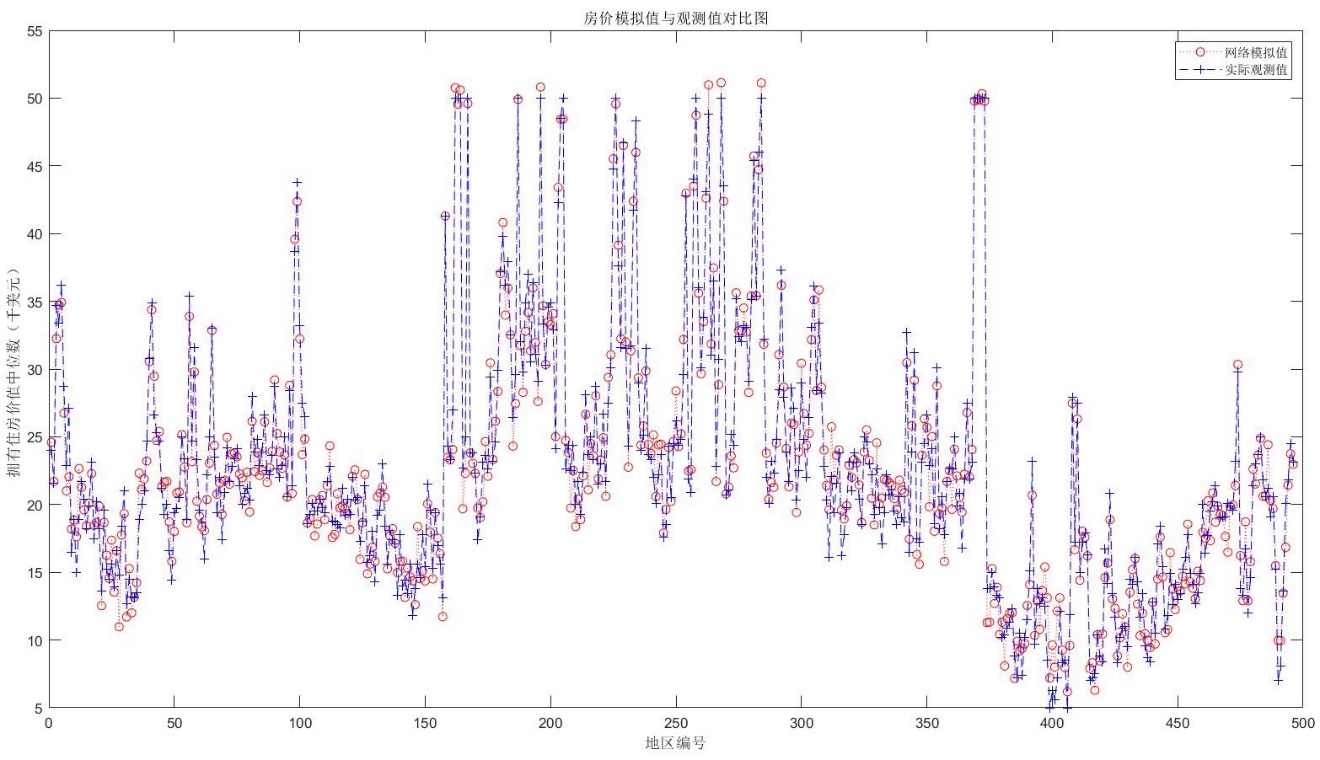


图5-4 训练结果与实际值之间的相关性

图5-5 训练样本的房价模拟值与观测值对比图

最后，进行预测。将已经归一化的用于预测的输入样本交由网络来进行仿真，获得以下结果。其中编号为497到506的房价预测值和真实值之间的对比如下：

表5-5 房价预测值和真实值的比较

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **497号** | **498号** | **499号** | **500号** | **501号** | **502号** | **503号** | **504号** | **505号** | **506号** |
| **预测值** | 18.4438 | 19.4087 | 20.4507 | 19.1264 | 18.7734 | 22.808 | 22.5962 | 21.8984 | 23.2051 | 21.1576 |
| **实际值** | 19.70 | 18.30 | 21.20 | 17.50 | 16.80 | 22.40 | 20.60 | 23.90 | 22.00 | 11.90 |

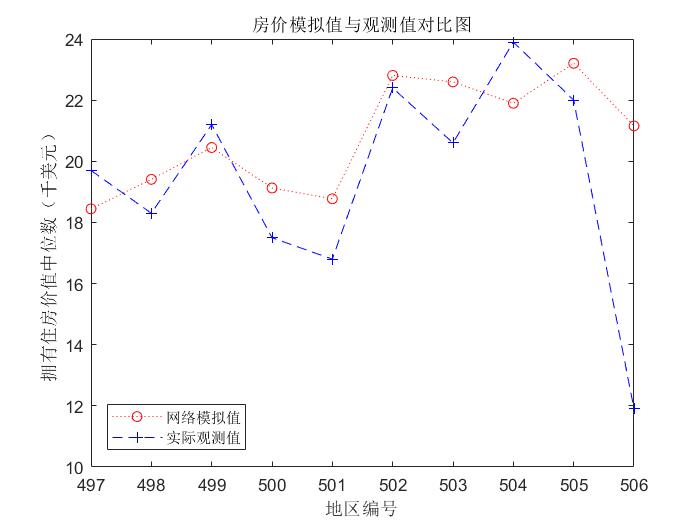


图5-6 房价模拟值与观测值对比图

5.2.2 模型评价部分

模型评价包括对训练数据的拟合效果的评价和对预测结果的评价。

首先，对训练数据的拟合效果进行评价。从图5-3可以看出在迭代到第1000次时的表现，MSE = 0.004315（公式5.3），均方误差很小；从图5-4可以看出训练结果与实际值之间的相关性 R = 0.98717，结果的相关性很大；从图5-5可以看出训练样本的房价模拟值与观测值拟合情况也是十分优秀的。

(公式5.3)

其次，对预测结果进行评价。从表5-5和图5-6可以直观地看出，相差不大。计算其均方误差得到 MSE = 18.4096。与训练数据相比，这个值较大。但在诸多次调参后，18.4096是相对优秀的状态。况且，训练数据由于神经元数量等原因，会让其表现产生过度优秀的状况。所以综上，预测结果是可以接受的。

5.3对问题三的建模与求解

5.3.1数据整理

文章从房屋内部环境和外部环境出发考虑宜居性，选取人均犯罪率、大块占地住宅区比例、非零售商业占地比例、查尔斯河虚拟变量、氮氧化物浓度、每户平均房间数、与五个波士顿劳动力聚集区的加权距离、与辐射式公路接近指数七个指标，并将指标进行整理汇总，如表5-6（部分）所示。

表 5-6

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **房屋编码** | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X8** | **X9** |
| **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX/10** | **RM** | **DIS** | **RAD** |
| 1 | 0.00632 | 18.00 | 2.31 | 0.00 | 0.538 | 6.575 | 4.090 | 1.00 |
| 2 | 0.02731 | 0.00 | 7.07 | 0.00 | 0.469 | 6.421 | 4.967 | 2.00 |
| 3 | 0.02729 | 0.00 | 7.07 | 0.00 | 0.469 | 7.185 | 4.967 | 2.00 |
| 4 | 0.03237 | 0.00 | 2.18 | 0.00 | 0.458 | 6.998 | 6.062 | 3.00 |
| 5 | 0.06905 | 0.00 | 2.18 | 0.00 | 0.458 | 7.147 | 6.062 | 3.00 |
| 6 | 0.02985 | 0.00 | 2.18 | 0.00 | 0.458 | 6.430 | 6.062 | 3.00 |
| 7 | 0.08829 | 12.50 | 7.87 | 0.00 | 0.524 | 6.012 | 5.561 | 5.00 |
| 8 | 0.14455 | 12.50 | 7.87 | 0.00 | 0.524 | 6.172 | 5.951 | 5.00 |
| 9 | 0.21124 | 12.50 | 7.87 | 0.00 | 0.524 | 5.631 | 6.082 | 5.00 |
| 10 | 0.17004 | 12.50 | 7.87 | 0.00 | 0.524 | 6.004 | 6.592 | 5.00 |
| 11 | 0.22489 | 12.50 | 7.87 | 0.00 | 0.524 | 6.377 | 6.347 | 5.00 |
| 12 | 0.11747 | 12.50 | 7.87 | 0.00 | 0.524 | 6.009 | 6.227 | 5.00 |
| 13 | 0.09378 | 12.50 | 7.87 | 0.00 | 0.524 | 5.889 | 5.451 | 5.00 |
| 14 | 0.62976 | 0.00 | 8.14 | 0.00 | 0.538 | 5.949 | 4.708 | 4.00 |
| ... |  |  |  |  |  |  |  |  |

5.3.2计算秩和比

文章选取数据中人均犯罪率、非零售商业占地比例、氮氧化物浓度、与五个波士顿劳动力聚集区的甲醛距离、与辐射式公路接近指数七个指标为低优指标，大块占地住宅区比例、查尔斯和虚拟变量、每户平均房间数为高优指标，将数据同一化处理。

根据公式：

(公式5.4)

得出RSR的值，并绘制RSR排序表，如表5-7（部分）所示。

表 5-7

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ZN/X2** | | **DIS/X8** | | **RAD/X9** | | **RSR** | **排序** |
|  |
| **A2** | **R2** | **A8** | **R8** | **A9** | **R9** |  |
| 18.00 | 384 | 4.090 | 184 | 1.00 | 496.5 | 0.7148 | 36 |  |
| 0.00 | 186.5 | 4.967 | 136.5 | 2.00 | 474.5 | 0.6187 | 112.5 |  |
| 0.00 | 186.5 | 4.967 | 136.5 | 2.00 | 474.5 | 0.6517 | 85 |  |
| 0.00 | 186.5 | 6.062 | 86 | 3.00 | 443.5 | 0.6709 | 66 |  |
| 0.00 | 186.5 | 6.062 | 86 | 3.00 | 443.5 | 0.6546 | 83 |  |
| 0.00 | 186.5 | 6.062 | 86 | 3.00 | 443.5 | 0.6445 | 89 |  |
| 12.50 | 377.5 | 5.561 | 103 | 5.00 | 257 | 0.5142 | 217 |  |
| 12.50 | 377.5 | 5.951 | 91 | 5.00 | 257 | 0.5117 | 226 |  |
| 12.50 | 377.5 | 6.082 | 84 | 5.00 | 257 | 0.4566 | 302 |  |
| 12.50 | 377.5 | 6.592 | 60 | 5.00 | 257 | 0.4839 | 274 |  |
| 12.50 | 377.5 | 6.347 | 70 | 5.00 | 257 | 0.5112 | 228 |  |

5.3.3确定RSR的分布

通过已给数据，根据excel功能计算出频数、平均秩次、向下累计频率、概率单位。绘制表格如表5-8（部分）所示：

表 5-8

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RSR** | **f** | **R** | **`R** | **(`R/N)\*100%** | **Probit** |
|  |
|  |
| 0.2853 | 1 | 1 | 1 | 0.1976% | 2.12 |  |
| 0.2883 | 1 | 2 | 2 | 0.3953% | 2.34 |  |
| 0.2904 | 1 | 3 | 3 | 0.5929% | 2.48 |  |
| 0.2906 | 1 | 4 | 4 | 0.7905% | 2.59 |  |
| 0.2914 | 1 | 5 | 5 | 0.9881% | 2.67 |  |
| 0.2915 | 1 | 6 | 6 | 1.1858% | 2.74 |  |
| 0.2921 | 1 | 7 | 7 | 1.3834% | 2.80 |  |
| 0.2926 | 1 | 8 | 8 | 1.5810% | 2.85 |  |
| 0.2927 | 1 | 9 | 9 | 1.7787% | 2.90 |  |
| 0.2929 | 1 | 10 | 10 | 1.9763% | 2.94 |  |
| 0.2940 | 1 | 11 | 11 | 2.1739% | 2.98 |  |

5.3.4计算回归方程

以累计频率所对应的概率单位值Probit为自变量，以RSR值为应变量，计算回归方程为：

y = 0.1315x - 0.1573 (公式5.5)

绘制回归图形如图所示：

图 5-6

根据回归方程求得RSR估计值，绘制表格如表（部分）所示：

表 5-9

|  |  |  |
| --- | --- | --- |
| **RSR** | **Probit** | ˆRSR |
|  |
| 0.2853 | 2.12 | 0.12 |  |
| 0.2883 | 2.34 | 0.15 |  |
| 0.2904 | 2.48 | 0.17 |  |
| 0.2906 | 2.59 | 0.18 |  |
| 0.2914 | 2.67 | 0.19 |  |
| 0.2915 | 2.74 | 0.20 |  |
| 0.2921 | 2.80 | 0.21 |  |
| 0.2926 | 2.85 | 0.22 |  |
| 0.2927 | 2.90 | 0.22 |  |
| 0.2929 | 2.94 | 0.23 |  |
| 0.2940 | 2.98 | 0.23 |  |
| 0.2943 | 3.02 | 0.24 |  |

5.3.5分档排序

根据秩和比分档排序七挡表得出排序结果，绘制成结果表格如表所示：

表 5-10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **类型** | **P** | **Probit** | **ˆRSR** | **分档排序结果** |
| 非常宜居 | <P1.618 | <2.86 | <0.22 | 415,363… |
| 宜居 | P1.618~ | 2.86~ | 0.22~ | 405,466… |
| 较宜居 | P10.027~ | 3.72~ | 0.33~ | 362,144… |
| 一般 | P33.36~ | 4.57~ | 0.45~ | 33,493… |
| 较不宜居 | P67.003~ | 5.44~ | 0.56~ | 332,2… |
| 不宜居 | P89.973~ | 6.28~ | 0.67~ | 187,81… |
| 非常不宜居 | P98.382~ | 7.14~ | 0.78~ | 103,284… |

5.4对问题四的建模与求解

5.4.1对问题四的分析

问题四是要针对不同人群给出买房建议，我们从多个角度出发对人群进行分类，对于人群的分类有很多种分类标准，我们在分析中不一一列出，综合分析人群对于房子的建议，对于购房者而言，对所购买的房子的宜居性要求就是购房者对房子的需求。根据我们之前的数据分析，购房者倾向于选择人均犯罪率低、不靠近河流、非商业占地比例低、空气中氮氧化物浓度低等环境条件较好的房子。周围环境是否安全影和空气等客观环境因素都会影响购房者对于房子的选择。

5.4.2对问题四的解决

其实问题四不难解决，我们可以从每个变量入手，对其进行详细的刨析。

城镇的人均犯罪率与房价的相关系数为-0.39，为负值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，人均犯罪率高的地区，房价相对低，并且相关程度中等。实际上，人均犯罪率高的地区治安差，人群受教育程度低，黑势力较为集中，房价就较低。

大块占地住宅区比例与房价的相关系数为0.36，为正值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，大块占地住宅区比例高的地区，房价相对高，并且相关程度中等。实际上LRA，反映的是与市中心的接近程度，显然越接近市中心的地区LRA越高，房价也越高，呈正相关，符合实际情况。

非零售商业占地比例与房价的相关系数为-0.48，为负值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，非零售商业占地比例高的地区，房价相对低，相关程度中等。根据资料显示，波士顿的批发厂主要分布在离市中心较远的地区，所以该变量同样能反映距离市中心的距离，比例越大，离市中心越远，房价越低。

查尔斯河虚拟变量与房价的相关系数为0.18，为正值且绝对值在0.3以下，说明在其它因素相同的情况下，靠近河流的地区，其房价相对高一点，相关程度较弱。实际上，靠近河流的地区，其自然环境相对较好，交通运输相对便利，但可能距离市中心较远，故综合起来看，房价受它的影响不大。更为关键的因素是，CR的取值只有0，1两种，故与其它变量的相关性不能明确的表现。

氮氧化物浓度与房价的相关系数为-0.43，为负值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，氮氧化物浓度高的地区，房价相对低，相关程度中等。实际上，氮氧化物浓度高的地区主要在大型工厂附近，自然环境相对差，且距离市中心较远，故房价相对低。

每户平均房间数与房价的相关系数为0.7，为正值且绝对值大于0.5，说明在其它因素相同的情况下，每户平均房间数高的地区，房价相对高，相关程度较强。实际上，房价最直接的影响因素就是房屋面积，每户平均房间数高的地区，其房屋面积大，甚至可能为别墅区，房价固然高并且影响程度较大。

1940年前建造的户主所有房比例与房价的相关系数为-0.38，为负值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，1940年前建造的户主所有房比例大的地区，房价相对低，相关程度中等。实际上，HOB能反映该地区城市化的程度，HOB越大，说明该地区老旧房屋占比大，城市化程度低，房价自然相对较低。

与五个波士顿劳动力聚集区的加权距离与房价的相关系数为0.25，为正值且绝对值在0.3以下，说明在其它因素相同的情况下，与五个波士顿劳动力聚集区的加权距离大的地区，房价相对高，相关程度较弱。理论上，WD5越大，说明该地区距离中心区相对远，房价应该相对低，但实际上，WD5不能准确的说明该地区是否在市区或者郊区，并且可能考虑到郊区别墅高房价、环境优的影响，使得WD5与房价表现了较弱的正相关性，与理论分析略有不同。

与辐射式公路接近指数与房价的相关系数为-0.38，为负值并且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，与辐射式公路的接近指数大的地区，房价相对低。相关程度中等。实际上，RH反映了该地区的交通情况，RH越小，说明附近交通便利，房价相对高。

每1万美元的全值财产税与房价的相关系数为-0.47，为负值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，每一万美元的全值财产税大的地区，房价相对较低。相关程度中等。实际上，TAX越大，说明财产税越大，更少人愿意去买，故而房价低。

学生/教师比例与房价的相关系数为-0.51，为负值且绝对值在0.5以上，说明在其它因素相同的情况下，学生与教师的比例大的区，房价相对低，相关程度较强。实际上，SFR反映了该地区的教育程度和学校的密集程度，SFR越小，说明教师相对学生多，教育程度高，并且附近学校较多，也可能为学区房，故房价高，呈现负相关。

1000（B-0.63）（B-0.63）I，B为非洲裔美国人比例，其与房价的相关系数为0.33，为正值且绝对值在0.3~0.5之间，说明在其它因素相同的情况下，POAA大的地区，房价相对较高，相关程度中等。实际上，POAA反映了非洲裔美国人的比例，POAA高的地区，非洲裔美国人的比例小，由于美国种族因素的影响，黑人比例大的地区，往往人均社会地位相对较低，房价也低，故POAA高的地区黑人比例小，房价相对高。

低社会地位人口的比例与房价的相关系数为-0.74，为负值且绝对值在0.5以上，说明在其它因素相同的情况下，低社会地位人口的比例高的地区，房价相对低，相关程度较强。实际上LSSP越高，说明该地区处于较低层次的社会，生活质量相对低，房价低。

总的来说，不同的地区房价不同，不同的阶层接触到的房价不同，不同的环境房价也不同，不同的性质房价房价更不同……根据以上的分析，居民可以进行更加细致全面的购房选择，这里也不再一一赘述，希望以上分析对与居民们有所作用。

5.4.3绘图与解析

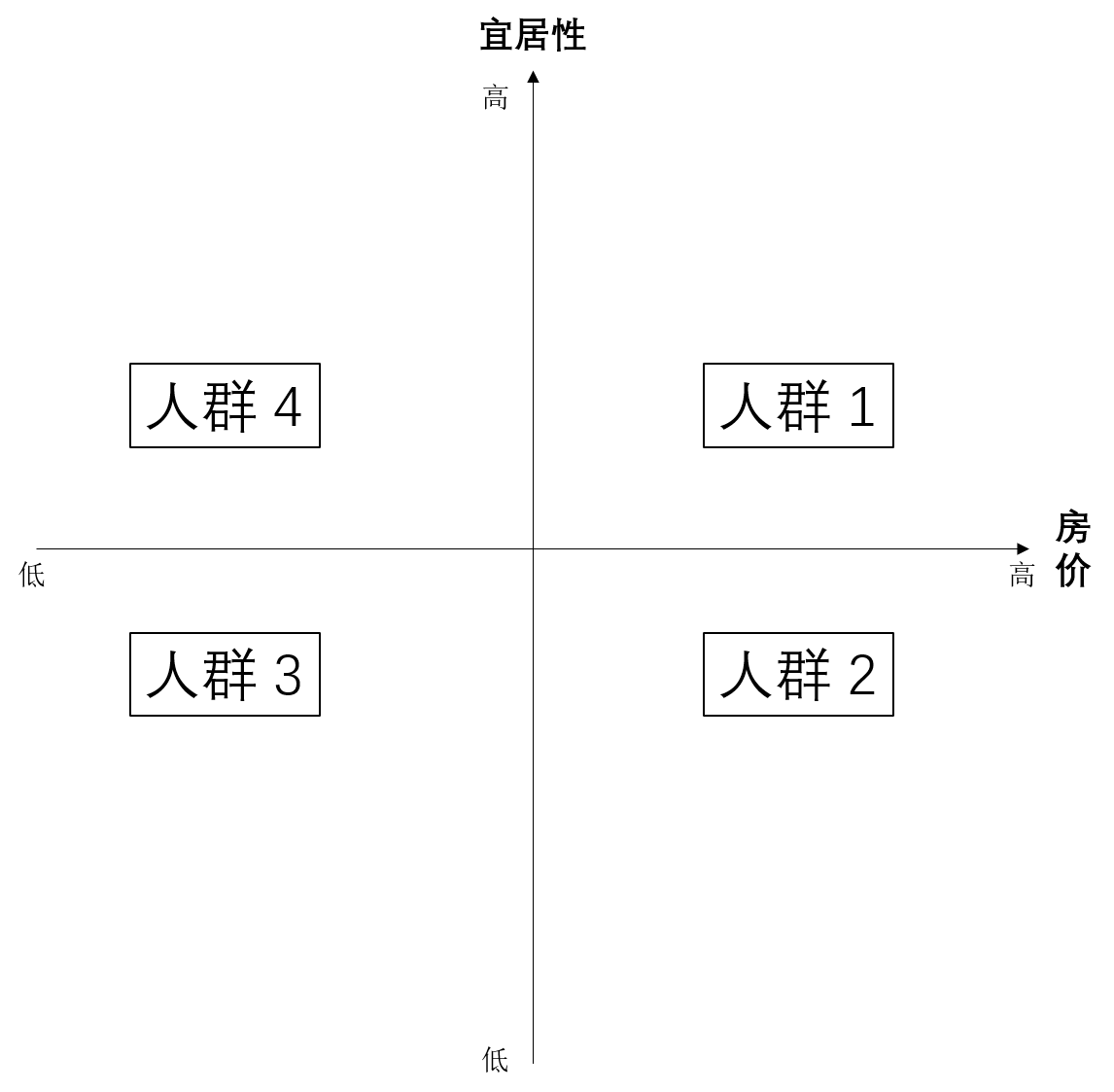


图5-7 划分图

文章将人群按照收入和生活态度划分为四种类别。人群1为富豪，人群2为中产阶级。人群3为无产阶级，人群4为对对自然生活有较强向往者等，如图5-7所示。

文章按照宜居性与房价作为划分依据绘制四象限表，假设房价和宜居性为人们考虑买房的决定性因素，划分情况如下：

高房价，高宜居性：如波士顿CBD的别墅  
高房价，低宜居性：如波士顿CBD附近的公寓  
低房价，低宜居性：如波士顿的部分的平民社区  
低房价，高宜居性：如波士顿靠近郊区的庭院式住宅区。

人群1为富豪阶层，该人群特点为收入高，房价影响买房权重小，该人群买房几乎无关价格，以宜居性为主要考虑指标，在不考虑房价的情况下，选择最宜居的房屋。该人群位于第一象限，高房价、高宜居性，一般为波士顿的别墅群。

人群2为中产阶级，该人群特点为收入水平较高，位于无产阶级和中产阶级之间，但是在发达城市生活，收入远不如房价，这类人群选择房屋主要考虑到房价和上班便捷程度，对宜居性要求不高，但往往工作地点附近房价较高，所以该人群位于第四想想，高房价、低宜居性，一般为波士顿CBD附近的公寓。

人群3为无产阶级，该人群特点为收入低，房价影响权重大，宜居性影响权重适中，主要考虑在经济承受范围内最适合的房子。此类人群以生活为主，考虑到宜居性但是又往往生活在较为便利的社区，由于收入水平低，经济能力有限，只能选择低房价、低宜居性的房屋，位于第三象限，多为波士顿的部分的平民社区。

人群4为对自然生活有较强向往者，此类人群经济水平较低，但是向往高质量生活，追求高宜居性的房屋，故需要牺牲市区的条件，选择位置偏远地区的房屋，此类房屋自然环境优美，宜居性较高，但是生活基础设施缺乏，故价格低。多为波士顿附近郊区的庭院式住宅区。

六、模型的优缺点及改进

6.1模型的优点：

1. 模型采用了神经网络、相关系数、多重共线性、回归分析、秩和比综合评价法等模型，建立求解，有充分的理论支撑，可信度高，实用性较强。
2. 用matlab、excel、spss等工具进行数据处理与模型建立，所得结果精确，可信度高。
3. 模型建立来源于实践，从生活角度出发，用贴近生活的思维完善模型，模型构建体现生活特色，具有较强的实用性与现实性。
4. 文章前后逻辑贯通，模型建立前后有思维的迭代，解释说明较为完善，对于读者来说通俗易懂，文章可读性高。

6.2模型的缺点：

1. 模型第三问数据选择凭主观臆断，并未借助任何公式模型，结果因人而异，会有误差。
2. 针对问题一仅仅分析了大概，并未做误差分析，可能会有较大的误差存在。
3. 模型三并未运用语言进行解决，而是运用excel较为简单。

参考文献

[1] D. Harrison, D.L. Rubinfeld, Hedonic prices and the demand for clean air. J. Environ. Econ. Manag. 5, 81–102 (1978)

[2]丁飞,江铭炎.基于改进狮群算法和BP神经网络模型的房价预测[J/OL].山东大学学报(工学版):1-9[2021-07-26].http://kns.cnki.net/kcms/detail/37.1391.T.20210513.1403.004.html.

附录

附录一：求解问题二用到的代码

MATLAB 实时脚本文件，problem2.mlx

clear;

clc;

close all;

middle\_point = 496;

districts = 1 : middle\_point; % 用于训练的地区

test\_districts = middle\_point + 1 : 506; % 用于测试的地区

load data.mat; % 加载数据

%load sorted\_data.mat

%data = sorted\_data;

data = data'; % 数据规整化

% 将前 districts 个数据作为训练数据

input = data(1:13, districts);

output = data(14:14, districts);

% 将后 test\_districts 个数据作为测试数据

test\_input = data(1:13, test\_districts);

test\_output = data(14:14, test\_districts);

[pn,minp,maxp,tn,mint,maxt] = premnmx(input, output); % 归一化

dx = [-1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1; -1 1]; % 取值范围

TF1='tansig';TF2='logsig';TF3='purelin'; % 设置传输函数

net = newff(dx, [5, 10, 1], {TF1, TF2, TF3}, 'trainlm'); % 创建网络

% 设置网络参数

net.trainParam.epochs = 50000;

net.trainParam.show = 1000;

net.trainParam.lr = 0.05;

net.trainParam.goal = 1e-4;

net.trainParam.mc = 0.9;

net.trainFcn = 'trainlm';

net = train(net, pn, tn); % 利用归一化后的数据pn,tn训练网络

a = sim(net,pn); % 仿真

answer = postmnmx(a,mint,maxt); % 反归一化，即还原

% 进行训练数据比较，绘制图表

figure(1);

plot(districts, answer, 'r:o', districts, output, 'b--+');

legend('网络模拟值','实际观测值');

xlabel('地区编号');ylabel('拥有住房价值中位数（千美元）');

title('房价模拟值与观测值对比图');

% 预测后 test\_districts 个地区的房价

test\_pn = tramnmx(test\_input, minp, maxp); % 归一化

tmp = sim(net, test\_pn); % 仿真

test\_answer = postmnmx(tmp, mint, maxt) % 反归一化

% 进行测试数据比较，绘制图表

figure(2);

plot(test\_districts, test\_answer, 'r:o', test\_districts, test\_output, 'b--+');

legend('网络模拟值','实际观测值');

xlabel('地区编号');ylabel('拥有住房价值中位数（千美元）');

title('房价模拟值与观测值对比图');

% 计算预测出来的结果的均方误差

ssr = sum(test\_output - test\_answer).^2; % 残差平方和

mse = ssr / (506 - middle\_point) % 均方误差