

Without PCA the premium MSe comes with the layer: 27:28

After running 4 neural network methods for 1600 times, the number of nodes in each hidden layer has been conducted.

mistakeda =

1.2594e+08

mistakedm =

1.2719e+08

mistakegd =

1.4335e+08

mistakerp =

1.2545e+08

gdaI =

36

gdaJ =

16

gdI =

2

gdJ =

10

gdmI =

9

gdmJ =

1

rpI =

1

rpJ =

35

Missing Data:

For data missing in question q101a2 to q101a6, which requests information about whether immigrates have certain family members travelling with them, a method simply refill the blank with 0 has been used in this project.

For records under question q203, which asking about the area that immigrates working in, a number 15 denoting that the immigrate worked in another unmentioned area.

For records under question q204, which asking about the kind of job that immigrates working in, a record 80 denoting that the immigrate worked in another unmentioned area is used in filling the blanks.

For records under question q205, which asking about the kind of cooperation that immigrates working in, a record 12 denoting that the immigrate did not work in any kind of cooperation as mentioned under the question is used in filling the blanks.

For records under question q206, which asking about the role inside the cooperation that immigrates working in, a record 12 denoting that the immigrate did not work as any kind of role as mentioned under the question is used in filling the blanks.

For records under question q207, which asking about the income last month for the immigrate, a record 0 denoting that the immigrate did not have any income is used in filling the blanks.

For records under question q208, which asking about the average working days for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q209, which asking about the average working hours per day for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q301, which asking about the time for the wedding of the immigrate, a record 201300 denoting that the immigrate has not married until 2012 is used in filling the blanks.

For records under question q304, which asking about the whether the family obtained the prof for one child family, a record 0 denoting that the immigrate has no child until 2012 is used in filling the blanks.

For records under question q306, which asking about the whether the the immigrate use any contraceptive method, a record 0 denoting that the immigrate has no sexual life during the last month for taking the questionnaire is used in filling the blanks.

Q308 q309

Q310: 9

Q311: 0

Q312-316:0

With the data after refilling, a better result can be conducted from examples with toCode 11:

>> mistake

mistake =

9.0610e+07

>> bI

bI =

15

>> bJ

bJ =

11

After PCA:

10^8

1.2700 1.2578 1.3106 1.2471

IList =

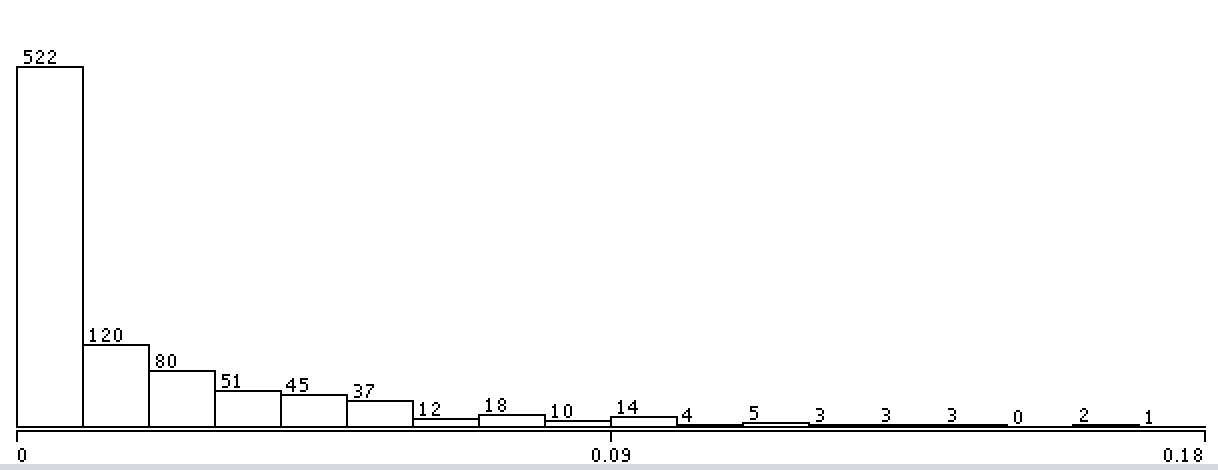
3 2 10 39

JList=

2 39 1 16

Interesting point: The increase rate of flat price and its relationship with immigration rate is not a positive relation as most academies expected. The MIC shows there is a certain relation between them with a correlation score as 0.63626. However, if the data is analyzed with linear model the score is -0.3043905, which denotes that, generally, there is a certain negative relation between.

Bagging Prediction:



1 Using Bagging Aggregating for predicting immigration rates

Bagging Aggregation is a method to select few subsets randomly from a whole data set for training the decision trees. With separate sets of data, the stability of prediction will be improved and the over-fitting problem can be generally avoided. Typically, in this analysis process, an iteration number of 10 has been chosen while in each iteration, of the whole data has been picked into the training set.

## 1.1 Immigration Data

All immigration Data sets are official data from the National Bureau of Statistics of China. There are mainly three parts of them used for training the model: Immigration investigation data, Geographic Data and The economic and environmental data.

With the recordings from thousands of immigration families taken from 2009 to 2014, the Immigration Investigation data sets contain ….. //数据集包括很多调查得到的个体信息。In the processing of building this model, only the rate of immigration between provinces and cities are used from calculating the numbers of immigration families from and to each places.

## 1.2 Selecting features

For predicting the future immigration rate, features must be selected to build the model. During the process of evaluating the relevance between feature and target variable, MIC scores are used to select the valuables that contribute more influence on immigration rate than others.

From Geographic Data set and The economic and environmental data set, the following feathers are conducted for analysis:

|  |  |
| --- | --- |
| dis | Distances for immigration |
| orgGDP | The GDP per capita of the place immigrates from |
| Population | The population of the destiny |
| Birth\_Rate | Birth Rate of the destiny |
| Avg\_Life | Life expectancy |
| GDP\_PC | GDP per capita of the destiny |
| Unemploy\_Rate |  |
| Flat\_Rates |  |
| Price\_Rates |  |
| City\_Income |  |
| Country\_income |  |
| Green\_Rate |  |
| Disasters |  |
| Stu\_Rate\_H |  |
| Low\_Income |  |
| Low\_Income\_Insurance |  |
| Flat\_Rates\_Change |  |

The features above obtains high MIC score among all features. The MIC scores of them are listed as following:

|  |  |  |
| --- | --- | --- |
| Y var | MIC (strength) | Linear regression (p) |
| toCode | 0.67098 | -0.23456322 |
| Population | 0.67098 | 0.5396318 |
| Birth\_Rate | 0.67098 | -0.29075608 |
| Avg\_Life | 0.67098 | 0.4073089 |
| Flat\_Rates | 0.67098 | 0.53999525 |
| City\_Income | 0.67098 | 0.5662188 |
| Country\_income | 0.67098 | 0.5470502 |
| Green\_Rate | 0.67071 | 0.30561876 |
| GDP\_PC | 0.66989 | 0.48178324 |
| Disasters | 0.66989 | -0.075478196 |
| Stu\_Rate\_H | 0.66989 | 0.3403138 |
| Flat\_Rates\_Change | 0.63526 | -0.3043905 |
| Low\_Income\_Insurance | 0.62499 | 0.45970377 |
| Low\_Income | 0.61332 | -0.226193 |
| Price\_Rates | 0.37312 | -0.12579323 |
| Unemploy\_Rate | 0.21025 | -0.2366798 |
| orgGDP | 0.14091 | -0.014871608 |
| fromCode | 0.13771 | 0.060742937 |

The increase rate of flat price and its relationship with immigration rate is not a positive relation as most academies expected. The MIC shows there is a certain relation between them with a correlation score as 0.63626. However, if the data is analyzed with linear model the score is -0.3043905, which denotes that, generally, there is a certain negative relation between. This might due to the influence from politics, since the government take actions for reducing the high flat rate in big cities for improving the living environment for locals. For this reason, the flat rates change is excluded from this model.

## 1.3 Building the model:

Using RMS(Root Mean Square) to comparing between different classifiers for predicting Immigration Rate:

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Bagging | 0.0537 | 0.0196 |
| MultilayerPerceptron | 0.0638 | 0.0345 |
| AdditiveRegression | 0.0587 | 0.0241 |
| REPTree | 0.0615 | 0.0223 |
| DecisionStump | 0.0656 | 0.0283 |
| Linear Regression | 0.0603 | 0.0273 |

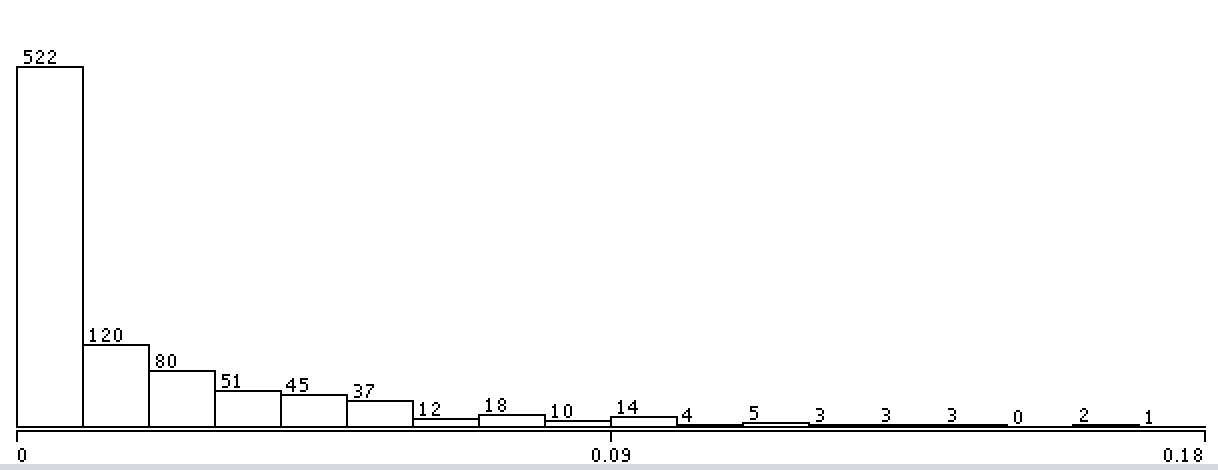
As shown in the table, the Bagging owns the best accuracy with RMS around 5.4%.

## 1.4 Predictions:

Predictions from the model according to values from features can be calculated to be as following format:

|  |  |  |
| --- | --- | --- |
| Actual | Prediction | Error |
| 0 | 0.001 | 0.001 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0.015 | 0.076 | 0.062 |
| 0.006 | 0.007 | 0.001 |
| 0.001 | 0.016 | 0.015 |
| 0 | 0 | 0 |
| 0.049 | 0.097 | 0.048 |
| 0.032 | 0.02 | -0.011 |
| 0.001 | 0.01 | 0.009 |
| 0.167 | 0.166 | -0.001 |
| 0 | 0.03 | 0.03 |
| 0.004 | 0.009 | 0.005 |

Error Bar:



2. Neural Network for individual prediction:

## 2.1 Methods:

Provinces in the data set are denoted by different codes. Obviously, codes are not good for identifying provinces.

In this section, using codes to denote provinces leads to larger errors. Codes fall in carrying much internal information or integrity for provinces considering that few provinces with vary different economic environment and geographic localization may have very close codes.

For this reason, GDP Per Capita is used in presenting different provinces. In the process of prediction, the Neural Network will output predicted GDP Per Capita of the destination that one family likely intended to travel to. If further prediction requires to know exactly the place rather than the GDP Per Capita of the destination, a province that near the original province which obtains close GDP Per Capita of the prediction will be chosen to be the place one probably travels to.

In the process of training, to find the number of nodes in each layers for best result, a training process, which running neural network for 40\*40 times are used to iterate every possible combination of layers.

10-fold cross-validation are used in validating.

To insure the result are compared with same test set ( since Matlab will chose different data for testing if using 10-fold cross-validation by default), this paper use the whole data set for testing:

y = net(InputData);

p = perform(net,InputData,OutputData);

## 2.2 Data Preparation:

43 feathers are used to represent the data. Most of them are from the interviews taken by Chinese government. Unfortunately, among the 43 feathers, 28 of them contains Null value. Most of them are caused by the designing of the questionnaire, which requiring some of people only take parts of questions. The paper proposes various default values for feathers valued as Null according to the questions themselves, which means different questions will have different default values.

Methods for Missing Data imputation:

For data missing in question q101a2 to q101a6, which requests information about whether immigrates have certain family members travelling with them, a method simply refill the blank with 0 has been used in this project.

For records under question q203, which asking about the area that immigrates working in, a number 15 denoting that the immigrate worked in another unmentioned area.

For records under question q204, which asking about the kind of job that immigrates working in, a record 80 denoting that the immigrate worked in another unmentioned area is used in filling the blanks.

For records under question q205, which asking about the kind of cooperation that immigrates working in, a record 12 denoting that the immigrate did not work in any kind of cooperation as mentioned under the question is used in filling the blanks.

For records under question q206, which asking about the role inside the cooperation that immigrates working in, a record 12 denoting that the immigrate did not work as any kind of role as mentioned under the question is used in filling the blanks.

For records under question q207, which asking about the income last month for the immigrate, a record 0 denoting that the immigrate did not have any income is used in filling the blanks.

For records under question q208, which asking about the average working days for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q209, which asking about the average working hours per day for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q301, which asking about the time for the wedding of the immigrate, a record 201300 denoting that the immigrate has not married until 2012 is used in filling the blanks.

For records under question q304, which asking about the whether the family obtained the prof for one child family, a record 0 denoting that the immigrate has no child until 2012 is used in filling the blanks.

For records under question q306, which asking about the whether the the immigrate use any contraceptive method, a record 0 denoting that the immigrate has no sexual life during the last month for taking the questionnaire is used in filling the blanks.

Q308 q309: similar above

Q310: 9

Q311: 0

Q312-316:0

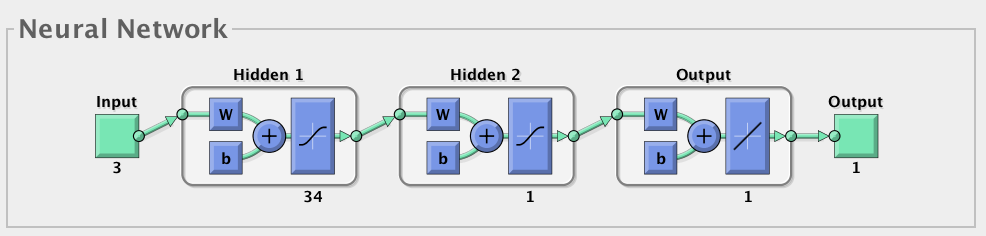
With the data after refilling, a better result can be conducted from examples with toCode 11:

Running Neural Network for hidden layers between [1:1] to [40:40] with the Data with adding 0 for all missing data:

MSe: 9.7 \* e ^ 7 found in layers: 27:28

After Missing Data Imputation as above:

MSe: 9.0610 \* e ^ 07



## 2.3 PCA

With 43 dimensions of data, running Neural Network through out all the data with dimensions from 1 to 40 for less than 3 layers and few algorithms would be time-expensive and make result over fitted. In this case, PCA would be useful for reducing dimensions.

After running PCA for the data after Missing Data Imputation, the top ten eigenvalues are:

|  |  |
| --- | --- |
| 1 | 378937318.737825 |
| 2 | 13077648.3401719 |
| 3 | 759810.916836644 |
| 4 | 274.561199927459 |
| 5 | 42.8332262317480 |
| 6 | 25.3276573980780 |
| 7 | 17.2908856398070 |
| 8 | 9.85532782440167 |
| 9 | 7.17570561084472 |
| 10 | 5.47016252574485 |

After Picking the top 3 eigenvalues, running the four Neural Network algorithms with the same part of data in section 2.2, the result is showing below:

MSe:

10^8

1.2700 1.2578 1.3106 1.2471

Hidden nodes in first layers=

3 2 10 39

Hidden nodes in second layers=

2 39 1 16

The four algorithms are:

1. Gradient descent backpropagation (traingd) – Parameter: learning rate (lr).
2. Gradient descent with adaptive learning rate backpropagation (traingda) – Parameters: learning rate (lr), ratio increase/decrease learning rate (lr\_inc, lr\_dec).
3. Gradient descent with momentum backpropagation (traingdm) – Parameters: learning rate (lr), momentum constant (mc).
4. Resilient backpropagation (trainrp) – Parameters: Increment/Decrement to weight change (delt\_inc/delt\_dec).

The result shows the square errors are around , which means real errors are around in Chinese Yuan, which can about to decide which province the immigrate about to travel to, since GDP Per Capita in provinces near to each other are mostly larger than .