MISCADA Core IIA (3) Classification Summative Coursework

Dataset

First of all, I select a practical dataset referring to https://github.com/ageron/handson-ml/blob/master/datasets/housing/housing.csv. I choose it because:

1. It can be accessed easily.
2. This dataset is a simple but practical collect about real estate, which is related to our real life.
3. Size of this dataset is medium, so I can use it to perform all techniques learned from class.

Executive summary

Housing price is an important factor affecting the quality of life, for young men and women. As we all know, housing price is a synthetic product under influence of many things, e.g., region, housing age, housing squares and so on. In this project, I'll research on the relationship between ocean proximity and other housing attributes. From dataset, there are 3 different positions represent space to the ocean. I need to model a mapping function that analyze how to mine the variables such as longitude, latitude, housing_median_age and total_rooms in the dataset. So it's a classic classification problem.

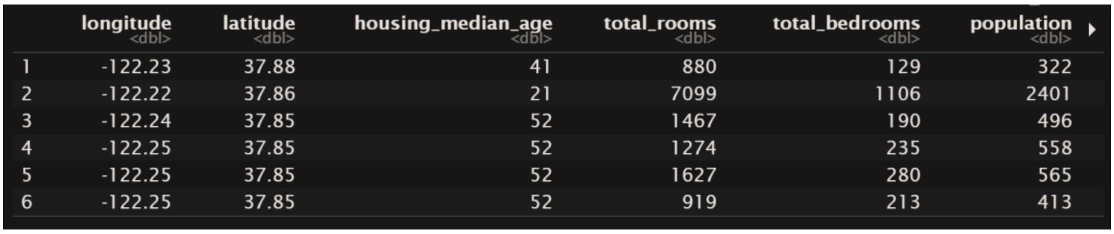
In order to classify each entry more correctly, I need perform many steps about cleaning data, feature engineering and model building. In terms of cleaning data, I'll visualize all data entries. Then fill all missing data. In terms of feature engineering, since all target labels are type of string, I'll transform them to numeric for the purpose of computing. Especially, I instruct two machine learning algorithms of decision tree and random forest. Decision tree is an extreme classic machine learning algorithm. It constructs a tree architecture to discover the classification rules implied in the data. How to construct a high-precision, small-scale decision tree is the core content of a decision tree algorithm. Decision tree construction can be performed in two steps. The first step is the generation of a decision tree: the process of generating a decision tree from a training sample set. In general, the training sample data set has a history and a certain degree of comprehensiveness according to actual needs and is used for data analysis and processing. The second step is the pruning of the decision tree. The pruning of the decision tree is the process of checking, correcting and repairing the decision tree generated in the previous stage. What's more, the implementation of decision tree in R is very effective to users

In machine learning, a random forest is a classifier containing multiple decision trees, and its output category is determined by the mode of the categories output by the individual trees. To evaluate these two models' performance, I test the accuracy on the testing dataset. Results demonstrate that decision tree can achieve a high accuracy to classify the ocean proximity

Technical Summary

First of all, I read the dataset *housing.csv* by read.csv, which is formatted as a dataframe. The first 6 columns are listed in Table 1.

Tab. 1 First 6 columns in housing dataset



	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	-122.23	37.88	41	880	129	322
2	-122.22	37.86	21	7099	1106	2401
3	-122.24	37.85	52	1467	190	496
4	-122.25	37.85	52	1274	235	558
5	-122.25	37.85	52	1627	280	565
6	-122.25	37.85	52	919	213	413

It is noted that 'ocean_proximity' is a type of string, so it can't be numerical calculus. It's a important point to process in the phase of feature engineering.

For the purpose of data exploration, I use the visualized function in ggplot2, a famous plotting library in R language, to show the distribution of each column in Fig. 1. Histogram is a brilliant manner to present the number of data fallen into a bin band

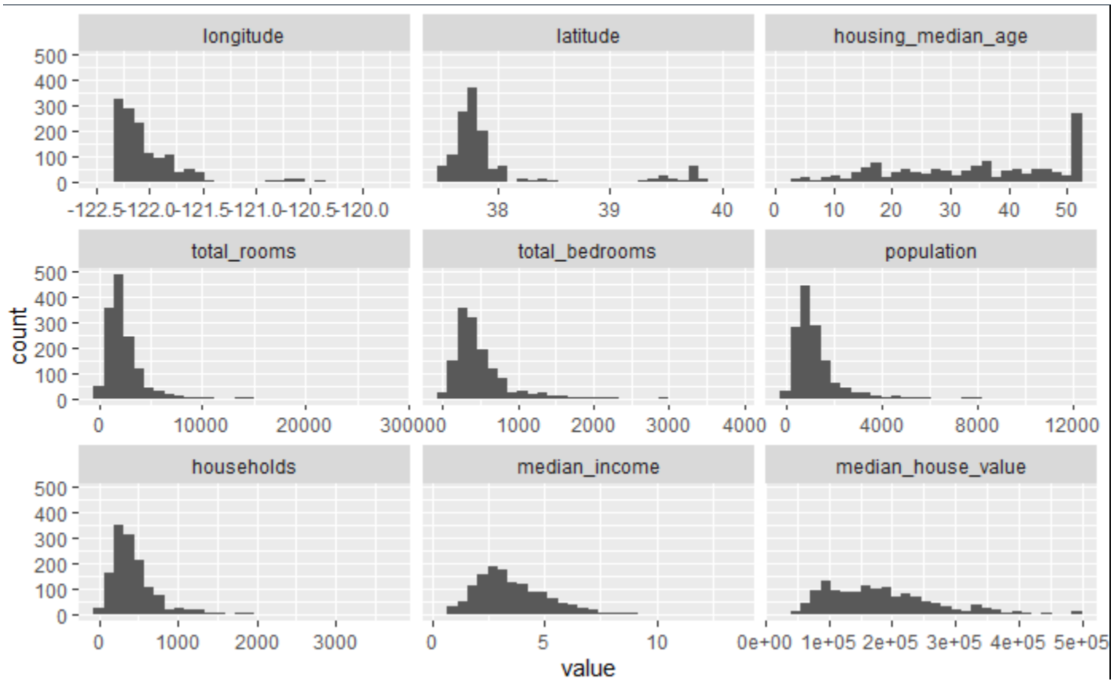
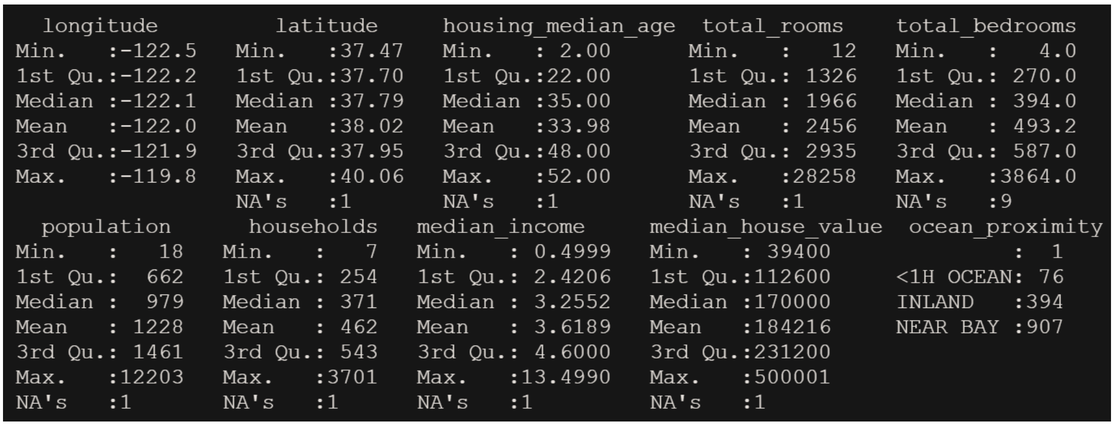


Fig. 1 Histogram of each column

As we can see, 'longitude' is bound up with 'latitude'. 'Longitude' is range from -122.5 to -121.5, while 'latitude' is range at 38 left and right. 'Housing_median_age' distributes averagely, the same as 'median_house_value'. The rest of other columns are unimodal in general.

In data mining, filling missing data is a significant step. This step is to ensure the integrity of dataset. So how to find missing data, and how to fill them is a research point in this project. To find missing data, I summary all columns in Fig. 2 to look attributes of them. Among of these attributes, number showed in NA’s represent the missing amount is each column.



```
  longitude         latitude     housing_median_age  total_rooms     total_bedrooms
 Min.   :-122.5   Min.   :37.47   Min.   : 2.00      Min.   :   12   Min.   :   4.0
 1st Qu.:-122.2   1st Qu.:37.70   1st Qu.:22.00      1st Qu.: 1326   1st Qu.: 270.0
 Median :-122.1   Median :37.79   Median :35.00      Median : 1966   Median : 394.0
 Mean   :-122.0   Mean   :38.02   Mean   :33.98      Mean   : 2456   Mean   : 493.2
 3rd Qu.:-121.9   3rd Qu.:37.95   3rd Qu.:48.00      3rd Qu.: 2935   3rd Qu.: 587.0
 Max.   :-119.8   Max.   :40.06   Max.   :52.00      Max.   :28258   Max.   :3864.0
                  NA's   :1       NA's   :1          NA's   :1       NA's   :9
  population       households    median_income     median_house_value  ocean_proximity
 Min.   :   18   Min.   :   7   Min.   : 0.4999   Min.   : 39400              :  1
 1st Qu.:  662   1st Qu.: 254   1st Qu.: 2.4206   1st Qu.:112600     <1H OCEAN: 76
 Median :  979   Median : 371   Median : 3.2552   Median :170000     INLAND   :394
 Mean   : 1228   Mean   : 462   Mean   : 3.6189   Mean   :184216     NEAR BAY :907
 3rd Qu.: 1461   3rd Qu.: 543   3rd Qu.: 4.6000   3rd Qu.:231200
 Max.   :12203   Max.   :3701   Max.   :13.4990   Max.   :500001
 NA's   :1       NA's   :1      NA's   :1         NA's   :1
```

Fig. 2 Summary of data

We can find the maximum missing data exist in ‘total_bedrooms’. Such ‘latitude’, ‘housing_media_data’, ‘total_rooms’ only have 1 missing value. I replace the missing data with the average value of each column, because all of them are numeric feature. In term of feature engineering, the main work I do in this stage is categorize them like label encoder. There are 3 different position in ocean proximity. I assign them as 1, 2 and 3.

Estimation of classification algorithm will execute on testing data. Thereby, I split the dataset into training set and testing set. The ratio I choose is 3: 1, that means, 75 percent of dataset will casted as training entries. Shuffle data is also a trick method in data mining, therefore I sample the indices of data entries in a disorder manner.

After preparing steps, I can build machine learning model to perform classification. Procedure can be separated as training phase and testing phase. Two classic models are used in this project, they are decision tree and random forest respectively.

As for decision tree, I feed training data and training target as inputting, model will select the optimal feature as splitting criteria. Finally, a tree shaped by the decision process use data feature. The final tree is shown as below:

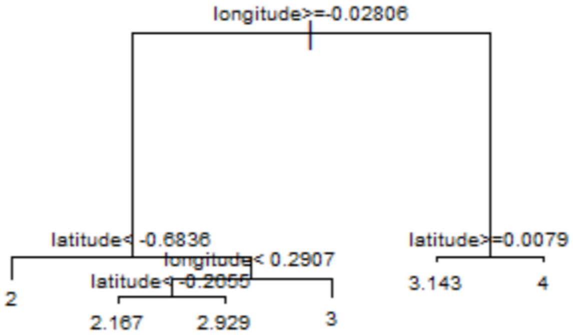
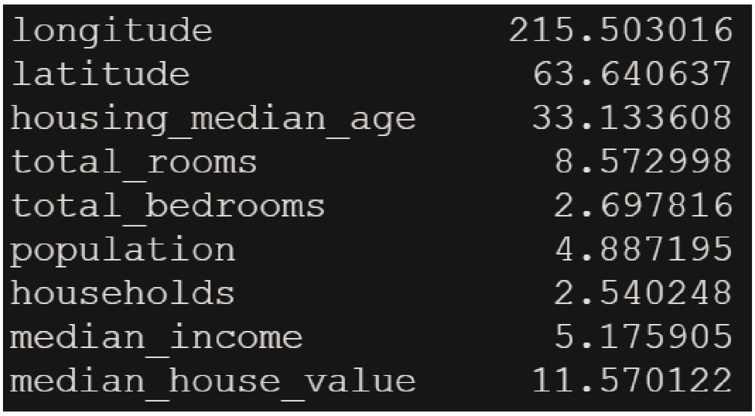


Fig. 3 Decision tree

From the tree, we can see that longitude and latitude are the most important feature to determine the position of ocean proximity. But how about other feature? How can they explain the ocean proximity? Here I introduce random forest, a boosting set of decision tree. It is very useful to research importance of each feature. Experimental result shows below:



```
longitude             215.503016
latitude               63.640637
housing_median_age     33.133608
total_rooms             8.572998
total_bedrooms          2.697816
population              4.887195
households              2.540248
median_income           5.175905
median_house_value     11.570122
```

Fig. 4 Importance of each feature

Except of longitude and latitude, houing_meadian_age is another vital factor.

The last step is testing the model. On testing dataset, accuracy is the major index to evaluate the model performance. Impressive accuracy on testing dataset is 92.4% with decision tree.