7/29/2021

Pranav Pai

REPORT

Regression and Classification using machine learning

Contents

[Regression 3](#_Toc78474578)

[Dataset 3](#_Toc78474579)

[Data exploration 3](#_Toc78474580)

[Linear Regression 4](#_Toc78474581)

[Results 5](#_Toc78474582)

[Ridge Regression 6](#_Toc78474583)

[Results 6](#_Toc78474584)

[Classification 6](#_Toc78474585)

[Dataset 6](#_Toc78474586)

[Logistic Regression 8](#_Toc78474587)

[Results 9](#_Toc78474588)

[Random Forest 10](#_Toc78474589)

[Results 11](#_Toc78474590)

[Unsupervised Learning (Clustering Analysis) 12](#_Toc78474591)

[k-means clustering 13](#_Toc78474592)

[Results 13](#_Toc78474593)

[References 14](#_Toc78474594)

# Regression

Regression is a machine learning technique which is used to predict a variable of our interest. The output variable which is also called target/dependent variable is continuous in nature. For this work, I have selected **linear regression ridge regression**. These techniques along with obtained results with be explained later. First, I am going to describe the dataset I have selected for regression analysis. The dataset is selected from UCI portal, which is a big repository for many datasets.

## Dataset

The dataset is taken from the below link <https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>

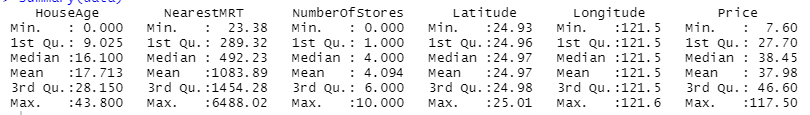
It is a dataset about real estate valuation [1]. The target is to predict the price of a property. The real estate market historical data collection was gathered from Sindian Dist., New Taipei City, Taiwan. The attributes of each property are given in table 1.

|  |  |
| --- | --- |
| Attribute | Description |
| X1 | The transaction date (e.g., 2013.250=March 2013, 2013.500=June 2013, etc.) |
| X2 | The age of the house (unit: year) |
| X3 | The distance between the nearest MRT station and your location (unit: meter) |
| X4 | The number of convenience stores within walking distance of the living circle (integer) |
| X5 | The geographic coordinate, latitude. (Unit: degree) |
| X6 | The geographic coordinate, longitude. (Unit: degree) |
| Y | House price per square meter (10000 NTD/Ping, where Ping is a local unit, and 1 Ping equals 3.3 meter squared) |

Table 1: Attributes

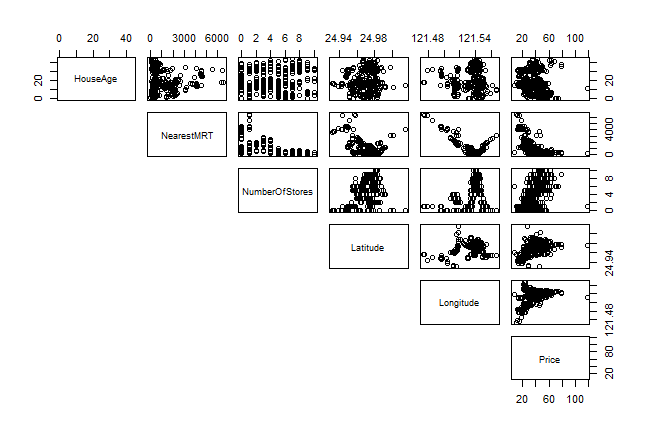
## Data exploration

First, I tried to understand the dataset. For this purpose, I used the summary function of R which produced the following information about the dataset (see figure 1).



**Fig.1:** Information of each attribute

There are 414 rows and 6 columns in the dataset.



**Fig.2:** correlation between attributes

The correlation analysis of the dataset is shown in figure 2. It can be seen from the that almost all attributes are positively correlated to the target variable price, except houseAge and NearestMRT.

## Linear Regression

A linear model is one in which the input variables (x) and the single output variable (y) are assumed to have a linear relationship (y). That y can be determined using a linear combination of the input variables is more detailed (x). The procedure is known as simple linear regression when there is only one input variable (x). When there are several input variables, the procedure is referred to as multiple linear regression in statistics literature. To construct or train the linear regression equation using data, various strategies can be applied, the most common of which is termed Ordinary Least Squares. Ordinary Least Squares Linear Regression, or simply Least Squares Regression, is a term used to describe a model created in this manner.

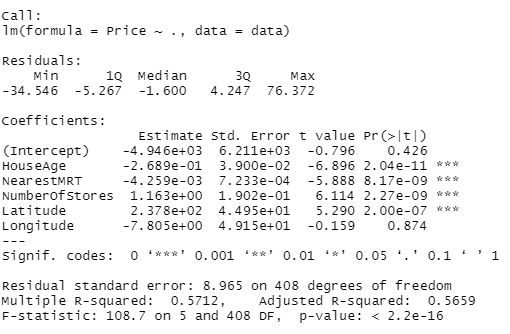
The representation is a linear equation that combines a collection of input values (x), with the solution being the projected output for that set of input values (y). As a result, both the input (x) and output (y) values are numeric. Each input value or column is assigned one scale factor, referred to as a coefficient and denoted by the capital Greek letter Beta in the linear equation (B).

One more coefficient is added, which gives the line an extra degree of freedom (for example, going up and down on a two-dimensional plot) and is known as the intercept or bias coefficient. For example, in a simple regression problem (with only one x and one y), the model would have the following form:

When there are several inputs (x) in higher dimensions, the line is called a plane or a hyper-plane. As a result, the representation is the equation's form as well as the coefficients' specific values (e.g. B0 and B1 in the above example). Normally, gradient descent algorithm is used to estimate the values of weights (bias and intercept). The goad is the minimize the error between predicted values and actual values. Mean squared error is a common error function used for linear regression.

### Results

The t-statistic examines whether there is a significant link between the predictor and the outcome variable, that is, whether the predictor's beta coefficient is significantly different from zero, for a given predictor. To calculate linear regression, I have used lm function, which produced the following results.



**Fig.3:** linear regression results

The squared error is **0.57** while p-value is less than **2.2e016**. When the p-value is smaller than this threshold, it shows the statistical significance of this linear model. The obtained weights and bias in enclosed in red rectangle in figure 3.

## Ridge Regression

Ridge regression is a variation of linear regression in which the loss function is changed to reduce the model's complexity. This is accomplished by adding a penalty parameter that is equal to the square of the coefficients' magnitude.

Here, OLS stand for Ordinary Least Square error. L2 regularization is another name for ridge regression.

### Results

The latter requires to tune a hyperparameter called lambda, which is one of the fundamental differences between linear and regularized regression models. The lambda values which are tested for these experiments ranges from **100 to 0.0001**. Best results were obtained for lambda equal to **0.158493**. The model is evaluated using RMSE (Root Mean Square Error) and R Squared Error. The obtained RMSE is equal to **8.90**, R Squared is equal to **0.5708**.

# Classification

Classification is also, a machine learning discipline which comes under the umbrella of supervised learning. The target variable for classification is categorical. For classification technique accuracy along with loss(error) is also calculated. I have used logistic regression and random forest for this task. The dataset is downloaded from UCI portal. First, I will try to explore the dataset and then will explain these classification methods and finally, I will discuss the results.

## Dataset

I have downloaded the dataset using the following link.

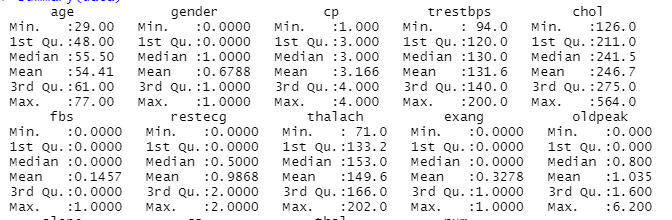
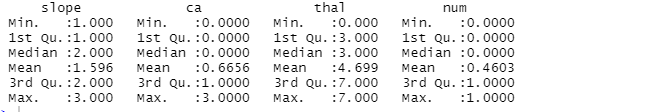
<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

The selected dataset is about heart disease [2]. It has two version. A processed version with 14 columns and a full version 76 version. Most people have used the processed version and it is also recommended by UCI portal to use this brief version. Therefore, I have also selected the processed version. Table 1 shows the attributes.

|  |  |
| --- | --- |
| Attribute | Description |
| Age | age in years |
| gender | 1 = male; 0 = female |
| Cp | chest pain type; there are 4 types |
| Trestbps | resting blood pressure (in mm Hg on admission to the hospital |
| Chol | serum cholestoral in mg/dl |
| Fbs | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false |
| Restecg | resting electrocardiographic results |
| Thalach | maximum heart rate achieved |
| Exang | exercise induced angina (1 = yes; 0 = no) |
| Old peak | ST depression induced by exercise relative to rest |
| Slope | the slope of the peak exercise ST segment |
| Ca | number of major vessels (0-3) colored by flourosopy |
| Thal | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| num | Target variable (diagnosis of heart disease (angiographic disease status)) |

**Table 2:** Attributes

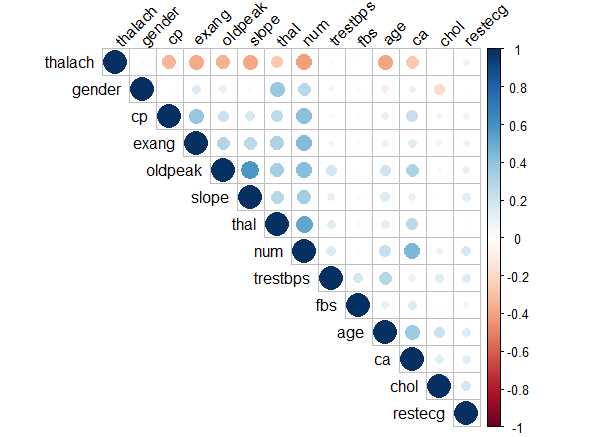
The target is to predict whether a person has heart disease or not (1 or 0). So, it is a binary classification problem. In this dataset 163 records are normal while 139 records belong to people with heart diseases i.e., **class 0: 163, class 1: 139**.

**Fig.4:** summry result of the selected dataset.

As shown in figure 4 the people whose data is collected were of age between 29 to 77 years. Their resting blood pressure ranges between 94 and 200.

Then I performed the correlation analysis, to understand that which attribute is high correlated with the target variable. The obtained results are shown in figure 5.



**Fig. 5:** Correlation analysis of the dataset

As shown in figure 5, gender, cp, exang, oldpeak, slope and thal are **positively correlated** with num (target variable) while thalach is **negatively correlated** with the target variable. The rest of the variables have very minor correlation, so actually, they can be ignored at classification level. But, in my case I used all 13 attributes.

The dataset is partitioned into training and testing using 70:30 ratio. For classification I have used logistic regression and random forest.

## Logistic Regression

Logistic regression is a binary classification technique. It is nonlinear in nature. The purpose of logistic regression is to determine the model that best describes the connection between a dependent variable and a set of independent variables. The coefficients of a formula to forecast a logit transformation of the likelihood of presence of the feature of interest (together with their standard errors and significance levels) are generated by logistic regression:

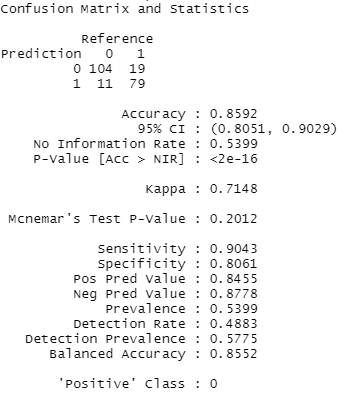
where p denotes the likelihood of the desired attribute being present. The logged odds are used to define the logit transformation.

And

Rather than minimizing the sum of squared errors (as in ordinary regression), logistic regression estimation maximizes the likelihood of witnessing the sample values.

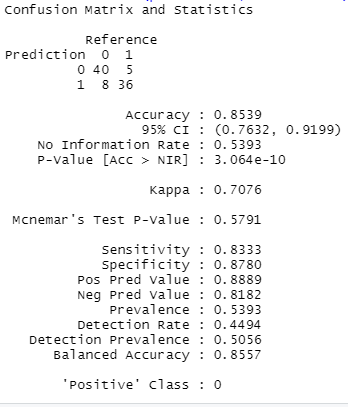
### Results

I have generated results for both training and testing dataset. The training accuracy is confusion matrix is shown in figure 6.



**Fig. 6:** Training results for logistic regression

The obtained results show that normal case is dealt with good accuracy as its sensitivity is more than specificity. The testing results are shown in figure 7.



**Fig. 7:** Testing results for logistic regression

The accuracy, sensitivity and specificity are compared in table 3.

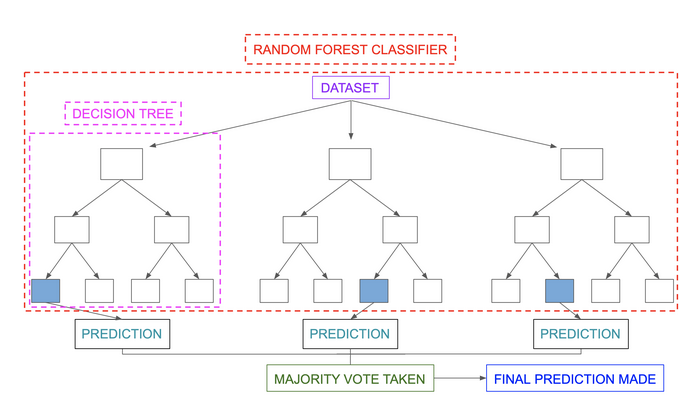
|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity |
| Training | 0.8592 | 0.9043 | 0.8061 |
| Testing | 0.8539 | 0.8333 | 0.8780 |

Table 3: performance comparison of logistic regression for training and testing dataset

Surprisingly, the trained model performed better for testing dataset for abnormal cases (heart disease cases).

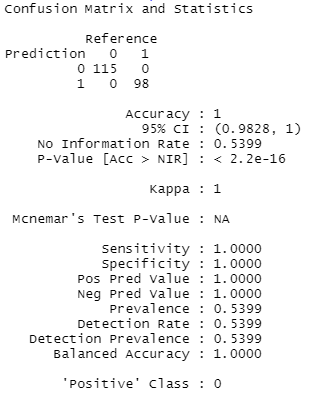
## Random Forest

A random forest is a supervised machine learning system that uses decision tree algorithms to build it. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. Many decision trees make up a random forest algorithm. The random forest algorithm's generated 'forest' is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them together. The (random forest) algorithm determines the outcome based on decision tree predictions. It forecasts by averaging or averaging the output of various trees. The precision of the result improves as the number of trees grows. Various decision trees are trained using the training data. This dataset contains observations and features that will be chosen at random when nodes are split. Random Forest is shown in the bellow figure.

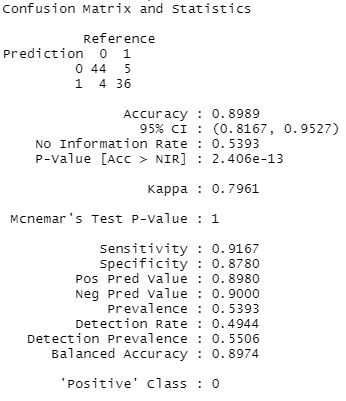


### Results

The obtained results are shown in figure 8 for training and figure 9 for testing.



**Fig. 8:** Training results for Random Forest



**Fig. 9:** Testing results for RF

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity |
| Training | 1 | 1 | 1 |
| Testing | 0.8989 | 0.9167 | 0.8780 |

**Table 4:** Results comparison of RF for training and testing datasets

The obtained results show that Random Forest has performed better than logistic regression on this dataset. For training dataset, its accuracy is 100%. Very high accuracy some time leads to overfitting.

# Unsupervised Learning (Clustering Analysis)

Unsupervised learning is a type of machine learning in which there is no target variable or class label. The model performed some statistical analysis of the dataset and derive useful information. Clustering is a technique for identifying groupings of observations in a set of data. When clustering, we require observations from the same group to have similar patterns, and observations from separate groups to have dissimilar patterns. In other words, the objects within same clusters are required to be same while objects in different clusters should be different from each other. The goal is minimizing the with in cluster distance and maximize the between clusters distance.

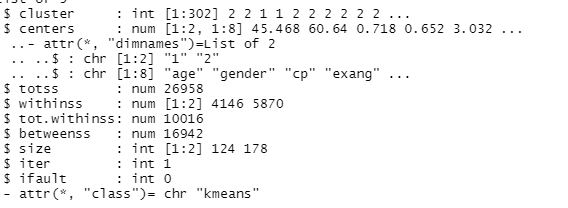
For this task I have selected k-means clustering and I am using the same dataset used in classification. The only difference is that I have discarded the target variable (num). Also, some of the columns which is very less correlation with the target variable are discarded.

## k-means clustering

When employing k-means clustering, the initial step is to specify how many clusters (k) will be formed in the final solution. The procedure begins by selecting k items at random from the data set to act as the clusters' initial centers. Cluster means or centroids are other names for the chose items. The remaining objects are then assigned to their closest centroid, which is determined by the object's Euclidean distance from the cluster mean. Now that the centers have been recalculated, each observation is double-checked to see if it belongs to a separate cluster. Iteratively repeat the cluster assignment and centroid update stages until convergence.

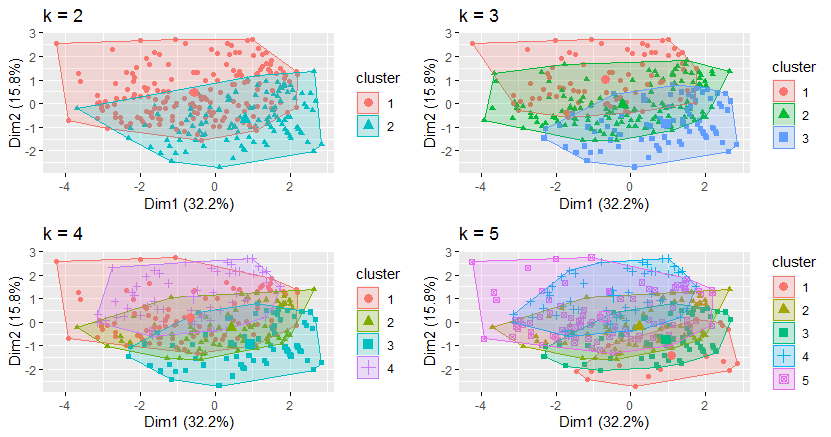
### Results

As mentioned before that the first issue that is faced by k-means algorithm is center point initialization. The second problem the selection of number of clusters. If k>> than the number of clusters naturally present in the dataset, it will break a compact cluster without any purpose. If k<< than the number of clusters present in the dataset then, two or more clusters may be combined, leading to uncompact clusters. The obtained results show that the k-means performed poorly on this dataset. The with in clusters and between clusters distance is almost same.



**Fig. 10:** Results of k-means clustering

I have checked multiple number of clusters. The results are shown in figure 11.



**Fig. 11:** Different values of k

# References

[1] Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

[2] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.