The research of regression model in machine learning field

Shen Rong[1] Zhang Bao-wen[2]

[1] School of information engineering of Ningxia University, YIN Chuan, China,750021
[2] School of mathematics and statistics,YIN Chuan,China,750021

Abstract: The paper herein will analyze the sale of iced products affected by variation of temperature. Firstly, we will collect the data of the forecast temperature last year and the sale of iced products and then conduct data compilation and cleansing. Finally, we will set up the mathematical regression analysis model based on the cleansed data by means of data mining theory. Regression analysis refers to the method of studying the relationship between independent variable and dependent variable. Linear regression model that corresponds to the practical situation is proposed in the paper, which is to set up simple linear regression model based on practical problem and then to implement the following with the help of the latest and most popular Python3.6. Python3.6 boasts the features of pure object-oriented, platform independence and concise and elegant language. So we will call the corresponding library function to predict the sale of iced products according to the variation of temperature, which will provide the foundation for the company to adjust its production each month, or even each week and each day. As a result, the situation of over-production can be avoided. Moreover, the other situation as the profit will be affected by the lack of production since the rise of temperature will also be avoided. So the regression model also has reference value for the other fields of marketing.

1 INTRODUCTION

With the increasing growth of the Internet, various kinds of data have been increasing in an explosive way. For example, twitter produced the data at the volume of 8TB each day[1] in 2011, Facebook produced the data at the volume more than 500TB a day in 2012[2] and moreover as the biggest Internet company, Google might work on the data at the volume of 20PB each day in 2008[3]. So, it will be so important to collect, analyze and model a large amount of data in order to conduct valuable prediction in various fields. Nowadays many machine learning systems running in data center have emerged continuously, such as Graph Lab[4] , Mahout[5] , Mad LINQ[6] and so on. Among them, Mahout based on distributed machine learning library has been widely used by lots of corporations such as Yahoo, Twitter, Linked In and so on. Under this background, Python, as the most popular programming language in the field of machine learning, has been used more and more widely. The paper herein will conduct the programming analysis on the model based on the latest Python3.6.

Python3.6 is a generalized programming language with strong function, that has been widely used and is easy to learn at the same time. Moreover, it fits for software programming of various scale [7]. Nowadays Python3.6 has mainly been used in fields of data mining, analysis and some others. Python boasts the features of unified writing form , strong readability, diverse algorithms and rich data structures, which means it can package both data structures and general algorithms. With its various advantages, it can not only integrate with other programming languages but also maximum its advantage of mixed programming language, leading to its great convenience of utility. Prediction analysis is one problem that has to be solved in data mining field. Generally speaking, we can choose the most suitable algorithm and model to conduct the analysis according to the practical case. The paper herein will mainly study on how to define the fitting relations between independent and dependent variables and then confirm parameters of the model, finally go back to the assumed equation to predict the variation tendency of dependent variable based on the typical linear regression model of predicting the sale of iced products according to the temperature variation. In the paper, we will establish linear regression model of one-variable with the help of Python3.6, which may predict the effect of temperature variation on the sale of iced products ideally.

2 PYTHON 3.6

2.1 Introduction

In recent years, further study [8-9] has gradually been the research focus of machine learning field, which may give rise to the prevailing of the object-oriented programming, Python3.6 in the field of machine learning and being well-received by learners.

There have been many new features added to Python3.6 on the basis of Python 2.7, such as formatting string and literals, variable comment syntax, underlining literals, asynchronous generator, asynchronous derivation and so on. Python3.6 has showed its great advantage on data analysis and data mining. Python3.6 has attracted great attention because of its convenience of learning and strong functions and boasted great advantage in data analysis and data mining.

2.2 Advantage

Python3.6 is an object-oriented language design program with pure script, which combines essence and designing rules of various design languages showing the features of interactive connections and type of explanation. Python has been focused by researchers due to its concise, elegant and clear language. Google was the first to use the Python as the development tool of network applications.

Python3.6 is a pure object-oriented language that can be used in the development of large-scale software due to its object-oriented mechanism, efficient execution and platform independence. It can express almost every comprehensive object with advanced data of tuple, list, dictionary and so on.

Python3.6 can write codes while running and each code can be tested immediately after being finished, which can dramatically improve the efficiency of engineers. Moreover, procedures written by c/c++ can be easily changed into the expansion mode of Python due to its strong expansibility. Python3.6 simplifies the duplicated codes in the software giving rise to the feature of permitting dynamic construction and execution of procedure.

The paper herein will predict the sale of iced products affected by the variation of temperature based on the library function such as linear regression by concise programming language.

3 DATA MINING

3.1 Introduce

The Huge amount of data triggers the swift development of data mining in the age of big data. One experienced data analyst will collect data first and then cleanse, analyze, and model it, which may have a close relation with data mining theory. The classic examples of application of data mining are the influenza trend forecast service pushed out by Google and "election of big data" [11] of Obama team. Domestic scholars also start related researches such as Meng Xiaofeng who systematizes and concludes the concepts, technologies and challenges of big data management; Hou Jingchuan who studies the quotation of data in the age of big data and has a deeper analysis and discussion on its current situation, latest development and future improvement.

Nowadays commonly used algorithms of data mining can be divided into several kinds of classification, cluster, association rules and time series prediction. Data mining may mainly be used in aspects of banking service, telecommunication, information security and scientific study. Furthermore, the popular tools of data mining are Weka, statistical analysis software Spass, Clementine, Rapidminer, Orange, Knime, Keel, Tanagra and so on.

Sorting algorithms that are often used in data mining are mainly linear regression algorithm, logistic regression algorithm, Bayesian decision theory and classifier, Support Vector Mouhime proposed by Cortes and Vupnik in 1995. Among these SVM algorithms, they can also be divided into several kinds according to its linear situation and the kernel function is used in the widest way.

Clustering algorithms commonly used in data mining are mainly hierarchical clustering algorithm, partition clustering algorithm, clustering algorithm based on density, clustering algorithm based on network, clustering algorithm based on model which may also include statistical method and neural network method.

3.2 Trend

Data The paper herein may mainly propose the linear regression algorithm belonging to sorting algorithm, including data collecting, data cleansing. We will set the forecast temperature as the independent variable and the sale of iced products as the dependent variable. The results of data analysis show that the factors which may influence the sale of the company are chosen totally correct.

4 LINEAR REGRESSION MODEL

4.1 Theoretical

Linear regression analysis can be divided into simple linear regression and multiple linear regression. The paper will mainly analyze simple linear regression model that is the analysis method of studying the relations between independent variable and dependent variable. We will set the model of dependent variable y and the independent variable x_i (i=1,2,3······) that will influence the variable y and the predict the development trend of y ,Simple linear regression model will be expressed as followed:

$$y = a_0 + a_1 x + e$$

y is the dependent variable and x is the independent variable. a_0 , the constant term, is the intercept of the regression line on the vertical axis and a_1 is regression coefficient that is the slope of the regression line. e is the random error which will be used to express the effect of random factors on dependent variable.

4.2 Methods

Regression analysis will evaluate a_0 and a_1 by observing the sample (x_i, y_i) in practical application. The paper will draw the scatter diagrams of dependent variable and independent variable by Python3.6 to evaluate the model parameter and establish regression model. After that, the regression equation and accuracy will be defined by the judgement of coefficient. For simple linear regression, coefficient of determination will be defined as:

$R^2 = cor^2(\hat{Y}, Y)$

$R^2 = ESS/TSS = 1 - ESS/TSS$

$$TSS = \sum (Y_i - \bar{Y})^2$$

$$TSS = \sum (\bar{Y}_i - \bar{Y})^2$$

$$TSS = \sum (Y_i - \bar{Y}_i)^2$$

$\hat{y}$ is the original value and y is the predicted value, so the equation will be showed here TSS is sum of squares for total and ESS is regression sum of squares. Moreover, RSS is residual sum of squares.

5 APPLICATION CASE ANALYSIS

5.1 Algorithm implementation

The paper will use Python3.6 to set up linear regression analysis model targeting at the effect of temperature variation on the sale of company. We introduce the unified operating system interface function in Python3.6 and all the functions in the numpy library that stores object variable. Moreover, we will also introduce Pandas analysis package and establish more advanced data structure and data analysis package of tool. The importing statements of Python3.6 will be showed as followed:

```
Import os;
Import numpy;
From pandas import  read_csv;
From matplotlib import pyplot as plt;
From sklearn linear_model import lineatRegression;
```

5.2 The comparison results

Draw the scatter diagram of the forecast temperature and sale of iced products after data cleansing, which will be showed as followed:

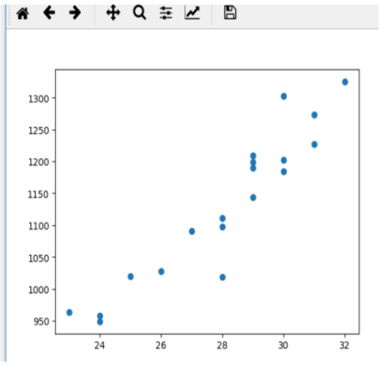
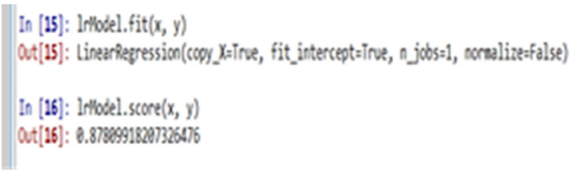


Fig.1 Scatter Diagram

Plt. ScatterEstimate model parameter andset up regression model, then introduce Linear regression function: LrModel=LinearRegression().

We can predict the linear relations between x and y by means of the sample of corresponding relations of variable x and variable y. The training model will be showed as followed: lrModel.fit(x, y) and lrModel.score(x, y), Test the regression model and the result will be showed in Fig.2.

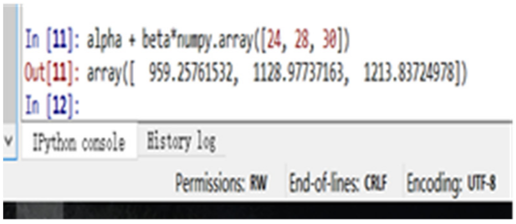


```
In [15]: lrModel.fit(x, y)
Out[15]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

In [16]: lrModel.score(x, y)
Out[16]: 0.8780991820732 6476
```

Fig.2 Test Result of Regression Model

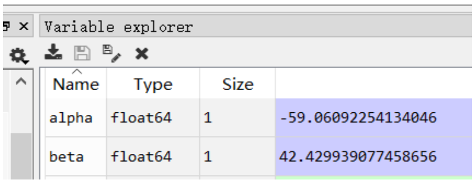
Predict by the regression model
lrModel.predict([[50], [40], [30]])
Intercept and parameter will be showed in Fig.3.
alpha = lrModel.intercept_[0]
beta = lrModel.coef_[0][0]



```
In [11]: alpha + beta*numpy.array([24, 28, 30])
Out[11]: array([  959.25761532,  1128.97737163,  1213.83724978])
In [12]:
```

Fig.3 Intercept and Parameter

Output the predicted values of sale under different temperatures and results will be showed in Fig.4.
alpha + beta*numpy.array([50, 40, 30])
alpha + beta*numpy.array([55, 45, 35)



Name	Type	Size	
alpha	float64	1	-59.0609225413404 6
beta	float64	1	42.42993907745865 6

Fig.4 Temperature Variation and the Sale of Iced Products
3

The result shows when the temperature is 24 degrees, the predicted value of sale will be 959.25; when the temperature is 28 degrees, the predicted value of sale will be 1128.977; when the temperature is 30 degrees, the sale will be 1213.837. So, the company can adjust the production according to the predicted value of sale, which will have great effect on the company. At the same time, the linear regression model used in the paper can obtain greater imitative effect of the predicted value of product sale based on the temperature variation and boast strong practicality. The latest Python3.6 used in the paper also shows its concise and elegant statements that fits for mathematics model. Python3.6 in data mining will be the first choice of learners of machine learning field.

6 CONCLUSION

The paper herein introduces the algorithm and model of the field of machine learning. Linear regression model is used to analyze the sale of iced products of company and the effect of temperature variation on the sale. Firstly, we cleanse the data collected one year ago and analyze data at the same time. Then we choose forecast temperature as the independent variable and the sale of iced products as the dependent variable to establish simple linear regression model for analysis. We use the object-oriented programming language Python3.6 and introduce the linear regression function. Programming language can also make data analysis in the field of data mining easier. The final result correctly leads the company to adjust the production and sale of iced products flexibly according to the variation of temperature, which definitely provides great commercial value and offers crucial theoretical foundation for the sale of other companies who produce iced products.

Acknowledgment

Fund project: Ningxia high school science research project: NGY2017032;

REFERENCES:

1. Krikorian R. Twitter by the numbers. 2010. http://www.slideshare.net/raffikrikorian/twitter-by-the-numbers.
2. Tam D. Facebook processes more than 500 TB of data daily.2012. http://www.cnet.com/news/facebook-processes-more-than-500-tb-of-data-daily/
3. Wikipedia,Petabyte.2014. http://en.wikipedia.org/wiki/Petabyte
4. Low Y,Gonzalez J, Kyrola A, Bickson D, Guestrin C, Hellerstein JM. Graphlab: A new framework for parallel machine learning. ar Xiv preprint ar Xiv:1006.4990. 2010.
5. The apache software foundation, what is apachemahout. 2014. http://mahout.apache.org/
6. Qian ZP, Chen XW, Kang NX, Chen MC, YuY,Moscibroda T, Zhang Z. Mad LINQ: Large-Scale distributed matrix computation for the cloud. In: Proc.of the 7th ACM European Conf. on Computer Systems. ACM Press, 2012. 197210. [doi: 10.1145/2168836.2168857]
7. Sun Q, Li JH, Li SH. Research on the development of text classification system based on Python[J].Computer application software,2011,28(3):13-14.
8. Hinton GE,Osindero S,Teh Y W.A fast learning algorithm for deep belief nets[J].Neural computation,2006,18(7):1527-1554.
9. Arel I,Rose D C,Karnowski T P.Deep machine learning-A new frontier in artifical intelligence esearch[J].Computational Intelligence Magazine,IEEE,2010,5(4):13-18.
10. Google Flu Trends. [EB/ O L]. http:
// www.google.org/ flutrends.
11. M. Scherer. Inside the Secret World of the Data
Crunchers Who Helped O bama Win. [EB/ O L].
(2012- 11- 07) [2013- 03- 06]. http:
//swampland.time.com/ 2012/ 11/ 07/ inside- the- secret- world- of- quants- and- data- crunchers- who- helped- obama- win/ .
12. Meng XF. Big data management: concept, technology and challenge.
[J]. Computer research and development, 2013,50 (1) : 146-169.
13. Hou JC, Fang JY. Data citation research: progress and prospect.[J]. Journal of Chinese library, 2013 (1) : 112-118.