A Study on 'Machine Learning Regression'

K Narayana Raju[1]*, K Pavan Kumar[2], Ch.S.V.Satyanarayana[3], D.Pujitha[4]

[1]Associate Professor, Dept. of Statistics, B.V.Raju College (A), Bhimavaram, A.P, India.
[2]Assistant Professor, Dept. of Statistics, B.V.Raju College (A), Bhimavaram, A.P, India.
[3]Associate Professor, Dept. of Mathematics, B.V.Raju College (A), Bhimavaram, A.P, India.
[4]Assistant Professor, Dept. of Statistics, B.V.Raju College (A), Bhimavaram, A.P, India.
*Corresponding Author

Received:05.07.2024 Revised: 20.07.2024 Accepted: 29.08.2024

ABSTRACT
Machine Learning Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes. Solving regression problems is one of the most common applications for machine learning models, especially in supervised machine learning. Algorithms are trained to understand the relationship between independent variables and an outcome or dependent variable. The model can then be leveraged to predict the outcome of new and unseen input data, or to fill a gap in missing data. Regression analysis is an integral part of any forecasting or predictive model, so is a common method found in machine learning powered predictive analytics. Alongside classification, regression is a common use for supervised machine learning models. This approach to training models required labelled input and output training data. Machine learning regression models need to understand the relationship between features and outcome variables, so accurately labelled training data is vital.

Keywords: algorithm, data, input, Machine Learning Regression, relationship, technique

INTRODUCTION

Regression is a key element of predictive modelling, so can be found within many different applications of machine learning. Whether powering financial forecasting or predicting healthcare trends, regression analysis can bring organisations key insight for decision-making. It's already used in different sectors to forecast house prices, stock or share prices, or map salary changes. This paper explores regression in machine learning, including what it is, how it's used, and the different types of regression in machine learning.

Machine Learning Regression

Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome. Outcomes can then be predicted once the relationship between independent and dependent variables has been estimated. Regression is a field of study in statistics which forms a key part of forecast models in machine learning. It's used as an approach to predict continuous outcomes in predictive modelling, so has utility in forecasting and predicting outcomes from data. Machine learning regression generally involves plotting a line of best fit through the data points. The distance between each point and the line is minimised to achieve the best fit line.

Alongside classification, regression is one of the main applications of the supervised type of machine learning. Classification is the categorisation of objects based on learned features, whereas regression is the forecasting of continuous outcomes. Both are predictive modelling problems. Supervised machine learning is integral as an approach in both cases, because classification and regression models rely on labelled input and output training data. The features and output of the training data must be labelled so the model can understand the relationship.

Usage of Regression Models

Machine learning regression models are mainly used in predictive analytics to forecast trends and predict outcomes. Regression models will be trained to understand the relationship between different independent variables and an outcome. The model can therefore understand the many different factors which may lead to a desired outcome. The resulting models can be used in a range of ways and in a variety of settings. Outcomes can be predicted from new and unseen data, market fluctuations can be predicted and accounted for, and campaigns can be tested by tweaking different independent variables.

In practice, a model will be trained on labelled data to understand the relationship between data features and the dependent variable. By estimating this relationship, the model can predict the outcome of new and unseen data. This could be used to predict missing historic data, and estimate future outcomes too. In a sales environment, an organisation could use regression machine learning to predict the next month's sales from a number of factors. In a medical environment, an organisation could forecast health trends in the general population over a period of time.

Supervised machine learning models are generally used for either classification or regression problems. Classification is when a model is trained to categorise an object based on its features. This could include facial recognition software, or to identify a spam email in a firewall. A model will be trained on labelled input and output data to understand the specific features which classify a labelled object. On the other hand, a regression problem is when a model is used to predict continuous outcomes or values. This could be a model that forecasts salary changes, house prices, or retail sales. The model is trained on labelled input and output data to understand the strength of relationships between data features and output.

Regression is used to identify patterns and relationships within a dataset, which can then be applied to new and unseen data. This makes regression a key element of machine learning in finance, and is often leveraged to help forecast portfolio performance or stock costs and trends. Models can be trained to understand the relationship between a variety of diverse features and a desired outcome. In most cases, machine learning regression provides organisations with insight into particular outcomes. But because this approach can influence an organisation's decision-making process, the explainability of machine learning is an important consideration.

Common use for machine learning regression models include:

a) Forecasting continuous outcomes like house prices, stock prices, or sales.
b) Predicting the success of future retail sales or marketing campaigns to ensure resources are used effectively.
c) Predicting customer or user trends, such as on streaming services or e-commerce websites.
d) Analysing datasets to establish the relationships between variables and an output.
e) Predicting interest rates or stock prices from a variety of factors.
f) Creating time series visualisations.

Types of Regression

There are a range of different approaches used in machine learning to perform regression. Different popular algorithms are used to achieve machine learning regression. The different techniques may include different numbers of independent variables or process different types of data. Distinct types of machine learning regression models may also assume a different relationship between the independent and dependent variables. For example, linear regression techniques assume that the relationship is linear, so wouldn't be effective with datasets with nonlinear relationships.

Some of the most common regression techniques in machine learning can be grouped into the following types of regression analysis:

a) Simple Linear Regression
b) Multiple linear regression
c) Logistic regression

Simple Linear Regression

Simple Linear regression is a linear regression technique which plots a straight line within data points to minimise error between the line and the data points. It is one of the most simple and basic types of machine learning regression. The relationship between the independent and dependent variables is assumed to be linear in this case. This approach is simple because it is used to explore the relationship between the dependent variable and one independent variable. Outliers may be a common occurrence in simple linear regression because of the straight line of best fit.

Multiple Linear Regression

Multiple linear regression is a technique used when more than one independent variable is used. Polynomial regression is an example of a multiple linear regression technique. It is a type of multiple linear regression, used when there is more than one independent variable. It achieves a better fit in the comparison to simple linear regression when multiple independent variables are involved. The result when plotted on two dimensions would be a curved line fitted to the data points.

Logistic Regression

Logistic regression is used when the dependent variable can have one of two values, such as true or false, or success or failure. Logistic regression models can be used to predict the probability of a dependent variable

occurring. Generally, the output values must be binary. A sigmoid curve can be used to map the relationship between the dependent variable and independent variables.

PYTHON 3.6

Introduction

In recent years, further study 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, a synchronous generator, a synchronous 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.

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 to 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, and 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.

Datamining

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” 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, Rapid miner, 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.

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.

Linear Regression Model

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 xx_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 follows:

$y$is the dependent variable and$x$is the independent variable.$a0$,the constant term, is the intercept of the regression line on the vertical axis and$a1$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.
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.

Evaluating a Machine Learning Regression Algorithm
Let's say you've developed an algorithm which predicts next week's temperature. The temperature to be predicted depends on different properties such as humidity, atmospheric pressure, air temperature and wind speed. But how accurate are your predictions? How good is your algorithm?
To evaluate the above predictions, there are two important metrics to be considered: variance and bias.

Variance
Variance is the amount by which the estimate of the target function changes if different training data were used. The target function f establishes the relation between the input (properties) and the output variables (predicted temperature). When a different dataset is used the target function needs to remain stable with little variance because, for any given type of data, the model should be generic. In this case, the predicted temperature changes based on the variations in the training dataset. To avoid false predictions, we need to make sure the variance is low. For that reason, the model should be generalized to accept unseen features of temperature data and produce better predictions.

Bias
Bias is the algorithm's tendency to consistently learn the wrong thing by not taking into account all the information in the data. For the model to be accurate, bias needs to be low. If there are inconsistencies in the dataset like missing values, less number of data errors in the input data, the bias will be high and the predicted temperature will be wrong.

The Bias-Variance Trade-off
Bias and variance are always in a trade-off. When bias is high, the variance is low and when the variance is low, bias is high. The former case arises when the model is too simple with a fewer number of parameters and the latter when the model is complex with numerous parameters. We require both variance and bias to be as small as possible, and to get to that the trade-off needs to be dealt with carefully, then that would bubble up to the desired curve.
Accuracy and error are the two other important metrics. The error is the difference between the actual value and the predicted value estimated by the model. Accuracy is the fraction of predictions our model got right.
For a model to be ideal, it's expected to have low variance, low bias and low error. To achieve this, we need to partition the dataset into train and test datasets. The model will then learn patterns from the training dataset and the performance will be evaluated on the test dataset. To reduce the error while the model is learning, we come up with an error function which will be reviewed in the following section. If the model memorizes/mimics the training data fed to it, rather than finding patterns, it will give false predictions on unseen data. The curve derived from the trained model would then pass through all the data points and the accuracy on the test dataset is low. This is called overfitting and is caused by high variance.
On the flip side, if the model performs well on the test data but with low accuracy on the training data, then this leads to underfitting.
There are various algorithms that are used to build a regression model, some work well under certain constraints and some don't. Before diving into the regression algorithms, let's see how it works.

Linear Regression in Machine Learning

Linear regression finds the linear relationship between the dependent variable and one or more independent variables using a best-fit straight line. Generally, a linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term (also called the intercept term). In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the nature of the regression line is linear. Mathematically, the prediction using linear regression is given as:

$$y = \theta_0 + \theta_1x_1 + \theta_2x_2 + … + \theta_nx_n$$

Here, y is the predicted value,

n is the total number of input features,

x_i is the input feature for i^{th} value,

θ_i is the model parameter (θ_0 is the bias and the coefficients are $\theta_1, \theta_2, … \theta_n$).

The coefficient is like a volume knob, it varies according to the corresponding input attribute, which brings change in the final value. It signifies the contribution of the input variables in determining the best-fit line.

Bias is a deviation induced to the line equation $y = mx$ for the predictions we make. We need to tune the bias to vary the position of the line that can fit best for the given data.

Drawing the Best-Fit Line

Now, let's see how linear regression adjusts the line between the data for accurate predictions.

Imagine, you're given a set of data and your goal is to draw the best-fit line which passes through the data. This is the step-by-step process you proceed with:

1. Consider your linear equation to be **$y = mx+c$,** where y is the dependent data and x is the independent data given in your dataset.
2. Adjust the line by varying the values of m and c, i.e., the coefficient and the bias.
3. Come up with some random values for the coefficient and bias initially and plot the line.
4. Since the line won't fit well, change the values of 'm' and 'c.' This can be done using the 'gradient descent algorithm' or 'least squares method'.

In accordance with the number of input and output variables, linear regression is divided into three types: simple linear regression, multiple linear regression and multivariate linear regression.

Least Squares Method

First, calculate the error/loss by subtracting the actual value from the predicted one. Since the predicted values can be on either side of the line, we square the difference to make it a positive value. The result is denoted by 'Q', which is known as the sum of squared errors.

Mathematically:

$$Q =\sum_{i=1}^{n}(y_{predicted}-y_{original})^2$$

Our goal is to minimize the error function 'Q." To get to that, we differentiate Q w.r.t 'm' and 'c' and equate it to zero. After a few mathematical derivations 'm' will be

$$m = \frac{cov(x,y)}{var(x)}$$

And 'c' will be,

$$c = y^{-} - bx^{-}$$

By plugging the above values into the linear equation, we get the best-fit line.

Gradient Descent

Gradient descent is an optimization technique used to tune the coefficient and bias of a linear equation.

Imagine you are on the top left of a u-shaped cliff and moving blind-folded towards the bottom center. You take small steps in the direction of the steepest slope. This is what gradient descent does — it is the derivative or the tangential line to a function that attempts to find local minima of a function.

Gradient descent help in minimizing the cost function

We take steps down the cost function in the direction of the steepest descent until we reach the minima, which in this case is the downhill. The size of each step is determined by the parameter α, called the learning rate. If it's too big, the model might miss the local minimum of the function, and if it's too small, the model will take a long time to converge. Hence, α provides the basis for finding the local minimum, which helps in finding the minimized cost function.

'Q' the cost function is differentiated w.r.t the parameters, m and c to arrive at the updated m and c, respectively. The product of the differentiated value and learning rate is subtracted from the actual ones to minimize the parameters affecting the model.

Mathematically, this is how parameters are updated using the gradient descent algorithm:

$$m = m - \alpha\frac{d}{dm}Q$$

$$c = c - \alpha\frac{d}{dc}Q$$
where $Q =\sum_{i=1}^{n}(y_{predicted}-y_{original})^2$.
This continues until the error is minimized.

Simple Linear Regression in Machine Learning
Simple linear regression is one of the simplest (hence the name) yet powerful regression techniques. It has one input (x) and one output variable (y) and helps us predict the output from trained samples by fitting a straight line between those variables. For example, we can predict the grade of a student based upon the number of hours they study using simple linear regression.
Mathematically, this is represented by the equation:
$$y = mx +c$$
where x is the independent variable (input),
y is the dependent variable (output),
m is slope,
and c is an intercept.
The above mathematical representation is called a linear equation.
Example: Consider a linear equation with two variables, 3x + 2y = 0.
The values which when substituted make the equation right, are the solutions. For the above equation, (-2, 3) is one solution because when we replace x with -2 and y with +3 the equation holds true and we get 0.
$$3 * -2 + 2 * 3 = 0$$
A linear equation is always a straight line when plotted on a graph.
In simple linear regression, we assume the slope and intercept to be coefficient and bias, respectively. These act as the parameters that influence the position of the line to be plotted between the data.
Imagine you plotted the data points in various colors, below is the image that shows the best-fit line drawn using linear regression.

Multiple Linear Regression in Machine Learning
Multiple linear regression is similar to simple linear regression, but there is more than one independent variable. Every value of the independent variable x is associated with a value of the dependent variable y. As it's a multi-dimensional representation, the best-fit line is a plane.
Mathematically, it's expressed by:
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$
Imagine you need to predict if a student will pass or fail an exam. We'd consider multiple inputs like the number of hours they spent studying, total number of subjects and hours they slept for the previous night. Since we have multiple inputs we would use multiple linear regression.

Multivariate Linear Regression in Machine Learning
As the name implies, multivariate linear regression deals with multiple output variables. For example, if a doctor needs to assess a patient's health using collected blood samples, the diagnosis includes predicting more than one value, like blood pressure, sugar level and cholesterol level.

CONCLUSION
Regression analysis is used to understand the relationship between different independent variables and a dependent variable or outcome. Models that are trained to forecast or predict trends and outcomes will be trained using regression techniques. These models will learn the relationship between input and output data from labelled training data. It can then forecast future trends or predict outcomes from unseen input data, or be used to understand gaps in historic data. As with all supervised machine learning, special care should be taken to ensure the labelled training data is representative of the overall population. If the training data is not representative, the predictive model will be overfit to data that doesn't represent new and unseen data. This will result in inaccurate predictions once the model is deployed. Because regression analysis involves the relationships of features and outcomes, care should be taken to include the right selection of features too.

REFERENCES
[1] Almasy LA, Dyer TD, Peralta JM, Kent JW, Jr, Charlesworth JC, Curran JE, Blangero J. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc. 2011;5(suppl 9):S2.
[2] Breiman L, Freidman JH, Olshen RA, Stone CJ. Classification and Regression Trees. CRC Press; Boca Raton, FL: 1984.
[3] Clarke B, Fokoue E, Zhang HH. Principles and Theory for Data Mining and Machine Learning. Springer; New York: 2009.

[4] Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinform. 2007;8:328. doi: 10.1186/1471-2105-8-328.
[5] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1–26.
[6] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.
[7] Elisseeff A, Evgeniou T, Pontil M. Stability of randomized learning algorithms. J Mach Learn Res. 2005;6:55–79.
[8] Evgeniou T, Pontil M, Elisseeff A. Leave one out error, stability, and generalization of voting combinations of classifiers. Mach Learn. 2004;55:71–97.
[9] Friedman JH, Hall P. Technical report. Stanford University; Stanford, CA: 2000. On bagging and non-linear estimation.
[10] Grandvalet Y. Bagging equalizes influence. Mach Learn. 2004;55:251–70.
[11] Hall DB, Shen J. Robust estimation for zero-inflated Poisson regression. Scand J Stat. 2010;37:237–52.
[12] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations. 2009;11(1):10–18.
[13] Hartigan JA, Wong MA. A K-means clustering algorithm. Appl Stat. 1979;28:100–8.
[14] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed Springer; New York: 2009.
[15] Hoerl AE, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67.
[16] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley; New York: 1990.
[17] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence; San Francisco, CA: Morgan Kaufmann; 1995. pp. 1137–1145.
[18] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34:1–14.
[19] Liu H, Zhang J. Estimation consistency of the group LASSO and its applications. J Mach Learn Res Workshop Conf Proc. 2009;5:376–83. [Google Scholar]
[20] McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed Chapman & Hall; New York: 1989.
[21] Meier L, van de Geer S, Buhlmann P. The group LASSO for logistic regression. J R Stat Soc Ser B. 2008;70:53–71.
[22] Meinshausen N, Yu B. LASSO-type recovery of sparse representations for high-dimensional data. Ann Stat. 2009;37(1):246–70.
[23] Nisbet R, Elder J, Miner G. Handbook of Statistical Analysis and Data Mining Applications. Academic Press; New York: 2009.
[24] Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann; San Francisco, CA: 1993.
[25] Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. J Comput Graph Stat. 2003;12:475–511.
[26] Ruczinski I, Kooperberg C, LeBlanc M. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. J Multivariate Anal. 2004;90:178–95.
[27] Schwarz DF, Konig IR, Ziegler A. On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics. 2010;26:1752–8. doi: 10.1093/bioinformatics/btq257.
[28] Sun YV. Multigenic modeling of complex disease by random forests. Adv Genet. 2010;72:73–99. doi: 10.1016/B978-0-12-380862-2.00004-7.
[29] Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV. Machine learning in genome-wide association studies. Genet Epidemiol. 2009;33(suppl 1):S51–7. doi: 10.1002/gepi.20473.
[30] Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B. 1996;58:267–88.
[31] Wilson AF, Ziegler A. A commentary on the lessons learned from the Genetic Analysis Workshop 17: Transitioning from genome-wide association studies to whole-genome statistical genetic analysis. Genet Epidemiol. 2011;X(suppl X):X–X. doi: 10.1002/gepi.20659.