Logistic Regression

A B M Abir Mahboob
Department of Electronic Engineering
Hochshule Hamm Lippstadt
Lippstadt, Germany
a-b-m-abir.mahboob@stud.hshl.de

***Abstract*—In this paper we will be looking at concepts related to logistic regression, as well as the statistical significance of individual regression coefficients. In this process we always get a binary outcome, which can have two values such as yes/no or true/false. When looking at modeling approaches we can use logistic regression to define a relationship between independent and dependent variables. We could use logistic regression to figure out which new sample fits where best therefore it can be a useful method for classification problems. We will also be looking at how robust logistic regression is as compared to linear regression and how it has helped solved problems that linear regression was not able to until now.**

I. Introduction

This section (including the sub-sections) is based on [1]. The field of Machine Learning emerged from the wider domain of Artificial Intelligence. Its goal is for machines to replicate human intelligence. The main question in the subject of Machine Learning is how to create machines capable of learning. In this context, learning is defined as logical interpretation, which occurs when one encounters instances that provide incomplete knowledge about a statistical fact. Unsupervised learning often involves trying to find hidden patterns or abnormalities in the data. Each sample in supervised learning has a label attached to it. It is intended to be the response to a query regarding the scenario. The task is referred to as a classification issue when the label is discrete. And the problem of real-valued labels is known as the regression problem. Based on these instances, it is especially tempting to guess the outcome of subsequent instances before they are actually witnessed. As a result, learning entails not just remembering but also generalizing to previously encountered circumstances. [1]

Let's look at some of the situations where machine learning is used in order to better comprehend its applications. Examples of machine learning in use include personalized adverts on Google or Instagram, the self-driving Tesla, YouTube video recommendations, cyberfraud detection, Amazon product recommendations, disease diagnosis, weather forecasts, and Netflix movie and television program recommendations. In todays modern data-rich environment, all of these instances highlight the critical role that machine learning has started to play. Machines can assist with the sifting of information to find the valuable bits that contribute to significant improvements. In a wide range of sectors, we are already witnessing the application of this technology. The applications, requirements, and significance of machine learning have increased as a result of the field's ongoing progress. In recent years, the term "big data" has gained a lot of popularity. It is partly because machine learning has become more sophisticated, which helps evaluate those large amounts of huge data. The methods for extracting and interpreting data have also altered as a result of machine learning. It was made feasible by utilizing automated sets of generic approaches that have taken the role of outdated statistical methods.[1]

Machine learning success is dependent on the algorithms that power it. Without being expressly taught to do so, ML algorithms by adopting sample data, generate a numerical model in order to make predictions or choices. This method sometimes classified as "training data". This can highlight patterns in the data that organizations can utilize to enhance decision-making, maximize productivity, and collect meaningful data at scale. AI solutions that automatically automate processes and resolve data-based business challenges are built on top of machine learning (ML). Companies can use it to supplement or replace some human skills.[8]

Three categories of approaches are used in machine learning.

A. *Supervised Learning*

Supervised machine learning aims to develop a model that creates predictions in the face of uncertainty based on data. Utilizing a collection of existing input data and recognised responses to the data(output), supervised learning trains a model to generate reliable prediction for the response to inbound data. When the data for the prediction outputs is already available, supervised learning is employed. Classification and regression algorithms are utilized in supervised learning to generate prediction models. Logistic regression is a supervised machine learning method.[1]

Discrete responses are predicted using classification techniques. Some instances include categorizing a tumor as malignant or benign and determining whether an email is real or spam. Models for classification cohorts the incoming data into categories. Examples of typical uses include fraud detection, voice recognition, and medical imaging.

Continuous responses are predicted via regression algorithms. Changes in temperature and fluctuations in power consumption are two examples of when regression techniques can be employed. When operating with a data range or when the nature of a response is a real number, regression techniques are required. Two prominent implementations of supervised learning include algorithmic trading and forecasting electricity load.[1]

B. Unsupervised Learning

Unsupervised learning detects basic data patterns or underlying patterns. It is designed to draw inferences from datasets that have input data but no labeled responses. In unsupervised learning, clustering continues to be the most utilized algorithm. To uncover hidden patterns or groups in data, it is employed in exploratory data analysis. Market analysis, object identification, and DNA sequence analysis are some usages for cluster analysis.[1]

C. Reinforcement Learning

Algorithms for reinforcement learning interact with their surroundings by taking actions, identifying mistakes, and learning from successes or failures. Trial-and-error learning and delayed rewards are two of reinforcement learning's most important features. With the help of this technique, machines and software proxies can automatically select the best course of action in a given situation to enhance performance. To learn which action is optimal, the agent requires simple reward feedback, often referred as the reinforcement signal. [8]

II. Logistic Regression

This section (including the sub-section) is based on [2]. A statistical analysis approach called logistic regression uses previous observations from a data set to predict a binary result, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. For instance, a logistic regression might be used to forecast whether a candidate for elections will win or lose, or whether a smoker would get lung cancer or not. These simple choices between two options allow for binary results. Multiple criteria for input can be taken into account using a logistic regression model. In the case of an election, the logistic function could take into account several more elements in addition to the popularity of the candidate and any current or past controversies. It then rates new instances according to their likelihood of falling into one of two result groups based on historical information about past outcomes using the same input criteria.[2]

The technique of logistic regression has grown in significance in the field of machine learning. It enables machine learning algorithms to categorize inbound input based on past data. The algorithms get more accurate at predicting classes within data sets when further relevant data is added. A mathematical equation that roughly estimates the correlation between the many variables being modeled is what regression models fundamentally reflect or embody. Input and output data are used to train machine learning models, which then utilize recent data to predict results.[2]

Mathematical calculations to determine the effect of several factors on a certain result are simplified by the use of logistic regression.The generated models can be used to dissect how successful various treatments are in relation to one another for distinct population groups such as male or female, healthy or ill etc.[2]

In order to produce characteristics for other kinds of AI and machine learning approaches, converting raw data streams can also be done by logistic models. In reality, logistic regression, which deals with issues with two class values, is one of the widely utilized machine learning techniques for binary classification problems such as yes or no, male or female, this or that etc. The odds of probabilities may also be estimated using logistic regression. This involves figuring out how attributes and outcomes' probability relate to one another. In other words, it may be implemented to categorization by constructing a model that links the number of hours of study to the odds that a student would pass or fail. On the other hand, if the number of study hours is given as an attribute and the response variable has two possible values (fail or pass) the same model might be used to predict whether a certain student would pass or fail.[2]

Because it diminishes ambiguous plausibility calculations to simple arithmetic problems, logistic regression is essential in different sectors. Thus, the contribution of numerous components to a particular outcome can be readily modelled and then studied.[2]

When utilizing logistic regression, a few assumptions must be kept in mind. The variables must initially be independent of one another. Postal code and age, for example, may be utilized in a model, but postal code and city would not. Because a city can have many postal codes, they are all interdependent. When logistic regression is applied as a starting point for intricate machine learning and data science applications, other less obvious correlations between variables may get overlooked.[2]

The raw data for logistic regression should reflect distinct or independent events for accurate prediction. For instance, different people's individual viewpoints should be represented in a customer satisfaction survey. But if someone took the survey more than once from multiple email addresses to be eligible for a prize, the outcomes would be distorted.[2]

A sizeable sample is required for logistic regression as well. This might only be 15 instances of each variable in a model. However, as the likelihood of each possibility decreases, the requirement increases.[2]

III. Difference between Logistic Regression and Linear Regression

This section is based on [7]. Among the most widely used models in data science are both linear and logistic regression. They can quickly and easily do the computation thanks to open-source programs like Python and R.

To determine their affiliation, one or more than one independent variables and a discrete dependent variable are evaluated using linear regression. It is regarded as simple linear regression when there is just one independent variable and one dependent variable. The term "multiple linear regression" is used when there are more independent variables. Each type of linear regression seeks to identify the optimum line to fit a given group of data points. Usually, the least squares approach is used to compute it.[7]

Logistic regression is used to determine how one or more independent variables and a dependent variable are related just like linear regression. However, the distinction is that as opposed to a continuous variable, it is utilized to predict a categorical one. True or false, yes or no, one or zero, etc. are all examples of categorical variables. The logit function converts the S-curve into a straight line, which is another distinction between logistic regression and linear regression in terms of the unit of measurement. In order to properly estimate the output for the continuous dependent variable, the best fit line must be found using linear regression whereas in logistic regression the weighted sum of the inputs is run via a logistic regression activation function, which can translate values between 0 and 1.[2]

While regression analysis uses both models to predict future events, linear regression is often simpler to comprehend. Furthermore, logistic regression requires a moderately large sample size to accurately portray values across all response categories, whereas linear regression does not. The model might not have enough statistical power to find a substantial influence in the absence of a larger, more representative sample.[7]

IV. Logistic Regression Model

This section is based on [6]. Using logistic regression, a binary dependent (outcome) variable and one or more independent (predictor) variables are identified. A categorical variable that can only have two possible values or levels is called a binary (or dichotomous) variable, such as 1 or 0, yes or no, positive or negative. Figure 1 depicts a basic scenario with X as the solitary independent variable and the dependent variable ranging from zero to one. The probability that the dependent variable has a value of one appears to rise as the value of the independent variable increases. Calculating the probability of a specific outcome considering the value of the independent variable or variables is possible using logistic regression, as it is more often known.

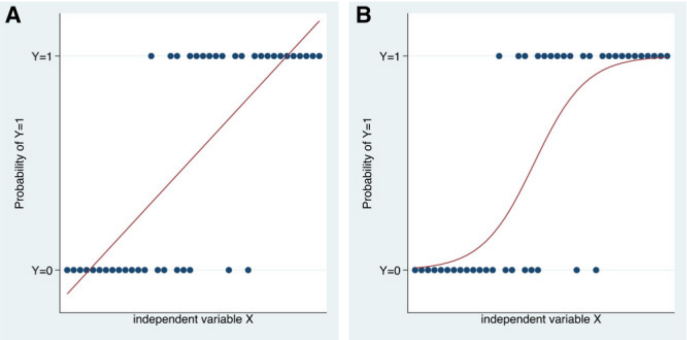

Fig. 1. The relation between a binary result Y, which can have values ranging from 0 to 1 and a continuous independent variable X.[6]

Logistic regression is basically a subset of linear regression. As shown in Figure 1(A), despite the fact that a linear correlation between the independent variable (X) and the probability of the outcome can be modeled; it is, however, unnatural since it would enable anticipated probabilities outside of the range of 0-1, but a prediction cannot be lesser than 0 or greater than 1. It presupposes a linear (straight line) correlation between the outcome and the logit (the natural logarithm of the odds). The slope (b1) and intercept (b0) of this line are represented by the regression coefficients.[6]

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

Figure 1(B) shows the sigmoidal relation between the probability and the independent variable that is created when this equation for the probability (P) is solved. As a result, the estimated probabilities are now suitably restricted to the range between 0 and 1.[6]

V. Advantages and Disadvantages of Logistic Regression

This section (including the sub-sections) is based on [3]. Some advantages and disadvantages of logistic regression are as follows:

A. Advantages

1. Easy to interpret, implement and analyze.

2. Allows to simply update models to incorporate new data.

3. Provides accurately well-calibrated probability as well as classification results.

4. Less prone to over-fitting in a low dimensional dataset.

5. Using a softmax classifier, this algorithm can easily be augmented to multi-class classification.[3]

B. Disadvantages

1. Logistic regression cannot tackle nonlinear problems since it has a linear decision surface.

2. It can only predict discrete functions.

3. Logistic Regression necessitates either average or no multicollinearity between independent variables.

4. Complex relationships are hard to determine when using logistic regression.[3]

VI. Application of Logistic Regression

This section (including the sub-sections) is based on [4]. Logistic regression is being used in a lot of different sectors of our daily life. In this section, applications of logistic regression in various sectors of life is discussed.

A. Medicine

Healthcare organizations can precisely pinpoint at-risk individuals who need a more individualized behavioral health strategy to assist them improve their everyday health behaviors by using logistic regression. In consequence, this creates the possibility for both patients' health and hospital expenses to improve.[5]

B. Gaming

The gaming business greatly benefits from logistic regression's speed, which is one of its advantages. In a game, speed is essential. The games that let you make in-game purchases to upgrade your character's gameplay abilities or to give them a more ostentatious appearance or to promote player interaction are quite popular right now. A recommendation system is best implemented in-game through purchases. The biggest video game company in the world is Tencent and it offers equipment recommendations for players using these platforms. In these situations, logistic regression is applied. The algorithm examines at a ton of user activity data and makes recommendations about what equipment a certain user would want to buy right away. Recommendation systems come in three different flavors. Based on user evaluations for comparable items they have previously purchased, as well as other activities, the interactive system can predict what the user might want to buy. The qualities listed in the item definition and the interests the user listed in the user's profile are what a content-based algorithm bases its judgment on. The hybrid is a blend of the two previous categories.[4]

C. Credit Scores

Logistic regression is one of the go to algorithm for credit scoring. A financial organization called ID Finance creates credit score prediction models. They require that their models be simple to interpret. A regulator can at any time inquire of them about a specific decision. Such a procedure as minimizing correlated variables is part of the data prepping for credit score models. Having a model with more than 15 variables is challenging. Finding out which factors have a greater and lesser impact on the predictions' outcome becomes simple using logistic regression. By exporting prediction results to an Excel file in the last step, analysts with or without technical skills can comprehend from the data.[4]

VII. Conclusion

In this paper, we first went through the what, how, and applications of machine learning. We spoke about several machine learning algorithms. Most significantly, we illustrate the fundamental logistic regression model, the similarities and differences between logistic regression and linear regression, and how logistic regression can be a strong analytical approach to apply when the outcome variable is dichotomous. The use of logistic regression has grown during the previous ten years. The JER and journals for higher education both clearly show the tendency. This popularity might be due to the ease with which academics can use sophisticated statistical programs that carry out thorough studies of this approach. The likelihood of increased adoption of the logistic regression approach is predicted. Given this possible expansion in usage, it is imperative that readers, editors, and researchers are instructed on what to anticipate from articles that employ the logistic regression method[9]. With an illustration of logistic regression applied to a data set and instructions and recommendations on a desired pattern of application of logistic techniques, it is anticipated that this research paper has answered some of the problems that have arisen.

VIII. Declaration of Originality

I, A B M Abir Mahboob, herewith declare that I have composed the present paper and work by myself and without the use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form have not been submitted to any examination body and have not been published. This paper was not yet, even in part, used in another examination or as a course performance. I agree that my work may be checked by a plagiarism checker.

03.07.2022& Hamm - A B M Abir Mahboob

Reference

1. Aery, M., and Ram, C. (2017). A Review on Machine Learning: Trends and Future Prospects. Research Cell: Int. J. Eng. Sci., 25, 89-96.

2. Lawton, G., Burns, E., amp; Rosencrance, L. (2022, January 20). What is logistic regression? - definition from Searchbusinessanalytics. SearchBusinessAnalytics. Retrieved July 2, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

3. Ranjan Rout, A. (2020, September 2). Advantages and disadvantages of logistic regression. GeeksforGeeks. Retrieved July 2, 2022, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

4. 5 real-world examples of logistic regression application: ActiveWizards: Data Science and Engineering Lab. ActiveWizards. (n.d.). Retrieved July 3, 2022, from https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application

5. Rojogan, B. (2018, March 26). Healthcare Analytics: Logistic Regression to reduce patient readmissions. Packt Hub. Retrieved July 3, 2022, from https://hub.packtpub.com/healthcare-analytics-logistic-regression-to-reduce-patient-readmissions/: :text=Using

6. Schober P, Vetter TR. Logistic Regression in Medical Research. Anesth Analg. 2021 Feb 1;132(2):365-366. doi: 10.1213/ANE.0000000000005247. PMID: 33449558; PMCID: PMC7785709.

7. What is logistic regression? IBM. (n.d.). Retrieved July 3, 2022, from https://www.ibm.com/topics/logistic-regression

8. Selig, J. (2022, April 11). What is machine learning? A definition. Expert.ai. Retrieved July 3, 2022, from https://www.expert.ai/blog/machine-learning-definition/

9. Peng, J. (2002, September). (PDF) an introduction to logistic regression analysis and reporting. ResearchGate. Retrieved April 6, 2022, from An Introduction to Logistic Regression Analysis and Reporting