Assessing the Risk of Having Heart Attacks with Machine Learning

Chenlisha Sun *

Cate School, 1960 Cate Mesa Rd, Carpinteria, CA, the USA

* Corresponding Author Email: sunnyyyyy0713@gmail.com

Abstract. Throughout the paper, what Machine Learning is and the Machine Learning technique known as Logistic Regression are explained. Afterward, this technique is used to analyze a set of data containing information about multiple patients. This information includes symptoms, as well as lab results such as cholesterol levels and whether the patient has a high chance of having a heart attack or not. The result of this analysis is a model that can be used on new patients to predict their risk of having a heart attack.

Keywords: machine learning, logistic regression, heart attack prediction.

1. Introduction

Heart Attack, a commonly known illness also called Myocardial Infarction, is a life-threatening condition caused by a blockage in the coronary arteries. The blockage will then lead to insufficient blood flow to the heart muscle. The blood flow then damages areas around the skin and even leads to necrosis, resulting in chest pain or discomfort, shortness of breath, sweating, and nausea. In this article, machine learning will be used to develop a model that predicts a patient's risk of having a heart attack.

Machine Learning is a subfield of artificial intelligence. It has been applied to numerous fields, and new applications are emerging constantly at a fast pace. This growth of Machine Learning has been possible due to two factors: the pervasive availability of data in different fields and the rapid growth of computational power during the last two decades. For more details, the reader is referred to the books [1-3].

The type of Machine Learning used in this article is Supervised Learning. In Supervised Learning, a data set is used to develop a computational model. This data set cannot be arbitrary. It has to consist of information about a collection of individual cases. These individual cases are referred to as examples in the Machine Learning jargon. In this study, the examples are the patients. The information about the examples is of two types, features, and a label. In this paper, the features of a patient are:

1. Age
2. Sex
3. The type of chest pain the patient experiences (there are four types: typical angina, atypical angina, non-anginal pain, and asymptomatic)
4. The resting blood pressure
5. The cholesterol level
6. The fasting blood sugar level
7. Resting electrocardiographic results
8. Maximum heart rate achieved
9. If the patient suffers from exercise-induced angina or not
10. Previous peak or old peak (ST depression induced by exercise relative to rest)
11. The slope of the peak exercise ST segment (there are three types: unsloping, flat, and down sloping)
12. The number of major vessels with cerebral amyloid angiopathy, which is a medical condition
13. Thalassemia rate (there are four types: null, fixed defect, normal, and reversible defect

14. Output, the target variable (it is divided into two types, less than 50% diameter narrowing means less chance of heart disease, and more than 50% diameter narrowing meaning more chance of heart disease)

The label is the chances of having a heart attack as either high chance or low chance.

In Supervised Learning, the development of the computational model requires the use of a data set known as the training set. Both the features and the labels of the examples in the training set are known. Once the computational model is developed, this model can be used to predict the label of new examples using only the features of these new examples as input. The labels of these new examples are not known, they are predicted by the computational model.

In this study, the data set used consists of 303 patients, all the features as well as the label of all the patients are provided by this data set. Once the computational model is developed, it can be used to predict the label of new patients. The label is the risk of the patient of having a heart attack. The features of the new patients will need to be provided to the model, but their labels are not known; the model will predict them. In other words, the model will play the role of a doctor diagnosing patients.

This article is organized as follows: In Section 2 describes the Binary Classification. Following, in Section 3, the supervised learning technique known as Logistic Regression is explained. Finally, in Section 4, Logistic Regression and the Heart Attack Data Set were used to develop the model. The article is then discussed and concluded in Section 5 and 6.

2. Background

2.1. Binary Classification

Problems such as the one considered in this paper, where the label of each example takes one of two values, are known as binary classification problems. In the problem of this article, the examples are the patients, and the two values that the label of each example can take are: high chance of having a heart attack and low chance of having a heart attack.

The first step in binary classification problems is to replace one of the possible labels with the number 1, and the other label with the number 0. From now on, the label of a patient will be 1 if the patient has a high chance of having a heart attack, but it will be 0, if the patient has a low chance of having a heart attack.

A model will be developed to predict the label of patients if the features of the patient are fed into the model as input. In other words, the model is a function, donated by y. The arguments of this function are the feature of the patient. The features are denoted by x_1, x_2 ..., x_{14}. The feature x_1 is the age of the patient, the feature x_2 is the sex of the patient, and the rest of the features are as described in the introduction.

Note that the sex of patients is not given as male or female. Instead, they are given as 1, if the patient is male, or 0, if the patient is a female. Similarly, all the other labels are also given as numbers. So, there is no need to worry about changing non-numerical information, such as the sex of each patient, into numbers.

As already mentioned, the model y is a function of the features and thus, is written as $y = y(x_1, x_2,...,x_{14})$. No matter the features of the example, a property of the model used for binary classification problems is that $0 < y(x_1, x_2,...,x_{14}) < 1$. Note that, while the real label of each example is either 0 or 1, the prediction the model gives is a number between 0 and 1. This number has to be interpreted as follows: if $0 < y(x_1, x_2,...,x_{14}) < 1/2$, the model predicts that the label of the example is 0. This means that the patient has a low risk of having a heart attack. If $1/2 < y(x_1, x_2,...,x_{14}) < 1$, the model predicts that the label of the example is 1. This means that the patient has a high risk of having a heart attack.

2.2. Binary Cross Entropy Error

This section describes a notion of an error. This error measures how well the prediction of the model is. Let y be the label of an example, and let y be the model's prediction. The binary cross-entropy error on this example is

$$bce(y,y) = ylog(y) + (1-y)log(1-y). \quad (1)$$

While this formula may seem complicated, what is essential is that the error *bce(y,y)* is always positive, and the closer y is to y, the smaller the error is. In fact, as y approaches y, the error *bce(y,y)* approaches zero. That is why *bce(y,y)* is a measure of the error the model makes in predicting the real label.

The error on a set of examples, not just on one example, is of particular interest. Assume there are n examples. Suppose y_i is the label of the i^{th} example and y_i is the prediction the model makes on the i^{th} example. Then, the binary cross-entropy error on this set is the average of the binary cross-entropy error on the examples of this set. In other words, the binary cross-entropy error on this set is

$$J = ln(bce(y_1,y_1) + bce(y_2,y_2) +...+ bce(y_n,y_n)) \quad (2)$$

The important property to remember is that, the smaller J, the binary cross-entropy error is, the better the model works on this set of examples. In fact, if the predictions were exact, the error would be zero.

3. Methodology

3.1. Logistic Regression

This section will explain the technique known as logistics regression. To that end, a function known as the sigmoid function needs to be introduced. This function is

$$(x) = 11 + e^{-x} \quad (3)$$

A plot of this function is given in the figure 1 below:

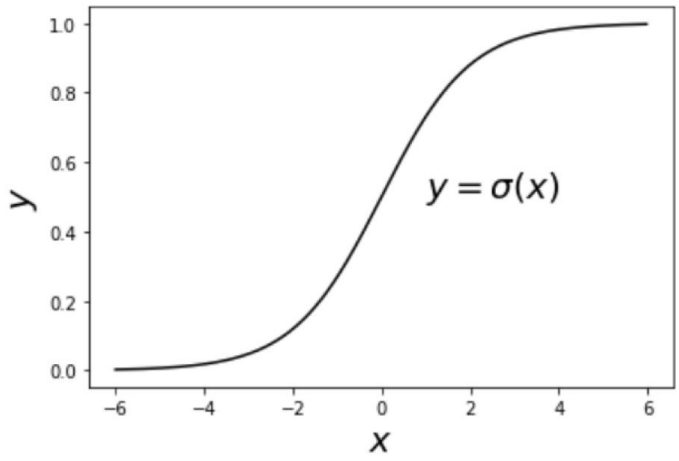

Fig 1. A plot of the sigmoid function.

The relevant properties of this function are:

1. The *sigmoid function = (x)* is increasing
2. The value of this function is always between 0 and 1. In other words, $0 < (x) < 1$ for all numbers x
3. As x increases, the value of *(x)* approaches 1
4. As x decreases, i.e. x becomes large in absolute value while remaining negative, *(x)* approaches 0
5. The sigmoid function at zero equals to one half, i.e. value of *(0) = 0.5*

Logistic regression assumes that the functional form of the model, i.e. function $y(x_1, x_2,...,x_{14})$ is

$$y(x_1, x_2,...,x_{14})=(w_1x_1+w_2x_2+...+w_{14}x_{14}+b) \quad (4)$$

where $w_1, w_2,....,w_{14}$, b are numbers known as parameters.

To find the parameters, a set of examples with known features and labels needs to be provided. This set of examples is known as the training set. Logistic Regression refers to the method that selects the parameters that make the binary cross-entropy error on the training set as small as possible. The details of the algorithm that is used to find such parameters will not be covered, but this algorithm is already embedded in popular coding libraries that are available to anyone.

3.2. Validation Set and the Accuracy of the Model

While the data set contains 303 examples, not all of the examples form part of the training set. A fraction of them, about 20%, selected randomly, are set aside and not used in the training of the model, i.e. in finding the parameters. These examples are set aside from a set of examples known as the validation set. Once the model is trained with the training set, it is tested on the validation set. The model's accuracy is the proportion of the examples in the validation set that the model predicts correctly. It is important to evaluate the accuracy of the model on the validation set instead of the training set, because the training set was used in the development of the model, but the validation set was not. Thus, the accuracy of the validation set provides how accurate the model will be when it is used to make predictions on new examples.

4. Applikation of Logistic Regression to Heart Attack Data Set

Logistic regression was implemented on the data set. The validation set contained 20% of the examples. The programming language Python was used, and more precisely, the popular TensorFlow and Keras libraries were employed for coding. The resulting model yielded an accuracy of 82% on the validation set.

Logistic regression is not the only technique to work on binary classification problems. Another popular technique is neural networks. Neural networks can be regarded as an extension of logistic regression. Neural networks were also tried on this problem, but the accuracy on the validation set decreased with the use of neural networks. This means that neural networks were too complex for the data set, and cause overfitting, an undesirable effect. The concept of overfitting will not be explained in this article. Suffy to say, that for this project, logistics regression works better than neural networks.

5. Discussion

Machine learning is a quickly expanding field with applications in many industries. In this article, machine learning applications in medicine were discussed, particularly in predicting the risk of heart attack. Although obtaining an accuracy of 82% is a promising start, it is important to think about the context of this finding in terms of the primary limitations to the study - a very small data set of only 303 patients. This limitation could have impacted the model's performance, and would be an area to develop in the future.

6. Conclusion

Overall, this research has constructed a logistic regression machine learning model that predicts a patient's risk of a heart attack. The model accuracy on the validation set was 82%, indicating that this approach has potential as a possible useful method for medical diagnostics. Clearly, future work with larger datasets, and more diverse datasets would help create more accurate and strong predictive models.

References

[1] Burkov A. The hundred-page machine learning book. Vol. 1. Quebec City, QC, Canada: Andriy Burkov; 2019.

[2] Carbonell JG, Michalski RS, Mitchell TM. An overview of machine learning. Machine Learning. 1983;3-23.

[3] Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd; 2019.