HIGH SCHOOL EDITION
Journal of Student Research

Volume 11 Issue 1 (2022)

Machine Learning as a Tool to the Diagnosis of Diabetes

Yosef Granillo[1] and Guillermo H. Goldsztein[#]

[1]John Hopkins University Center for Talented Youth, MD, USA
[#]Advisor

ABSTRACT

Machine learning is the field of computer science that uses data to make predictions and decisions. The problem we consider in this article belongs to the class known as supervised learning and the technique we use is logistic regression. After explaining supervised learning and logistic regression, we use a data set to develop a computational model able to give a diabetes diagnosis to patients. We discuss the accuracy of the model developed.

Introduction

Machine learning, also known as artificial intelligence, is a field of computer science that uses data to make predictions and decisions [1,2,3,4]. Machine learning has found applications in numerous fields. Examples include applications in the medical field, where machine learning is used for the diagnosis of diseases, such as heart disease, diabetes and pneumonia; applications in the banking business, where machine learning is used to make decisions on loan applications; applications to the real estate business, where machine learning is used to price real estate; applications to self-driving cars, where machine learning is at the core of the software used by self-driving cars; machines experts in playing chess; robots that can carry out numerous tasks.

Diabetes is a serious and debilitating chronic disease. It is becoming very common in the developed world. Thus, the efficient and early diagnosis of diabetes is very important, as changes in lifestyle, that includes changes in diet, increase of exercise, and use of medication when appropriate, can have an impact on the outcome of patients with diabetes, especially if these adjustments in lifestyle are implemented early, in the onset of the disease.

In this article, we use machine learning to develop a computational model to diagnose diabetes. The computational model is built using a data set that we obtained from the website Kaggle [5]. This is a website that has a large collection of data sets, available to the public, that can be used to develop machine learning models.

This article is organized as follows. We first explain what supervised learning is. This is a class of problems within the larger class of problems of machine learning. Our diabetes example belongs to this category of supervised learning. We explain the structure of the data sets in supervised learning problems, and explain the concept of examples, features and labels. We explain these concepts in general, as well as in our diabetes data set. Next, we explain what logistic regression models are. This is the class of models we use to diagnose diabetes. We explain the notion of parameters, training set, error on a set of examples, and how the parameters are selected by minimizing the error on the training set. We finish this article illustrating the concepts explained by developing a model to diagnose diabetes and we discuss the accuracy of the model on a set of examples that are not part of the training set. This set is called the validation set. We finish the article with a small discussion.

Methods

Supervised learning and the data set

The specific problem we address is the diagnosis of diabetes. Our data set consists of information about several patients. The information about each patient is: if the patient has high blood pressure or not; if the patient has high cholesterol or not; the body mass index of the patient; if the patient is a smoker or not; if the patient has heart disease or not; if the patient is physically active or not; if the patient eats fruits or not; if the patient eats vegetables or not; if the patient consumes a large amount of alcohol or not; the general health of the patient (a number between 1 and 5); the sex of the patient; the age of the patient; as well as whether the patient has diabetes or not. Part of this data set is illustrated in Table 1. The entries in the first row are abbreviations of the information in that column. They have the following meaning:

1. The entry in the column HBP has a 1 if the patient has high blood pressure and a 0 if the patient does not have high blood pressure.
2. The entry in the column HCH has a 1 if the patient has high cholesterol and a 0 if the patient does not have high cholesterol.
3. The entry in the column BMI has the body mass index of the patient.
4. The entry in the column Smo has a 1 if the patient smokes and a 0 otherwise.
5. The entry in the column HD has a 1 if the patient has heart disease and a 0 otherwise.
6. The entry in the column PA has a 1 if the patient is physically active and a 0 otherwise.
7. The entry in the column EF has a 1 if the patient eats at least one fruit per day and has a 0 otherwise.
8. The entry in the column EV has a 1 if the patient eats at least one serving of vegetables per day and has a 0 otherwise.
9. The entry in the column Alc has a 1 if the patient consumes at least 14 glasses of alcohol per week and a 0 otherwise.
10. The entry in the column Hea rates the general health of the patient from 1 = poor to 5 = excellent.
11. The entry in the column Sex has a 1 if the patient is male and a 0 is the patient is a female.
12. The entry in the column Age has the age transformed to a number between 1 and 13, where 1 means the age of the patient is between 18 and 24, 9 means the age of the patient is between 60 and 64 and 13 means the patient is 80 years old or older.
13. The entry in the column Dia is 1 if the patient has diabetes or 0 if the patient does not have diabetes.

In Table 1 we show the information about only two patients, but our dataset contains information about 70692 patients.

HBP	HCH	BMI	Smo	HD	PA	EF	EV	Alc	Hea	Sex	Age	Dia
1	0	26	0	0	1	0	1	0	3	1	4	0
1	1	25	0	1	1	1	0	0	2	0	9	1

Table 1. Data of two of the examples in our data set.

The problem we consider in this article belongs to the class of problems known as supervised learning. A first characteristic of this class of problems is that the data set consists of information about a collection of units. In our data set, the units are the patients. In the language of machine learning, the units are called examples. Thus, the examples are the patients in our data set.

A second characteristic about supervised learning problems is that the information the data set contains about each example is of two types: the label or target variable, and the features. The label is what we eventually want to predict for examples that are not in our data set. In the data set we consider; this information is whether the patient has diabetes or not. The rest of the information about each example are called features. Thus, in our data set, the features are the information stored in the columns HBP, HCH, BMI, Smo, HD, PA, EF, EV, Alc, Hea, Sex and Age.

The objective of the rest of this article is to use the data set of the patients to develop a computational model that can predict if a new patient, not in the data set we use to develop the model, has diabetes. To make its prediction, we need to provide the model with the features of the patient. In the rest of the article we will explain the theory behind the development of the model as well as the results we obtain.

Binary Classification Problem

Each patient either has diabetes or not. In other words, the label takes one of two values: 1 if the patient has diabetes and 0 otherwise. Problems where the label takes one of two possible values are known as binary classification problems. We say that each example belongs to one of two categories, according to the value of its label. One of the categories is identified with the number 0 and the other with the number 1. We call the categories category 0 and category 1, respectively. In our case, 1 means the patient has diabetes and 0 means the patient does not have diabetes.

A model for binary classification problems is a function that takes as input the features of an example and gives as output a number between 0 and 1. As is the common practice, we denoted this number by $\hat{y}$. As it will be explained soon, $\hat{y}$ is a prediction of the label of the example. Note that $\hat{y}$ is a function of the features of the example. In our case, each example has 12 features. We denote these features by $x_1, x_2, \ldots, x_{12}$ and the meaning of the features are as in the columns of Table 1. Thus, given a patient, $x_1 = 1$ if the patient has high blood pressure, but $x_1 = 0$ if the patient does not have high blood pressure. Similarly, $x_2 = 1$ if the patient has high cholesterol, but $x_2 = 0$ if the patient does not have high cholesterol. The meaning of the other features, $x_3, \ldots, x_{12}$, is explained similarly from Table 1. Since $\hat{y}$ is a function of the features, we write $\hat{y} = \hat{y}(x_1, x_2, \ldots, x_{12})$. The prediction of the model is that the example with features $x_1, x_2, \ldots, x_{12}$ belongs to the category 1 if $\hat{y}(x_1, x_2, \ldots, x_{12}) > 0.5$ or to the category 0 if $\hat{y}(x_1, x_2, \ldots, x_{12}) < 0.5$.

We have not explained how the function $\hat{y}(x_1, x_2, \ldots, x_{12})$ is selected. We will do so in subsequent sections. For now, consider the following example. Assume that a patient has the following features:

1. $x_1 = 1$ (the patient has high blood pressure)
2. $x_2 = 0$ (the patient does not have high cholesterol)
3. $x_3 = 24$ (the patient has a body mass index of 24)
4. $x_4 = 0$ (the patient does not smoke)
5. $x_5 = 0$ (the patient does not have heart disease)
6. $x_6 = 1$ (the patient is physically active)
7. $x_7 = 0$ (the patient does not eat a fruit per day)
8. $x_8 = 1$ (the patient eats at least one serving of vegetable per day)
9. $x_9 = 0$ (the patient does not consume 14 glasses of alcohol per week)
10. $x_{10} = 5$ (the patient is in excellent health)
11. $x_{11} = 0$ (the patient is female)
12. $x_{12} = 13$ (the patient is older than 80 years old)

Assume that when feed to the model these features, the output is 0.2, i.e. $\hat{y}(1,0,24,0,0,1,0,1,0,5,13) = 0.2$. This means that the model predicts that the patient does not have diabetes. In the next sections, we describe how the function $\hat{y}(x_1, x_2, \ldots, x_{12})$ is constructed.

Logistic regression

Logistic regression is a machine learning technique that is used to develop models in binary classification problems. This is the technique that we use in this article and we explain in this section. We first need to explain what the sigmoid function is.

The sigmoid function is the function

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

The graph of the sigmoid function is displayed in Figure 1.

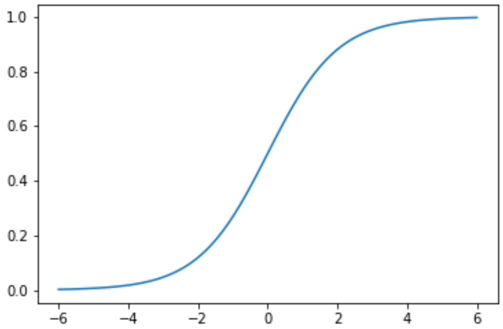

Figure 1. Plot of the graph of the sigmoid function.

The important properties of the sigmoid function are:

1. $0 < \sigma(x) < 1$ for all x.
2. $\sigma(x)$ is an increasing function of x.
3. $\sigma(x)$ becomes arbitrarily close to 0 as x becomes large in absolute value but negative.
4. $\sigma(x)$ becomes arbitrarily close to 1 as x increases.
5. $\sigma(0) = 0.5$.

In the rest of this article, we assume that each example has 12 features, even when we talk in general terms, not just referring to our diabetes problem. Logistic regression is a machine learning technique that assumes the prediction of the label to have the functional form

$$\hat{y} = \hat{y}(x_1, x_2, \dots, x_{12}) = \sigma(w_1x_1 + w_2x_2 + \dots + w_{12}x_{12} + b),$$

where as before, $x_1, x_2, \dots, x_{12}$ are the features of the examples, but $w_1, w_2, \dots, w_{12}, b$ are some numbers known as parameters. Note that we have not explained how the parameters are to be selected. We will get to that in subsequent sections. For now, note that the model is determined by the parameters. If we change the parameters, the model changes and thus, the predictions made by the model.

As an example, assume that the parameters are: $w_1 = 0.2$, $w_2 = 0.1$, $w_3 = -0.01$, $w_4 = 0.1$, $w_5 = 0.02$, $w_6 = -0.1$, $w_7 = 0$, $w_8 = -0.2$, $w_9 = 0$, $w_{10} = -0.02$, $w_{11} = 0$, $w_{12} = 0$ and $b = 0$; and the features are $x_1 = 1$,

$x_2 = 0$, $x_3 = 24$, $x_4 = 0$, $x_5 = 0$, $x_6 = 1$, $x_7 = 0$, $x_8 = 1$, $x_9 = 0$, $x_{10} = 5$, $x_{11} = 0$ and $x_{12} = 13$. The model predicts

$$\hat{y} = \sigma(0.2(1) + 0.1(0) - 0.01(24) + 0.1(0) + 0.02(0) - 0.1(1) + 0(0) - 0.2(1) + 0(0)$$
$$-0.2(1) + 0(0) - 0.02(5) + 0(0) + 0(13) + 0) = 0.39$$

and thus, the model predicts that this patient does not have diabetes. In the next section we explain how the parameters $w_1, w_2, \ldots, w_{12}, b$ are selected.

Binary cross entropy error

Assume that the features of an example are $x_1, x_2, \ldots, x_{12}$. Assume that we know the label of that example and this label is y. Note that y is either 1 or 0. On the other hand, our model predicts the label of this example to be $\hat{y}$. Note that $0 < \hat{y} < 1$. The binary cross entropy error on this example is defined to be

$$BCE(y, \hat{y}) = -(y \, log\, (\hat{y}) + (1 - y)\, log\, (1 - \hat{y}))$$

While we will not go into the details of the binary cross entropy error, we list here its properties that are most relevant to us:

1. $BCE(y, \hat{y}) \geq 0$.
2. If $\hat{y} = y$, then $BCE(y, \hat{y}) = 0$.
3. The closer $\hat{y}$ is to y, the smaller $BCE(y, \hat{y})$ is.

For the reasons listed above, $BCE(y, \hat{y})$ is a measure of the difference between y and $\hat{y}$. Thus, $BCE(y, \hat{y})$ can be considered as a measure of the error the model makes in predicting the label of the example. For example, assume that $y = 1$ and $\hat{y} = 0.7$, then

$BCE(y, \hat{y}) = BCE(1, 0.7) = -\, log\, (0.7) = 0.15.$

On the other hand, if $y = 1$ and $\hat{y} = 0.9$, then

$BCE(y, \hat{y}) = BCE(1, 0.9) = -\, log\, (0.9) = 0.05.$

We see that the better prediction of $\hat{y} = 0.9$ gave the smaller cross entropy error.

The mean binary cross entropy error on a set of examples, is the average of the binary cross entropy errors on the examples in the set. We illustrate this with the help of Table 2, where we display the labels y, the predicted labels $\hat{y}$ and the binary cross entropy errors $BCE(y, \hat{y})$ of three examples. We also show the average of those errors, which is the mean binary cross entropy error on this set of three examples.

y	$\hat{y}$	$BCE(y, \hat{y})$
1	0.9	0.05
0	0.2	0.1

0	0.1	0.05
Mean $BCE(y, \hat{y})$		0.67

Table 2. Binary cross entropy errors of three examples and the mean binary cross entropy error on the set of these three examples together.

Training and validation set

The examples on the data set given to us to develop the model are split into two sets: the set of training examples, or the training set, and the set of validation examples, or the validation set. As is common practice, our training set will contain 75% of the examples and thus, our validation set will contain 25% of the examples. This split is done randomly. In other words, given an example in our original data set, the probability that this example will belong to the training set after the split is 75%. Note that we have both the features and the labels of the examples in both the training and the validation set. The reason for this split is described in later sections.

Results

Selection of the parameters

Note that this binary cross entropy error on the training set depends not only on the values of the features and labels of the examples in the training set, but also on the parameters $w_1, w_2, \ldots, w_{12}, b$. If we change those parameters (keeping the training set the same), the binary cross entropy error also changes.

In logistic regression, the parameters that are selected are those that make the mean binary cross entropy error on the training set as small as possible. We will not go into any details on the algorithms used to find those parameters. In practice, these parameters are usually found using software libraries that are available to be used by the public at no cost.

To illustrate the above discussion, consider Table 3, where we show a training set with only six training examples. Each example has only one feature, so this Table is unrelated to the diabetes data set we consider in this paper, where each example has 12 features. In that table, MBCE means the mean binary cross entropy error. Note that, with the parameters $w = 1$ and $b = 0$, the mean binary cross entropy error is 4.56. On the other hand, with the parameters $w = 3.83$ and $b = -0.89$, the mean binary cross entropy error is 0.33. This means that the model with the parameters $w = 3.83$ and $b = -0.89$ is better than the model with the parameters $w = 1$ and $b = 0$. This is evident by looking at the column with the predictions $\hat{y}$ from each model. In fact, the parameters $w = 3.83$ and $b = -0.89$ gives the smallest mean binary cross entropy error, i.e. a model with other parameters gives a larger mean binary cross entropy error. Note also that we have not, and will not, explained how these optimal parameters, $w = 3.83$ and $b = -0.89$ are found. We only mention that we use the library Keras to find these optimal parameters.

x = feature	y = feature	$\hat{y}$ = predicted label with $w = 1$ and $b = 0$	$\hat{y}$ = predicted label with $w = 3.83$ and $b = -0.89$
-1	0	0.27	0.01
-0.8	0	0.31	0.02

0.2	1	0.55	0.47
0.4	0	0.60	0.66
0.8	1	0.69	0.90
1	1	0.73	0.95
MBCE		4.56	0.33

Table 3. Example that illustrates that the parameters that lead to the smallest possible mean cross entropy error leads to better predictions.

We now go back to our diabetes data set. We used the corresponding training set to find the optimal parameters, i.e. the parameters that minimize the mean binary cross entropy error on the training set. We find the optimal parameters to be the ones that we list in Table 4

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	b
0.76	0.60	0.08	0.02	0.31	-0.06	-0.06	-0.12	-0.81	0.60	0.20	0.17	-6.18

Table 4. Optimal parameters

Validation set and accuracy

The validation set is used to evaluate how good the model is. The validation set was not used in the development of the model; thus, the validation set gives an accurate prediction of how well the model will work on new examples, these are examples where the label is not known. The accuracy of the model is defined as the number of correct predictions of the model on the examples in the validation set, divided by the number of examples in the validation set. This number gives the expected percentage of times that our model will give the right prediction. We obtained:

Accuracy on the validation set $=$ 0.74.

Thus, our model is expected to give the correct diagnoses 74% of the time.

Discussion

In this article we gave an overview of supervised learning and logistic regression. We applied these concepts and techniques to a data set of patients and we developed a model to diagnose diabetes. We find that our model is %74, which is certainly much better than just guessing, which would be %50, but not satisfactory enough for the model to be the sole diagnosis tool. To improve the accuracy of the model we propose to explore more complex types of models, such as neural networks, and/or include more features. These research directions will be pursued in the future and reported in a future article.

Conclusion

In conclusion, we can say that machine learning techniques can be used to diagnose or predict the risk of patients of developing diseases such as diabetes. This promises to be a very valuable tool that can help in the prevention of diseases and in lowering the cost of health care.

Limitations

While promising, the calculations in this article have their limitations. These limitations are likely to be due to the fact that we have used logistic regression instead of the more complex technique on neural networks.

Acknowledgements

I would like to thank my mentor, Guillermo H. Goldsztein, Ph. D., Professor, School of Mathematics, Georgia Institute of Technology, for his guidance with this project.

References

[1] Ethem Alpaydin. 2010. Introduction to Machine Learning (2nd ed.). MIT Press.

[2] Andreas C. Müller, Sarah Guido, 2016. Introduction to Machine Learning with Python. O'Reilly Media, Inc.

[3] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, 2018, Foundations of Machine Learning (2nd ed.). MIT Press.

[4] Kevin P. Murphy, 2012, Machine Learning A Probabilistic Perspective, MIT Press.

[5]https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv