Machine Learning and the Detection of Diabetes through AI

Author: Bill Liu

Author Bio: Bill Liu is currently a high school student at Manhasset high school. He is interested in math and sports. He is currently taking a calculus course and finds it extremely interesting. He also loves the sports of wrestling and MMA. During free times he likes to hangout with friends and play video games. However, he has a very limited amount of free time due to a heavy schedule in academic activities. As of right now, he is focusing on both school and sports with his goal of getting into his dream college.

Abstract

Diabetes is a serious illness which could result in serious consequences if not properly treated. One way to detect diabetes early is with the use of machine learning, in which a subject would provide his/her information and the machines will predict whether the subject has diabetes. In this study, four types of machine learning models were used to predict the result. Each model is unique in their own way and a comparison could reveal the advantages and disadvantages of each model.

Keywords

Machine learning, diabetes, disease prediction, modeling, category, neural network, blood sugar, glucose

Introduction

Diabetes is a disease in which the blood sugar level is too high. This disease can be fatal if not properly treated. On average 10.5% of people in the US have diabetes. Detection of diabetes requires visits to clinics or hospitals. Even if early detection was accomplished, treatment can still be expensive.

Since not many people are able to afford a proper exam, Machine learning would be vital since it can predict the outcome with the filled out information. Machine learning is the study of fine limited sequences. Algorithms can be used to predict an outcome. In this case we would take a list of examples like the person's age, glucose level, blood pressure, and so on. We would then use the information given to calculate the probability of diabetes. Machine learning will not always produce the exact right answers since there could be a small percentage error. However the error can be reduced by training the model with more data.

Machine learning could be classified into 3 different categories. Which include supervised learning, which is predicted from the training data, unsupervised learning in which the system predicts the unlabeled data, and finally reinforcement learning is in which the system predicts from dynamic data.

Supervised learning:targets function have to be learned by the system function. This is basically data described by the model. Dependent and output variables are also used to predict the values and variables. The sets of possible functions in its domain are called the instances. In supervised learning there are two kinds of learning called classification and regression. Classification predicts distinct classes such blood groups, while the regression models predict actual number values.

Unsupervised learning: the system discovers the similarities between the data and variables. As a result, training data consists of instances without any corresponding labels.

Reinforcement learning: This learning is when the system adapts to the environment so that it maximizes some notion of cumulative reward. It is very important to note that the system had no interaction with the knowledge about the environment and the only way to determine the outcome is through trial. Reinforcement learning is often related to autonomous systems.

Methods

Data collection and preparation

We obtained data from (https://www.collaborat.com/pima-diabetes-data-discovery-predictive-model/).

The data set consists of information of 768 female patients. The information about the patients is: number of pregnancies, glucose levels. blood pressure, skin thickness, insulin level, diabetes pedigree function, age, as well as whether they have diabetes or not. This data set is illustrated in Table 1, where we only show the information about four patients. The dataset was split into a training set and a validation set.

In our case, the features are the number of pregnancies, glucose levels. blood pressure, skin thickness, insulin level, diabetes pedigree function, age, and the label is whether the patient has diabetes or not. The label is in the column Outcome. It is 1 if it has diabetes and 0 if it does not.

Table 1 - **Diabetes sample (including the missing data)**

Pregnancies	Glucose	Blood Pressure	Skin thickness	Insulin	BMI	Diabetes pedigree	age	Outcome
6	148	72	35	NaN	33.6	0.627	50	1
1	85	66	29	NaN	26.6	0.351	31	0
8	183	64	NaN	NaN	23.3	0.672	32	1
1	89	66	94	96	28.1	0.167	21	0

https://www.collaborat.com/pima-diabetes-data-discovery-predictive-model/

Note that some of the entries on Table 1 read NaN. This means that the corresponding data is missing. We removed the examples with missing data (Table 2), resulting in a total of 392 samples

Table 2 Diabetes sample (Excluding missing samples)

Pregnancies	Glucose	Blood Pressure	Skin thickness	Insulin	BMI	Diabetes pedigree	age	Outcome
1	89	66	94	96	28.1	0.167	21	0
3	78	50	32	88	31	0.248	26	1
2	197	70	45	543	30.5	0.158	53	1
1	189	60	23	846	30.1	0.398	59	1

Model training

Neural network was selected as the desired model. We used Python and its popular libraries, like numpy to train the logistic regression model.

We denote by k the number of features of each example. In our problem, $k = 8$. Logistic regression is a machine learning technique that is used in problems as the one described in this article. To explain the model, we first need to introduce the sigmoid function (Equation 1).

$$\sigma(x) = 1/(1 + e^{-x}) \quad (1)$$

The model assumes that the prediction $\hat{y}_i$is of the form in equation 2.

$$\hat{y}_i = \sigma(w_1 x_1 + w_2 x_2 + \ldots + w_k x_k + b) \quad (2)$$

The numbers $w_1, w_2, ..., w_k, b$ are called parameters. The parameters that are selected for the model are those that make the binary cross entropy error on the training set as small as possible.

The trained model takes an input of given features (e.g. $x = [x_1, x_2, ..., x_k]$) , and makes a predictions $\hat{y} = \hat{y}(x)$. This number $\hat{y}$ is a number between 0 and 1, which represents the probability of the example belonging to category 1. In our case, 1 indicates that the patient has diabetes. For example, a patient with $x = [1{,}70,\ 90, 96, 27, 0.2, 30]$ means that, this patient, the number of pregnancies was 1, glucose levels was 70, blood pressure was 90, skin thickness was 27, insulin level the diabetes pedigree function was 0.2 and her age was 30. If $\hat{y}([1{,}70,\ 90, 96, 27, 0.2, 30]) = 0.7$ This means there is a 70 percent chance the patient will have diabetes.

After training the model, the validation set was used to measure the performance of the model. Accuracy and precision were used as evaluation metrics. Besides that, there are also several other measures on the performance of the model. Assume that we have n examples and let x_i be the (one dimensional numpy array with the) features of the i^{th}example. Let $y_i$be the label of that example and $\hat{y}_i$be the prediction from our model. The binary cross entropy error on this example is $-(y_i log(\hat{y}_i) + (1 - y_i)log(1 - \hat{y}_i))$. This quantity is a measure of the error the model makes because it is always non-negative. The closer $\hat{y}_i$ is to y_i , the smaller the error is, and the error is 0 if $\hat{y}_i = y_i$.

On the whole set of examples, the mean binary cross entropy error is the average of the binary cross entropy error on the examples in the set , i.e.

Binary cross entropy error = $-\frac{1}{n}\sum_{i=1}^{n} (y_i log(\hat{y}_i) + (1 - y_i)log(1 - \hat{y}_i))$,

where the $\sum_{i=1}^{n}$ means adding over all i between 1 and n.

In this study, we have trained four models with different parameter settings. Model 1 is a sequential network,where the activation function is sigmoid and the loss function is binary entropy. Model 1 fitted the data with 400 epcohes. Model 2 is a sequential network as well as another layer of relu activation and the model consists of 800 epochs.Model 3 has a hidden layer with 2 nodes and another layer of relu activation with the function being sigmoid. Model 4 is the same as model 2 except that L2 regularization penalty was applied.

Results

Model 1 has achieved a weighted average precision of 0.79, recall of 0.80, and f1-score of 0.79 on the training set. For the testing set, the precision, recall, f1-score are 0.75, 0.76, and 0.75, respectively. It was observed that precision, recall, and f1-scores for non-diabetes are consistently higher than that for diabetes patients, indicating a higher prediction power for non-diabetes than diabetes. This trend is consistent in the training set and testing set.

Model 2 has achieved a weight average precision of 0.82, recall of 0.82, and a f-1 score of 0.82 on the training set. Now the precision, recall, f1-score are 0.7,0.7,0.7 for the validation set, respectively. It was observed the precision, recall, and f1-score for non diabetes are consistently higher than their counterparts, the diabetes patients. The trend is consistent in both the training set and testing set.

Model 3 has achieved a weight average precision of 0.79, recall of 0.8, and a f-1 score of 0.8 on the training set. Now the precision, recall, f1-score are 0.76, 0.77, 0.76 for the validation set, respectively. It was observed the precision recall, and f1-score for non diabetes are consistently higher than their counterparts the diabetes patients. Both sets are consistent in the trend.

Model 4 has achieved a weighted average precision of 0.81 , recall of 0.82, and f1-score of 0.81 on the training set. For the testing set, the precision, recall, f1-score are 0.70, 0.71, and 0.70, respectively. It was observed that precision, recall, and f1-scores for non-diabetes are consistently higher than that for diabetes patients, indicating a higher prediction power for non-diabetes than diabetes. This trend is consistent in the training set.

Overall model 1 has achieved 79% precision on the training set and 75% on the testing set. In model 2 the model has achieved 82 % on the training set however only 70% on the testing set. Model 3 the model achieved 79% one training set and 76% in the testing and at last model 4 has achieved a 81% on the set and a 70% on the testing set.

Accuracy and the precision of the data depend on the amount of numbers in the data and the amount of samples taken. The lower the sample taken the lower the accuracy will be and consequently the higher the sample the higher the accuracy will be.

Discussion

This study compared several machine learning methods in predicting diabetes. Each model showed different prediction results. Results indicated that the training and testing accuracy on model 1 was close, indicating that the model has not suffered from overfitting or underfitting. This is very similar to the results of model 3. In comparison, model 2 and model 4 showed a 12% and 11% difference between training and testing sets, indicating a potential overfit. This can be explained by the nature of the models. Since model 2 uses another layer of relu activation and uses 800 epochs and model 4 uses L2 regularization, it is possible that these features contributed to the overfit.

Compared to other models in the literature, our model is similar to most studies with around 80% precision. It seems to be unlikely that the number can be further increased without introducing more data. Given the nature of the dataset where only eight predictors are present and the large amount of missing data, this dataset is inherently limited. First, the data does not show any pre-existing diseases, which is extremely important in predicting any disease. Second, the key information, insulin, was missing in many cases, which further decreased the applicability of this dataset.

While various studies including the current study attempted to use neural networks to fit the data, this approach lacks the power of explanation. It is hard to understand which is the most important factor in determining diabetes. Previous studies have indicated that glucose is the most important factor. Future studies should focus on applying more explainable methods in model development, e.g. logistic regression.

Future studies should also focus on applying neural networks in more complicated solutions. The fundamental goal of this study is to build an algorithm that could help with fast diabetes detection. The end implementation of this algorithm would most likely be in the form of a smartphone application. Therefore, it is beneficial to include image data in the model training process.

In conclusion, this study demonstrated that machine learning can be used to predict diabetes given ample training data. We made a systematic effort to use Machine learning and applied it to a real world problem. Diabetes is a problem faced by many people and to this date we still haven’t been able to 100 percent cure the diseases therefore it is vital to detect this disease as early as possible. With many families dealing with financial support, machine learning will be their best hope of a cheap and reliable result. This is why Machine learning is a powerful and useful tool for a variety of problems. This will eventually lead to the fast detection of diabetes and decrease the chance of complications due to delayed diagnosis.

These values are different since the amount of samples taken are different: the more samples taken the more accurate the result will be however there will always be an error we can only reduce the error. The future of the research might be to increase the percentage even higher. My scientific conclusion is that model 4 achieved the highest percentage on the training set and model 1 achieved the highest on the testing set.

References

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017, January 8). *Machine Learning and data mining methods in diabetes research*. Computational and Structural Biotechnology Journal. Retrieved January 4, 2022, from https://www.sciencedirect.com/science/article/pii/S2001037016300733

Singla, R., Singla, A., Gupta, Y., & Kalra, S. (2019). *Artificial Intelligence/machine learning in diabetes care*. Indian journal of endocrinology and metabolism. Retrieved January 4, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/