CLASSIFICATION OF DIABETIC PATIENTS USING TWO MACHINE LEARNING APPROACHES

BY

DABERECHI CORNELIUS AHANONU

(UNIVERSITY OF SALFORD)

Introduction

Diabetes is a condition that affects both the young and the elderly and is characterised by excessive blood sugar levels (Chentli, Azzong, and Mahgoun, 2015). It can result in a range of complicated disorders, including stroke, renal failure, heart attack, etc. In 2014, over 422 million individuals were diagnosed with diabetes globally. In 2040, the number will reach 642 million. Based on these alarming figures, it is necessary to build a prediction system for early illness identification. Therefore, a technique based on machine learning has been presented for the categorization, early detection, and prediction of diabetes (Maniruzzaman, Rahman, and Ahammed, 2020). For this work, logistic regression and the support vector machine were employed as machine learning techniques. Other portions of this report will detail the remaining tasks.

Datasets

This data set was gathered from the India Kaggle (2020) competition hosted by the National Institute of Diabetes and Digestive and Kidney Diseases. The goal here is to use diagnostic parameters as a predictor for diabetes. The dataset will be trained using supervised learning since it has labels that can be used to directly map input variables to the desired output.

Explanation and preparation of datasets

The variables for the dataset include:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)

- BMI: Body mass index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)

The above variables are the independent variables.

- Outcome: Class variable (0 or 1)

While the outcome is the dependent variable.

Given the range of possible values for the independent variables, we found it necessary to normalise the collected information. The purpose of normalisation is to make all of the characteristics uniform in size (Google 2020). This enhances the model's efficiency and reliability throughout training. We used z-score normalisation to achieve our goals here. This is given by

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

We checked for missing values and there was none. In terms of class balance, we observed that the class is not highly imbalanced as it might not have much influence on the performance of our learning algorithms.

Implementation in Python / Azure Machine Learning Studio

Description of the algorithms used

Logistic regression is a sort of supervised learning that assigns a label to a set of input variables (Simplilearn, 2022). The dependent variables for linear regression are continuous variables, whereas the dependent variables for logistic regression are dichotomous variables, or binary variables. Similar to conventional linear regression, logistic regression provides a non-linear function that captures complicated patterns in the data set. For instance, if we need to determine if a patient is sick or not, logistic regression is the superior method. The sigmoid function is a nonlinear function, often known as the activation function. The sigmoid function is responsible for compressing the entire real number input into values between 0 and 1. (Lazara, 2020).

The sigmoid function is given by:

$$s(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

where x can have range $[-\infty, +\infty]$, while s has range [0,1].

When we are on a decision boundary, the probability is equal for both outcomes. When

we move away from the decision boundary; we have certain outcome more likely.

$$\Pr(y|x) = \frac{1}{1+e^{-y(wx+b)}} \qquad (3)$$

Sigmoid function is linear near 0 and has sharp slopes towards the ends. It squashes the outliers towards 0 or 1. The data is fitted with the linear regression model, then a sigmoid function is used to predict the categorical target.

1. Support Vector Machine (SVM)

SVM is another classification technique employed in machine learning (Kumari and Chitra, 2013). The SVM determines the decision boundary that maximises the margin in order to correctly classify an unobserved data point. The optimal decision boundary is referred to as a "hyperplane." SVM is a two-classifier algorithm. Suppose we have N data points, each with attributes $x = [x_1, x_2]^T$ and target t=± 1, then a linear decision boundary can be represented as a straight line:

$$w^T x + b = 0$$

Our task is to find w and b. Once we have the values of w and b, classification is easy.

$$w^T x_{new} + b > 0 : t_{new} = 1$$

$$w^T x_{new} + b < 0 : t_{new} = -1$$

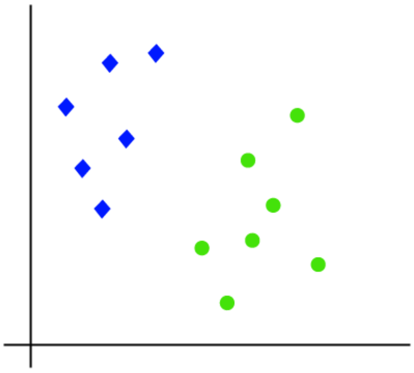
That is $t_{new} = sign(w^T x_{new} + b)$

In the SVM space, we can look at a situation where the dataset is linearly separable or not.

Types of SVM

1. **Linear SVM**

Consider a dataset with two characteristics, x1 and x2, that can be identified by two tags (green and blue). Consider the below image. We need a classifier capable of linearly classifying the pair of coordinates (x1, x2) as either green or blue.

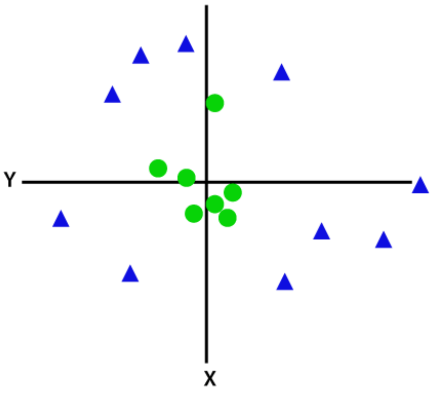


In a two-dimensional space, by just using a straight line, we can easily separate these two classes.

Consequently, the SVM method assists in locating the optimal line or decision boundary; this optimal border or region is referred to as a hyperplane. Therefore, the SVM algorithm identifies the nearest point between two classes' lines. These are known as support vectors. Margin refers to the distance between the vectors and the hyperplane. Moreover, the objective of SVM is to increase this margin.

2. Non-Linear SVM

Sometimes data may appear in a ring form and hence not linearly separable, hence, to solve this, we project the data onto a high dimensional space. This is shown in the diagram below:



For linear data, we have utilised dimensions x and y, but for non-linear data, dimension z will be added. It can be determined by:

$$z = x^2 + y^2 \qquad (4)$$

The implementations of the logistic regression and support vector machine in Azure Machine learning studio can be found here: logistic regression and Support Vector Machine

The application of data-mining techniques to selected datasets that you choose using Python.

Classification is the data mining approach employed in this instance. It categorises data set elements or variables into preset categories or types (Sharma, 2022). Data mining utilises linear programming, statistics, decision trees, and artificial neural networks, among other approaches. Classification is used to construct software that can be modelled to categorise elements inside a data set into distinct classes.

For instance, we might use it to divide all of the applicants who went to an interview into two distinct groups: the first group would be a list of the candidates who were chosen for the position, and the second group would be a list of the candidates who were not chosen for the position. This categorization work may be done with the help of tools designed for data mining.

Explanation of the experimental procedure, including the setting and optimisation of model hyperparameters during training, and your approach to validation (for supervised learning tasks).

We used the hold-out strategy. We separated the dataset into training and testing halves. In accordance with the criterion that the proportion of training data must be greater than that of test data, we allocated 80% of the dataset for training and the remaining 20% for testing. Regarding the hyperparameters, we studied the C parameter, which controls error in SVM, and the kernel, which determines whether or not the datapoints are linearly separable. We also used a grid search to get the ideal hyperparameter values. Later, we also implemented a cross-validation strategy for partitioning the dataset such that every portion of the dataset was represented throughout the training procedure.

Visualisation of the results.

We visualized the class labels to have a feel of how imbalanced the dataset is. The ROC_AUC curve was plotted to check the overall performance of the two models. The ROC_AUC curve is a plot of the true positive rate against the false positive rate at different thresholds.

Results analysis and discussion

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Logistic Regression	0.75	0.75	0.76	0.75	0.81
Support Vector Machine	0.75	0.74	0.75	0.74	0.82

The accuracy for both logistic regression and support vector machine are the same with value of 75% but the accuracy is always hampered by imbalanced dataset. However, the dataset is not highly imbalanced as to have much influence on the dataset. The dataset was up-sampled and down-sampled, but their accuracies did not improve significantly. We also compared the two models in terms of being able to correctly predict the positive cases, we examined the precision and recall measures. Here, there's not significant difference with regards to the logistic regression and support vector machine. We also looked at the overall model performance at different thresholds using the AUC value, with the support vector machine performing a little bit better than the logistic regression. For the SVM, we carried out a grid-search to determine the optimal value of the SVM parameters, we result showed that the best value for C is 10 with a linear kernel.

Conclusions

In the work, we have successfully carried out a task on classification of diabetic patients using two machine learning approaches. Both learning algorithms achieved a 75% accuracy, but the SVM performed a little bit better than the logistic regression in terms of predicting the labels at each probability threshold in terms of the AUC value.

References

Chentli F, Azzoug S, Mahgoun S., 2015. Diabetes mellitus in elderly. *Indian J Endocrinol Metab.* Nov-Dec;19(6):744-52. Available at: doi: 10.4103/2230-8210.167553.

Dorian Lazara., 2020. Understanding Logistic Regression. Available at: https://towardsdatascience.com. (Accessed: 12 November, 2022).

Google., 2020. Normalization. Available at: https://developers.google.com/machine-learning/data-prep/transform/normalization. (Accessed: 12 November 2022).

Kaggle., 2020. Diabetic Dataset. Available at: www.kaggle.com. (Accessed: 12 November 2022).

Kumari, V.A. and Chitra, R., 2013. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2), pp.1797-1801.

Maniruzzaman, M., Rahman, M.J., Ahammed, B. *et al.*, 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst* **8**, 7. Available at: https://doi.org/10.1007/s13755-019-0095-z.

Pranabdas., 2022. Logistic Regression. Available at: Pranabdas.github.io/machine-learning. (Accessed: 12 November, 2022).

Rohit Sharma (2022). Data Mining Techniques: Types of Data, Methods, Applications. Available at: https://www.upgrad.com/blog/data-mining-techniques/. (Accessed: 12 November, 2022).

Simplilearn., 2022. An Introduction to Logistic Regression in Python. Available at: www.simplelearn.com. (Accessed: 12 November, 2022).