Predicting Diabetes Using Perceptron: A Machine Learning Approach

Saurab Kharel - a1919868

October 5, 2024

Abstract

Diabetes is a disease that occurs when the body cannot properly regulate blood sugar levels. Predicting diabetes is crucial in the medical field. In this article, a single-layer perceptron has been used to process patient data and predict the likelihood of developing diabetes. The analysis utilizes data from Pima Indian women, focusing on key metrics like accuracy, precision, recall, and F1 score.

1 Introduction

Diabetes mellitus is a metabolic disorder characterized by elevated blood sugar levels. It occurs when the pancreas does not produce sufficient insulin or when the body's cells do not respond effectively to insulin. It is typically classified into different types, including Type-1 and Type-2 diabetes. Early diagnosis and treatment are crucial for managing the disease and preventing long-term complicationsSahu, M. Kumar, and R. Kumar 2021.

The single-layer Perceptron algorithm, introduced by Frank Rosenblatt, works as a linear classifier that computes a weighted sum of input features and generates a binary output**rosenblatt1958perceptron**. In this study, we aim to apply the Perceptron algorithm to a binary classification problem, predicting whether patients have diabetes based on a variety of medical features.

2 Dataset

The dataset used is provided by the National Institute of Diabetes and Digestive and Kidney Diseases and includes data collected from Pima Indian women. The dataset features multiple predictor variables, including:

- Number of pregnancies
- Glucose level
- Blood pressure
- Skin thickness
- Insulin level
- BMI (Body Mass Index)
- Diabetes pedigree function
- Age
- Outcome (0 for non-diabetic, 1 for diabetic)

The outcome variable is binary i.e 0 or 1, indicating whether or not a patient has been diagnosed with diabetes.

3 Methodology

3.1 Perceptron Model

The Perceptron is a linear classifier that can be expressed mathematically as:

$$f(x) = \text{sign}(w^T \cdot x + b)$$

Where:

- w represents the weights assigned to the input features.
- x is the input feature vector.
- b is the bias term.

The sign function outputs either 1 (indicating the presence of diabetes) or 0 (indicating the absence of diabetes). The training process involves iterative updating the weights w and the bias b to minimize classification errors**hastie2009elements**.

3.2 Training Process

The dataset is divided into training and test sets using the `train_test_split` function from the `scikit-learn` library, with 20% of the data reserved for testing. The model is trained using a learning rate (η) of 0.20 and a maximum of 200 iterations and stratify=b ensures that the proportion of samples for each class is roughly the same in both train and test set along with the random state set is 3 for reproducible. The following steps outline the Perceptron learning algorithm:

1. Initialize the weights w and bias b to small random values.
2. For each misclassified data point (x_i, y_i), update the weights and bias:

$$w = w + \eta y_i x_i$$

$$b = b + \eta y_i$$

3. Repeat the process until convergence or until the maximum number of iterations is reached.

4 Evaluation Metrics

The performance of the Perceptron is evaluated using the following metrics:

- **Accuracy**: The proportion of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: The proportion of true positive predictions out of all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: The proportion of actual positives that are correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score**: The harmonic mean of precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5 Results

The performance of the Perceptron model on the test dataset is summarized in the following classification report:

- **Accuracy:** 0.64
- **Precision:** 0.46
- **Recall:** 0.24
- **F1 Score:** 0.32

Class	precision	recall	f1-score	support
0	0.67	0.85	0.75	100
1	0.46	0.24	0.32	54
Accuracy			0.64	154
macro avg	0.57	0.55	0.53	154
weighted avg	0.6	0.64	0.6	154

Table 1: Classification Report for Perceptron Model

The values in the classification report represent:

- Class 0: Non-diabetic patients
- Class 1: Diabetic patients
- Precision: The proportion of true positive predictions out of all positive predictions for each class.
- Recall: The proportion of actual positives that are correctly predicted for each class.
- F1-score: The harmonic mean of precision and recall for each class.
- Support: The number of instances in each class.

These results suggest that the Perceptron is reasonably accurate in predicting diabetes but has limitations in recall, indicating that it misses some positive cases of diabetes.

6 Discussion

The accuracy of 0.64 indicates that the model correctly classifies 64% of the instances overall. However, the low precision (0.46), recall (0.24), and F1-score (0.32) for Class 1 suggest that the model struggles to correctly identify diabetic patients. In contrast, the model performs better for Class 0, with higher precision (0.67), recall (0.85), and F1-score (0.75). These results highlight the limitations of the single-layer Perceptron in capturing the complexity of the diabetes prediction task. The model tends to have a higher false negative rate, misclassifying diabetic patients as non-diabetic more frequently than the reverse.

7 Conclusion

The single-layer Perceptron is a simple yet powerful linear classifier that can be applied to the task of predicting diabetes. While it performs well in terms of accuracy and precision for non-diabetic cases, its recall score for diabetic cases suggests that it may not be the best model for detecting all cases of diabetes. Future work could explore the use of more sophisticated models and feature engineering to improve performance, particularly in identifying diabetic patients more accurately.

References

Auer, Peter, Harald Burgsteiner, and Wolfgang Maass (2008). "A learning rule for very simple universal approximators consisting of a single layer of perceptrons". In: *Neural Networks* 21.5, pp. 786–795.

Ding, Jingfei et al. (2019). "Deep learning for diagnosis of chronic kidney disease from postcontrast CT images". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1496–1502.

Mankar, Aditya (2023). *DiabetesPredictor-SLP*. `https://github.com/aditya-524/DiabetesPredictor-SLP`.

Raudys, Šarūnas (1998). "Evolution and generalization of a single neurone: I. Single-layer perceptron as seven statistical classifiers". In: *Neural Networks* 11.2, pp. 283–296.

Sahu, Sushree Sangeeta, Manoj Kumar, and Rajeev Kumar (2021). "Diabetes Prediction Using Machine Learning Techniques". In: *International Research Journal of Engineering and Technology (IRJET)* 8.2, pp. 1104–1108. URL: `https://www.irjet.net/archives/V8/i2/IRJET-V8I262.pdf`.

Smith, J.W. et al. (2020). *Pima Indians Diabetes Database*. `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`. Accessed: [Insert access date here].

Zhang, Yuxin et al. (2024). *A Deep Learning Approach to Diabetes Diagnosis*. arXiv: `2403.07483 [cs.LG]`.

Raudys Raudys 1998 discusses the evolution of a single neuron.

Deep learning has been applied to various medical diagnosis tasks Ding et al. 2019.

Auer et al. Auer, Burgsteiner, and Maass 2008 proposed a learning rule for simple universal approximators.

An implementation of a diabetes predictor using SLP is available on GitHub Mankar 2023.

Zhang et al. Zhang et al. 2024 proposed a deep learning approach for diabetes diagnosis.

Smith et al. 2020