Comparison of Machine Learning Models for Predicting Cancerous Breast Cancer Tumours

Lucas Samson

Department of Computing Science

University of Alberta

jlsamson@ualberta.ca

Introduction

The purpose of this experiment is to determine if a machine learning model can accurately predict the cancer diagnosis of a tumor in breast cancer patients. My study compares the accuracy of a logistic regression model, a softmax regression model, and a neural network model. This paper explores the strengths and weaknesses of each model with respect to the prediction of cancerous tumour diagnosis, as well as the design and specificities of each model.

Dataset

In order to accurately predict the cancer diagnosis of a tumour, a lot of data is required to train a machine learning algorithm. The “Breast Cancer Wisconsin (Diagnostic)” dataset, obtained by the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg contains 569 records, a sufficient number to train the models. The dataset contains 30 features in which the model will consider including tumour radius, concavity, and symmetry.

Problem Description

Each model in this study will take the same dataset as an input. The model will analyze every feature of each record to find trends and classify each tumour. The data will be split into a training set and a test set. The training set consists of 455 records, making up 80% of the dataset, and the test and validation sets consists of 57 records each, making up 10% of the dataset for validation and 10% for testing.

The dataset contains a feature labeled “Diagnosis”, represented by “M” for malignant (cancerous) or “B” for benign (non-cancerous). For the purpose of binary classification, these are converted to 1 and 0 for malignant and benign, respectfully.

Evaluation Metric

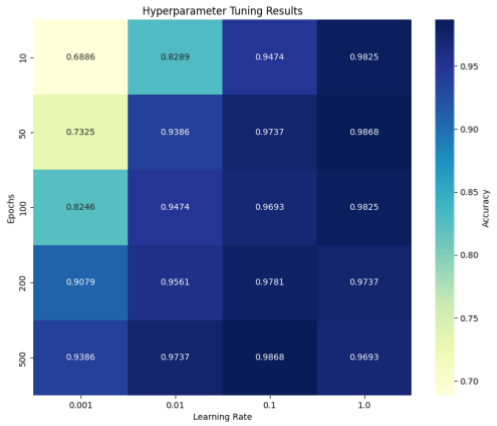
The measure of success for this experiment is the accuracy of each model’s predictions. In any medical diagnosis, accuracy is extremely important in order to prescribe the correct medication or procedures. An incorrect diagnosis can be extremely dangerous, if not fatal in some cases. For this reason, any machine learning model used for medical diagnosis must be extremely reliable.

Logistic Regression

The first method I chose was to implement a Logistic Regression model. Logistic Regression is a commonly used machine learning algorithm to solve binary classification problems because the outcome is a probability. Therefore, since the response variable is binary, Logistic Regression can compute the probability of a benign tumor reliably.

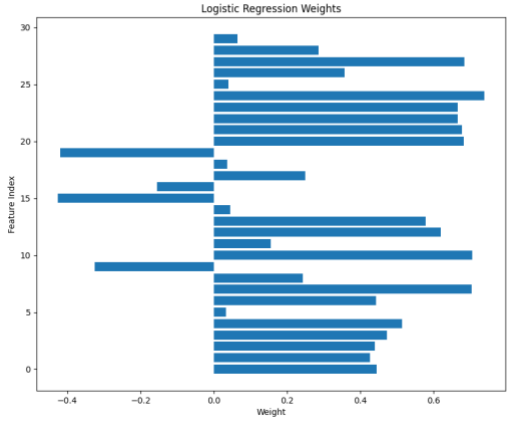
During model training, I had the algorithm repeat the training, looping through multiple sets of hyperparameters and storing the hyperparameters yielding the most accurate results. The hyperparameters used for training were the learning rate (0.001, 0.01, 0.1, or 1) and the number of epochs iterated through (10, 50, 100, 200, or 500).

Hyperparameter Tuning Results:



As shown above, the model achieves an accuracy greater than 95% when the learning rate is 0.01 or higher with at least 200 epochs. The model remains consistently accurate as the number of epochs grows and the learning rate grows, peaking at an accuracy of 98.68% at 500 epochs with a learning rate of 0.1, as well as at 50 epochs with a learning rate of 1.0.

From Logistic Regression, we can also see what features of the data are attributed the most to the diagnosis prediction.

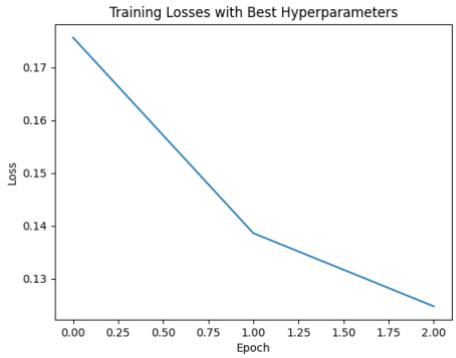


As shown above, we can see that the features at indexes 8, 11, and 25 are weighted the heaviest while features 10 and 16 have the least impact on the diagnosis. Cross-referencing the dataset, we see that the concavity, radius, and smoothness of the tumour have the greatest impression on the diagnosis while the fractal dimension and compactness are not good indicators as to whether or not a tumour is cancerous.

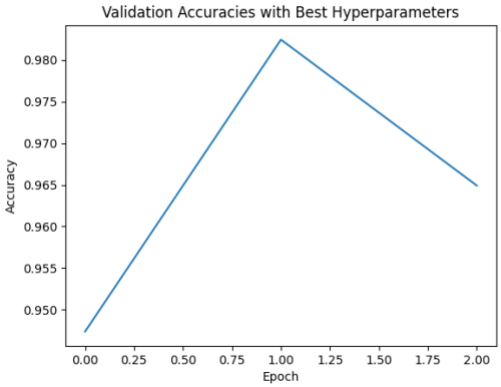
Softmax Regression

The second method I chose to try was using Softmax Regression. Softmax Regression is another commonly used algorithm for classification problems. This type of model is capable of multi-label classification, however since this is a binary classification problem we only have 2 labels. Therefore, rather than classifying the tumor as a probability of being cancerous, we find the probability of each class (cancerous vs non-cancerous). For this reason, softmax is usually applied to problems with more than 2 classes.

Similarly to Logistic Regression, I implemented an algorithm to loop through different combinations of hyperparameters in order to tune to the most accurate combination. The hyperparameters I tuned were the number of epochs to train on (10, 50, or 100), learning rate (0.001, 0.01, 0.1, or 1.0), batch sizes (32, 64, or 128), and decay (0.01, 0.1, 0.5, or 1.0).



The plot above shows the decreasing training loss through model training with the best hyperparameters.



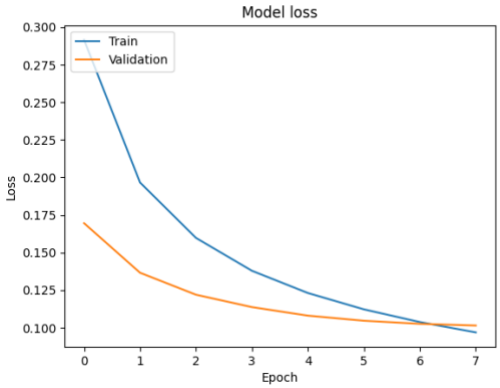
Shown above is the plot of the validation accuracy per epoch with the best hyperparameters. In order to combat issues of overfitting through training, I implemented a stopping condition that breaks the training loop once the accuracy stops improving.

Overfitting was a major problem through the implementation of this machine learning model. To counter this, I tried many strategies such as adding a stopping condition, using PCA from the sklearn library to trim the data to 10 features instead of the original 30, and increasing regularization. While I was able to improve it, the model remained continued to overfit to the training data but was still able to make a somewhat accurate prediction on the test data.

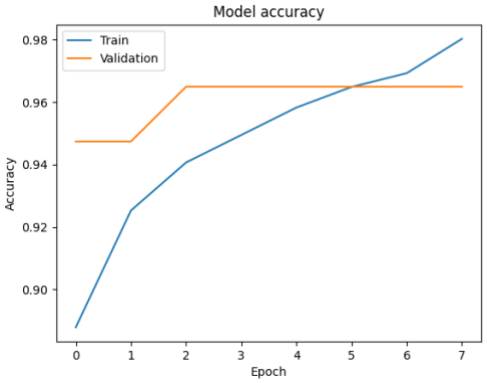
Neural Network

The last machine learning model I decided to compare is a Neural Network. I leveraged the Keras API from TensorFlow to build an algorithm to predict the cancer diagnosis. The complexity of the Neural Network requires a GPU to run so this test was completed in Google Colab.

The hyperparameters tuned in order to get the best accuracy were the learning rate (0.01, 0.1, or 1.0), number of epochs (10, 50, or 100), hidden layer activation function (tanh or ReLU), and number of hidden units (10, 50, or 100).



The plot above demonstrates the training and validation loss through each epoch with the best hyperparameters.



The plot above shows training and validation accuracies through each epoch with the best hyperparameters.

Findings

Model	Accuracy	Learning Rate	Number of Epochs	Batch Size	Decay	Hidden Layer Activation Function	Hidden Units
Logistic Regression	98.68%	0.1	500	-	-	-	-
Softmax Regression	98.25%	1.0	10	32	0.01	-	-
Neural Network	92.98%	1.0	10	-	-	ReLU	10

Conclusion

The most accurate model I tested to predict the breast cancer diagnosis of tumors was the Logistic Regression model with an accuracy of 98.68%. The Softmax Regression model delivered a similarly accurate prediction at 98.25%, and the Neural Network performed the poorest with 92.98%.

While a percentage around 99% is very impressive, it is insufficient for medical diagnosis. This means that 1 in every 100 patients can be misdiagnosed, resulting in two possibilities; either the cancer goes undetected and the patient does not get the treatment they desperately need, or the patient is falsely diagnosed with cancer. This second option could result in the patient receiving unnecessary dangerous cancer treatment such as chemotherapy that could make them worse off than before.

Therefore, while these machine learning models can predict the diagnosis of breast cancer with sufficient data to a very high probability, medical professionals cannot rely on them alone to provide a diagnosis. However, it is possible to use these models in conjunction with other methods to make the prediction or use them to support a diagnosis.

References

Dataset – https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Loading Dataset - https://www.kaggle.com/code/ashikm96/wdbc-primary-dataset-analysis

Sklearn Logistic Regression - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Heatmap - https://seaborn.pydata.org/generated/seaborn.heatmap.html

Neural Network Implementation - https://www.kaggle.com/code/karthik7395/binary-classification-using-neural-networks