Breast Cancer Diagnosis using Machine Learning

Muneera Cadersa, Dr Edward Kim

Drexel University, College of Computing & Informatics

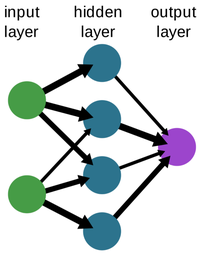
Introduction

Breast cancer is the prevalent cause of death, widespread among women globally. Several techniques exist to diagnose and treat cancer. Over the last decade, the use of IT and modern technologies have been employed to facilitate breast cancer detection and eventually deploy its cure. As we continue to evolve in an era backed by technology, machine learning techniques are being used extensively to provide early detection and prediction of this widespread disease.

Methods

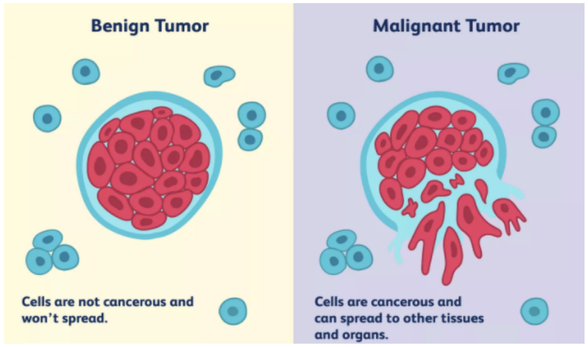
Machine Learning algorithm's aim is to develop a mathematical model that fits data. The algorithm is exposed to enough of these medical labeled data, which allows them to move into a model designed to then derive conclusions. **Logistic Regression** is used as an important predictor of breast cancer using odd ratios and generate confidence intervals, providing additional information for decision-making. **Neural network** is used to determine whether a tumor present in a woman's breast is benign or malignant.

A simple neural network



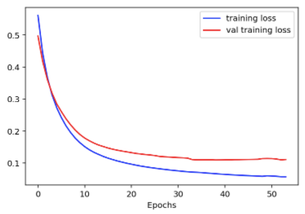
Data/Results

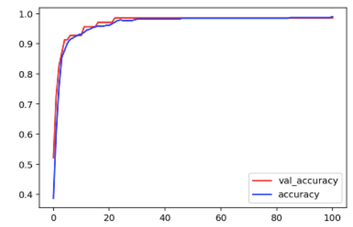
The dataset is imported from the UCI Machine learning repository. The output parameter which is the classification comes either as benign (B) or malignant (M).



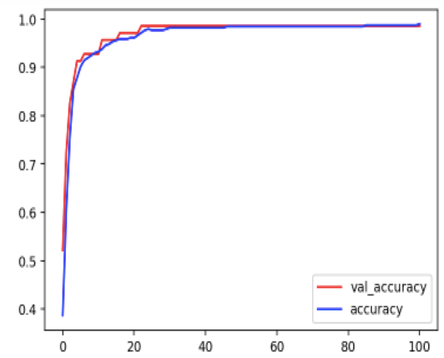
Input data are scaled. Inputs and outputs are split into training and testing data, with training data being 80% of the total data and testing data being 20 % of the testing data.

A shallow logistic regression model is developed. An input layer with 30 neurons, corresponding to the input features from the data set is added, connecting to a single hidden layer with an arbitrary amount of neurons. Each hidden layer neurons has the 'relu' activation function applied to it. The learning rate is set to 0.001 and the model also keeps track of the accuracy metric. The earlystopping function is defined. The performance and accuracy of the model are plotted.





The ROC curve plots the possible thresholds of a trained model, created by plotting the true positive rate against the false positive rate. If the threshold is set at 0.3, any probability below 0.3 would cause the model to predict a benign tumor while anything above would predict that the tumor is malignant. In this implementation, sigmoid is used as the activation function while binary cross entropy is used for the loss. Softmax regression is also used as a second approach and it turns out that both types of regression works well for our model. The model is trained over 3000 epochs with the early stopping callback, and the loss and accuracy over the epochs are plotted.




```
# Calculate loss and accuracy of testing data
loss, acc = model.evaluate(X_test, y_test)
print("Test loss: ", loss)
print("Test accuracy: ", acc)
```

```
4/4 [==============================] - 0s 2ms/step - loss: 0.0603
accuracy: 0.9737
Test loss:  0.06029871478676796
Test accuracy:  0.9736841917037964
```

Conclusion

It can be concluded that logistic regression does a great job at predicting whether a tumor is malignant or benign with an accuracy rate of 97%.

References

Wisconsin Cancer Data 2016
UCI Machine Learning Repository

Acknowledgements

Dr Edward Kim