Deep Learning

Module 4 project
Predicting Breast Cancer

LaShanni Butler

Flatiron School

7/26/19

Agenda

- Overview
- Background
- Problem Statement
- Machine & Deep Learning models
- Conclusion
- Recommendations & Future work

Overview

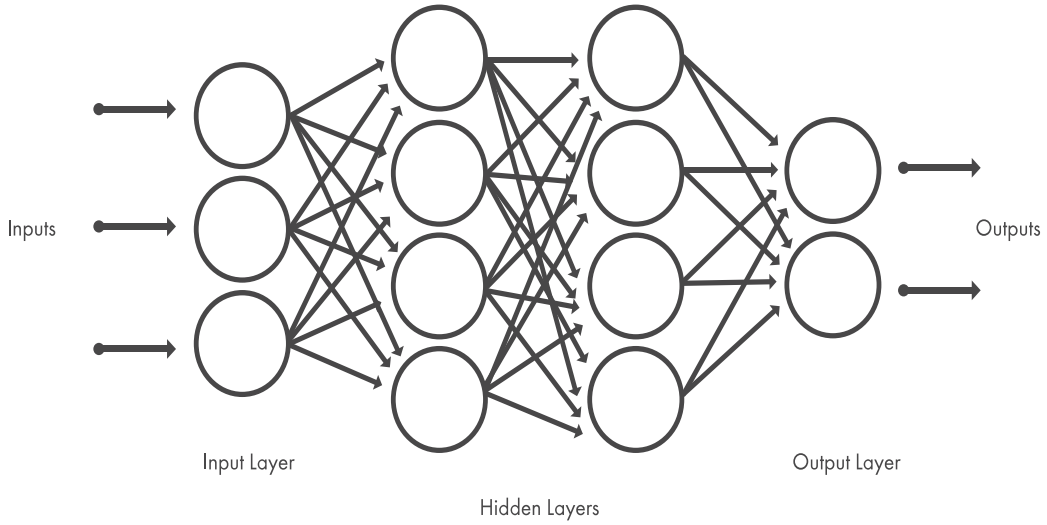
- What is Deep Learning?
  - It's a machine learning technique that teaches computers to do what comes naturally to humans: learn by example
  - In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound
  - These models can achieve state-of-the-art accuracy
  - Examples include:
    - Driverless cars
    - Voice control technology in devices (like Siri)
    - Cancer detection

Sources: https://www.mathworks.com/discovery/deep-learning.html

Background

How Deep Learning Works:

- Deep learning methods use neural network, which resembles the structure (or neurons) of the human brain
- The structures are connected by nodes
- The data passes through the nodes in each layer



ources: https://www.mathworks.com/discovery/neural-network.html

Problem Statement

- The dataset: Breast cancer data obtained from University of California, Irvine. The data contains 569 patient samples and was stained to determine if the patient had cancer (i.e. - malignant) or no cancer (i.e. – benign)

- The Ask: With the dataset provided, can we use a deep learning model to accurately predict breast cancer?

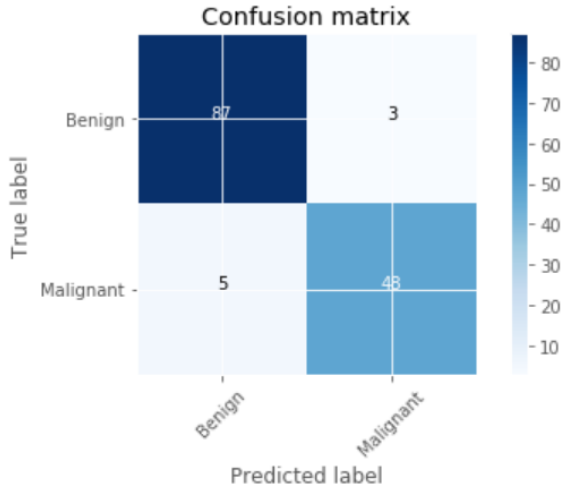
Sources:
Kaggle dataset: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

K-Nearest Neighbors

- KNN is generally used to predict categorical values based on the nearest datapoints of interests

- Confusion matrix: a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

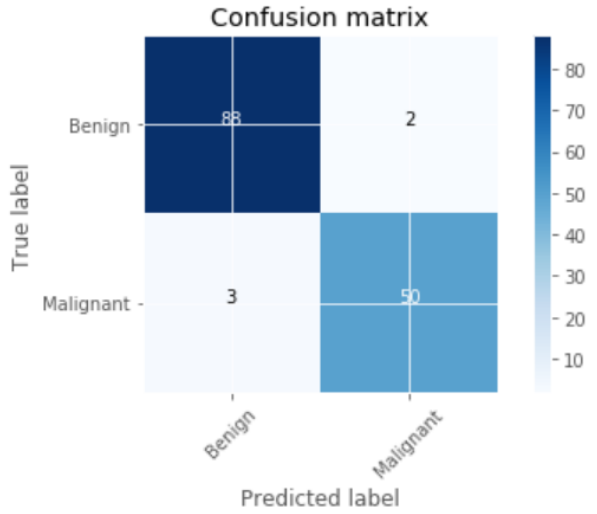
```
kNN Accuracy is 0.94
Cross Validation Score = 0.93
              precision    recall  f1-score   support

           0       0.95      0.97      0.96        90
           1       0.94      0.91      0.92        53

   micro avg       0.94      0.94      0.94       143
   macro avg       0.94      0.94      0.94       143
weighted avg       0.94      0.94      0.94       143
```



Logistic Regression

- The simplest classification algorithm used for binary or multiclassification problems (datasets where y = 0 or 1, where 1 denotes the default class).

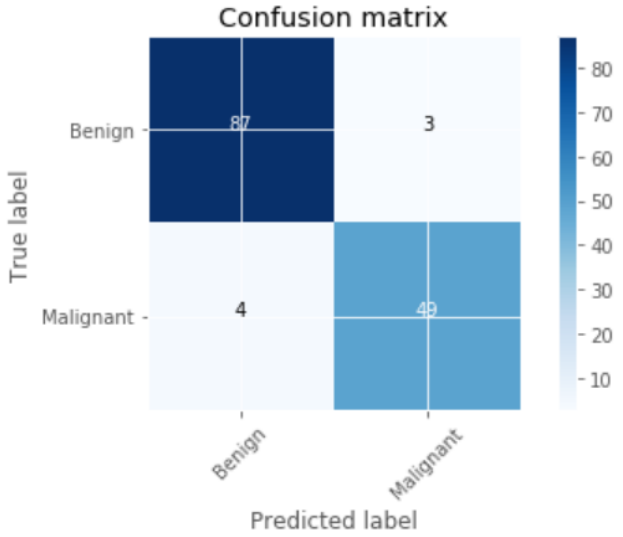
```
Logistic Accuracy is 0.97
Cross Validation Score = 0.95
              precision    recall  f1-score   support

           0       0.97      0.98      0.97        90
           1       0.96      0.94      0.95        53

   micro avg       0.97      0.97      0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.96      0.97      0.96       143
```



Support Vector Machine (SVM)

- Is widely used classification algorithm. SVM creates a separation line which divides the classes in the best possible manner. Ex - dog or cat, disease or no disease.

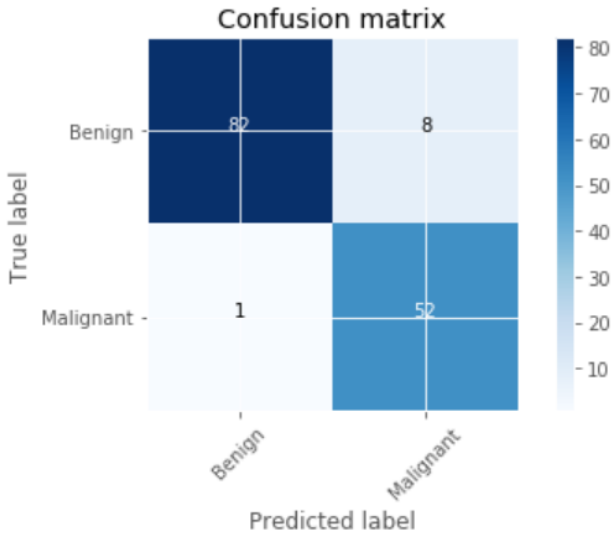
```
SVM Accuracy is 0.95
Cross Validation Score = 0.63
              precision    recall  f1-score   support

           0       0.96      0.97      0.96        90
           1       0.94      0.92      0.93        53

   micro avg       0.95      0.95      0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143
```



Decision Tree

- Is an inverted tree shaped algorithm used to determine a course of action. Each tree branch represents a possible decision

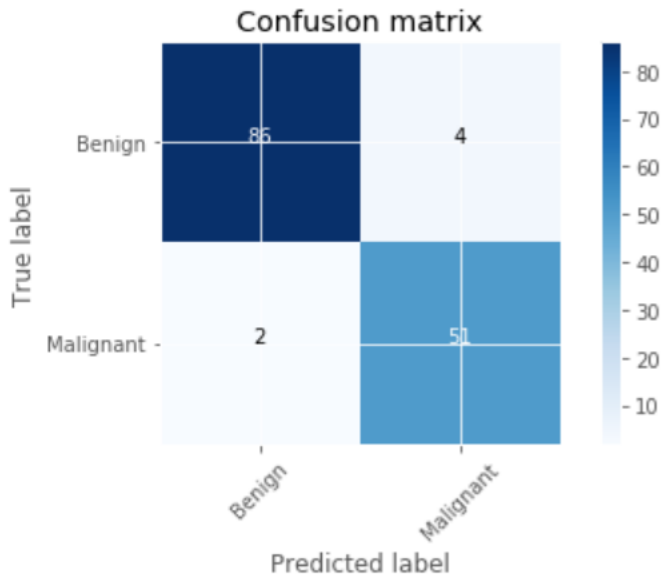
```
Decision Tree Accuracy is 0.94
Cross Validation Score = 0.95
              precision    recall  f1-score   support

           0       0.99      0.91      0.95        90
           1       0.87      0.98      0.92        53

   micro avg       0.94      0.94      0.94       143
   macro avg       0.93      0.95      0.93       143
weighted avg       0.94      0.94      0.94       143
```



Random Forest

- Similar to Decision Tree, but multiple trees are used. Each "tree" observation is classified. Usually increased the accuracy when DT is used as an algorithm.

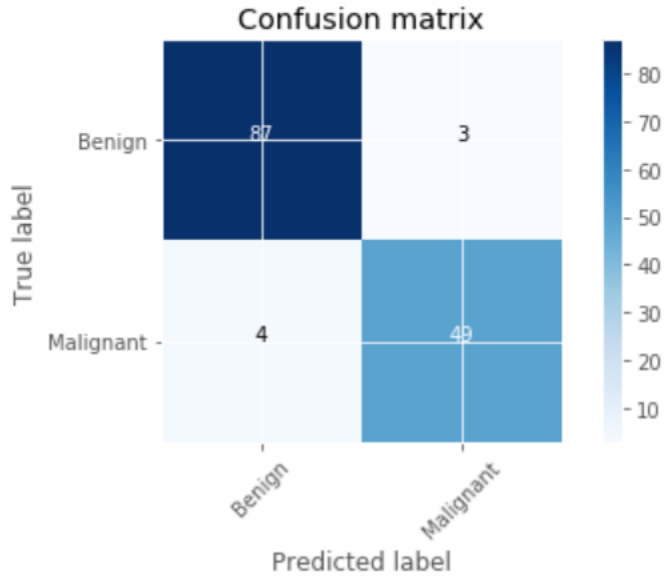
```
Random Forest Accuracy is 0.96
Cross Validation Score = 0.96
              precision    recall  f1-score   support

           0       0.98      0.96      0.97        90
           1       0.93      0.96      0.94        53

   micro avg       0.96      0.96      0.96       143
   macro avg       0.95      0.96      0.96       143
weighted avg       0.96      0.96      0.96       143
```



Keras
(Deep learning)

- Keras is a high-level neural networks focused on enabling fast experimentation.

```
[[87  3]
 [ 4 49]]
Our accuracy is 95.1048951048951%
```



Conclusions

- Logistic regression had the highest level of accuracy (97%) in predicting malignancy
- Random Forest and Keras came in second (96%) in accurately predicting malignancy
- This methodology could help reduce physician burnout and speed up detection
- Accurate detection can reduce the likelihood of a missed diagnosis by a human eye
- Conversely, this methodology isn't perfect. A physician will still have to double check positive results

Recommendations & Future work

- Use Logistic regression for detection, and improve parameters in Keras deep learning model to improve accuracy
- Possibly adjust training and testing set of data to see if that will yield higher accuracy
- Add more features to the dataset, making it more robust for testing
- Explore other deep learning models to determine level of accuracy

Thank You!

- Questions/Concerns/Comments?