BREAST CANCER PREDICTION

Domain Background

Breast cancer occurs when abnormal cells in the breast begin to grow and divide in an uncontrolled way and eventually form a growth (tumour).

Breast cancer starts in the breast tissue, most commonly in the cells that line the milk ducts of the breast. It is the most common cancer in the UK. It mainly affects women, but men can get it too [1].

Sometimes a small sample of breast cells or breast tissue may be taken to help make a cancer diagnosis. This will usually be done using either [2]

1. a core biopsy.
2. a fine needle aspiration (FNA) or another procedure, such as
3. a punch biopsy

FNA uses a fine needle and syringe to take a sample of cells. The samples can then be examined under a microscope.

Problem Statement

Features are computed from a digitized image of a fine needle aspirate (FNA) on a breast mass. They describe characteristics of the cell nuclei present in the image in the 3-dimensional space, full details can be found described in [3].

These features are then used to predict whether the breast mass is malignant or benign.

The Datasets

The dataset is available from Kaggle [4] and is described as the Breast Cancer Wisconsin (Diagnostic) Data Set.

This database is also available through the UW CS ftp server [5] and on the UCI Machine Learning Repository [6].

Kaggle [4] gives the following Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter

d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

Solution Statement and Design

After examination of the dataset for missing values and basic characteristics, variable reduction will be investigated using Principal Component Analysis.

The reduced variable dataset will then be separated into build and test datasets.

Machine Learning techniques available within Amazon Sagemaker will be deployed on the reduced variable datasets in order to initially assess accuracy. Potential methods include

1. Linear Learner
2. XG Boost
3. Neural Network

A further examination of the Linear Learner model focusing on the Model recall as the dominant metric will be conducted, reflecting the importance of avoiding a False Negative result

A Benchmark Model

A possible solution to the task on Kaggle is presented by the Kaggle expert Nisa Soylu [7].

This uses Logistic Regression on normalised variables and the solution has an accuracy 97.7%

Evaluation Metrics

The following metrics will be utilised

1. False Positive Rates
2. False Negative Rates
3. True Positive Rates
4. True Negative Rates

This will additionally allow

5. Precision
6. Recall
7. Accuracy

to then be measured

Accuracy will form the comparison with the benchmark model however Recall will also be explored.

Refereneces

[1] https://www.cancerresearchuk.org/about-cancer/

[2] https://breastcancernow.org/information-support/have-i-got-breast-cancer/core-biopsy-fine-needle-aspiration-fna?gclid=EAIaIQobChMI54rj9-KO7AIVSebtCh2eVQctEAAYASAAEgI0D_D_BwE

[3] K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34

[4] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[5] ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

[6] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[7] https://www.kaggle.com/nisasoylu/machine-learning-implementation-on-cancer-dataset