Machine Learning Engineer Nanodegree

Capstone Proposal

Akshit Gupta
April 23rd, 2020

Domain Background

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed each year, representing about 25 percent of all cancers in women. It is the fifth most common cause of death from cancer in women, with an estimated 522,000 deaths (6.4 per cent of the total). Belgium had the highest rate of breast cancer in women, followed by Luxembourg. So it can help a lot of women to predict breast cancer based on certain features and attributes.

Problem Statement

This is a binary classification problem, based on certain features and attributes the goal of the model is to classify if the woman has malignant or the benign tissues cells.

Datasets and Inputs

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.

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 center 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.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

The data will also be scaled using a scaler.There are Certain NaN which I will remove and I will convert the M and B for malignant and benign to 0 and 1.

Solution Statement

The solution will be predictions of either benign or malignant tissue cell in the test dataset. First I will use sklearn libraries to process all the texts and do some visualization of the data to get some understanding. For training models, I will use Keras and tune the parameter for better accuracy.

Benchmark Model

For this problem, the benchmark model will be a logistic regression classifier. I will try to beat its performance with other algorithms. Evaluation Metrics I will take accuracy and confusion matrix as my evaluation metrics to compare with other models.

Project Design

Before start training models, I will preprocess the dataset and remove null values.I will convert the M and B for malignant and benign to 0 and 1. I will use a Standard scaler to do data

preprocessing. I may perform some graph visualization for a better understanding of the data distribution. I plan to build an ANN using Keras and I think I will use Relu Activation and will try to tune my parameters to generate high accuracy from the model. I plan to visualize the data using a heatmap for the evaluation. This heatmap will show the true positive, true negative, false positive, and false negative, hence helping me to evaluate my model.

References

1. https://www.bcrf.org/breast-cancer-statistics-and-resources
2. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data