MACHINE LEARNING ENGINEER NANODEGREE

BREAST CANCER PREDICTION MODEL

By: Mahmoud Galal Ahmed Abo-Hamda

Date: 23-12-2018

OVERVIEW

1. Domain background

Breast cancer is one of the most commonly spreading types of cancer between women in my home country Egypt and worldwide, and it affected a close people to me, and as this article and the Breast Cancer Foundation of Egypt (BCFE) say that the breast cancer incidences in Egypt has grown to 32% percent and it's predicted to grow more as the population grows. So it's a very important problem to solve and I personally would like to make it my project subject. And it will be a classification model that predicts if the patient have a benign or malign tumor that needs to be cured or examined carefully again.

2. Problem statement

The problem as we said in the previous section is the growing number of the breast cancer incidences in Egypt and worldwide and the difficulties in diagnosing the tumor early. The solution will be training a model to detect if the tumor is malign or benign to ease up the process of identifying the disease.

3. Datasets and input

I'll use a breast cancer dataset that is available for free in the UCI Machine learning repository. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at here. And the attribute information and the relevant papers can be found at the dataset link.

4. Solution statement

The solution is to construct a model that takes the features of the tumor and predicts if it is a malign or benign, and we will train the model on the dataset and use various supervised learning classification algorithms like SVM and decision trees etc. and compare between them and construct the model with the best one like we did in the finding donors for charity.

5. Benchmark model

I found several articles and blog posts that discusses the breast cancer with machine learning and compares several algorithms and chooses the best of them in sites like:

- **Kaggle**
- **Towards data science**

6. Evaluation metric

I'll use the confusion matrix and probably will use also f beta score and will make it close to the recall part of the line as we should be interested more in the false negative area of the confusion matrix.

7. Project Design

- **First I'll gather the data from the dataset and put it in a *pandas* data frame.**
- **Then I'll prepare my data like the scaling process to optimize my model performance.**
- **Then I'll choose my model by comparing between the various algorithms I'll use like SVM and decision tree and I'll use also the ensemble method like gradient boosting, then I'll use the best model and work with.**
- **Then we will train the appropriate model with the data after splitting it into training, validation and testing sets to test the model at last.**
- **Then after training the model we will evaluate its performance and tweak its hyper parameters manually or using grid search techniques.**
- **Then we will use our model in the prediction stage in which we will provide a new data to the model and see the results.**