Providing an optimal predictive machine learning model for the breast cancer tumour diagnosis

Prepared by Sahar Nouri

# Introduction

Breast cancer is the most common malignancy among women. It is the second leading cause of cancer death for women. Breast cancer occurs as a result of abnormal growth of cells in the breast tissue commonly named as a tumor A tumor does not mean cancer. It can be benign (not cancerous), pre-malignant (pre-cancerous) or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are usually used to diagnose breast cancer perform. Nowadays, machine learning has become more accurate than human medical professionals in the diagnosis of cancerous and not cancerous tumors.

# 2. Objective

The goal of this project is to use alternative methods of predictive machine learning algorithms and evaluate their performance to provide an optimal diagnostic model. The model can thereby help clinicians make better decisions for the treatment of breast tumors. In the next section, a description of the data set was provided. In section 4, alternative machine learning methods that will be used are proposed. Section 5 discussed how the models will be evaluated and compared with each other's.

# 3. Data set

The data contains 569 records. This data set was created at the University of Wisconsin Hospital in Madison, Wisconsin, USA. To create the dataset fluid samples taken from patients with solid breast masses were used and an easy-to-use graphical computer program called Xcyt, which is capable of performing the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image.

## 3.1 Attribute Information

The variables of the Wisconsin breast cancer data set are shown below. The ten real-valued features computed for each cell include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The mean, standard error and the largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

* radius (mean of distances from the center to points on the perimeter)
* texture (standard deviation of gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness (perimeter^2 / area - 1.0)
* concavity (severity of concave portions of the contour)
* concave points (number of concave portions of the contour)
* symmetry
* fractal dimension (“coastline approximation” - 1)

The outcome variable is Diagnosis (M = malignant, B = benign).

# 4. Methods

Various machine learning techniques including logistic regression, classification trees and random forests will be conducted to detect existing patterns of the data and predict the outcome variable (Diagnosis). A cross-validation method will be used to avoid model overfitting and hyperparameter tuning will be implemented using Grid Search and/or Random Search to increase the performance of the model. All the analysis will be conducted in a Python (version 3.7) Jupyter Notebook.

# 5. Evaluation

Each model will be evaluated to indicate how successfully the model has been trained to predict the class of outcome correctly. The performances of alternative models will be compared using some criteria such as the accuracy of models, confusion matrices, Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) value. The evaluation of the performance of models on the training data leads to biased results. To avoid this, the data will be split into training and test data. The training and prediction process are described as follows: First, the data is divided into two parts using component splitting. In this experiment, data is split based on a ratio of 80:20 for the training set and the prediction (testing) set. The training set is used in the component for model training while the prediction set data is used in the prediction component. This process is also known as supervision and learning. The trained model will then be used to predict the risk of a malignant tumor.

# 6. References

The data are available at <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>.