**Case Study: Breast Cancer Detection**

Author: Sanjay Jaras

Bellevue University DSC-630

**Overview**

Cancer starts when human cells begin to grow out of control. Breast cancer is a type of cancer that originates in the breast. Breast cancer usually forms a lump of cells in the breast. This lump can be detected through an x-ray image or felt through manual breast inspection. Breast cancer generally occurs in women. Men can get breast cancer rarely. Breast cancer is the most common type of cancer in American women. Approximately 13% of women or one in eight women have chances of developing breast cancer in their lifetime. In recent years, breast cancer incidence rates have increased by 0.5% per year. Breast cancer is the second leading cause of cancer death in women; lung cancer kills more women each year[1].

The University of Wisconsin has collected several digitized images of a fine needle aspiration(FNA) of breast masses with labels. FNA is a type of biopsy procedure. In this process, a fine needle is inserted into an abnormal body growth[2]. Some features are computed from these digitized images. A dataset with computed features and label is prepared. This dataset can be used to train a machine learning model and can be used to detect the nature of the cancer lump.

**Business Understanding**

The features in the dataset describe the characteristics of the cell nuclei present in the FNA image. The dataset contains 570 records with labels[3]. The dataset has the following features.

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. Column 3-32 Characteristics of image

Ten real-valued features are computed for each observation:

1. radius (mean of distances from the center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. 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 example, field 3 is Mean Radius, field 13 is Radius Standard Error, field 23 is Worst Radius.

**Defining a problem**

With this case study, we need to prepare a prediction model that will predict if a given characteristic falls in the Benign or Malignant category. We can treat this problem as the logistic regression problem or classification problem. In this problem, the cost of Type-2 error with Malignant cases(Malignant as positive) is very high. So, we have to be careful while rejecting the null hypothesis(Malignant lump) in the case of Malignant cases.

**Defining the Target Variable**

For this case study, the target variable is very obvious; the Diagnosis variable. This variable has two values B(Benign) and M(Malignant). Benign means non-cancerous lump, and malignant means cancerous lump. In the dataset, we have 357 observations with the Benign label and 212 observations with the Malignant label. We can convert the target variable to a binary variable 0 for Benign and 1 for malignant. We will refer to the Malignant as positive and Benign as the negative from now onwards. Along with the high accuracy, we need to additionally take care that the False-Negative number or Type II error should be less.

**Data Understanding:**

In data understanding, we can observe all the features for data types. For this dataset, except for the Diagnosis column, all columns are of numeric type. The Id column is only for observation identification purposes. Diagnostic column as two values B and M. We can check the distribution of values for each column. We can also use Feature support to find out the p-score; so that we decide to remove any feature that is not having any impact on the target variable.

**Data Preparation**

This dataset needs very little data preparation. We can remove the id column as it is only for identification purposes. Before removing the identity property, we can check for duplicate records and get rid of them. We can check for the data type of each column value. If needed, we can convert them to the numeric type. We can transform the diagnostic column to binary values. We will replace the value B with 0 and value M with 1. We will inspect the distribution for all columns and will remove records with outliers if there are any. We can replace all missing values with Iterative Imputers. The Iterative Imputer will observe all other column values and compare them with the observation with the missing value. Imputer then decides the appropriate value to replace for the absent value. We can use the standard scaler to scale values as values are having different ranges. The Standard Scaling will reduce the skew impact on algorithms like the logistic regressions.

**Feature Selection**

We can use SelectKBest from Scikit learn to find out P-Scores for all features. According to the p-scores, we can get rid of some of the features that do not have a major impact on the target variable.

**Modeling**

Before modeling, we will split our dataset into training and test data sets. We can use an 80:20 ratio for the split as this dataset is small. We can use Stratified splitting based on the target variable to split the dataset as there are fewer records of the Malignant label as compared to the Benign label. We can build several prediction models for this problem. We can start with logistic regression. Then we can train other models like decision tree classifier, Random Forest classifier, K-Neighbors classifier, and XGB classifiers, etc. While preparing the model, we can use k-fold cross-validation as the dataset is small. For the cross-validation split, we can use a Stratified split to remove bias due to one type of observation.

**Model interpretation**

We will compare all trained models for accuracy as well as the confusion matrix by predicting values for the test dataset. As previously mentioned, the model with the small Type II error we need to select. We can finalize a model with high accuracy and small Type II error as our final model for deployment. The comparison of the ROC curve can help us deciding the best model selection.

**Model Deployment**

For all data preparation, we will use the sklearn pipeline. The pipelines will help in applying the same operations on unseen data before applying model prediction. Because this model is used as a medical application, we have to observe the model results frequently. We also need to retrain the model regularly as we are starting with a small dataset. It will also help us in accommodating characteristic changes with time.

**References**

1. <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>
2. Fine Needle Aspiration - <https://www.webmd.com/a-to-z-guides/fine-needle-aspiration#1>
3. Breast Cancer Wisconsin (Diagnostic) Data Set - <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>
4. <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>
5. Dean Abbott – Applied Predictive Analytics – 2016 – Wiley Publication
6. Eric Siegel – Predictive Analytics – 2016 - Published by John Wiley & Sons