Breast Cancer (Wisconsin Dataset) Classification using Stacked Ensemble of Logistic Regression, SVM and Neural Networks

The problem starts with this important question, “Can you identify breast cancer before it occurs?” Early diagnosis significantly increases the chances of survival, but the key challenge is to be able to detect/classify tumors before it becomes malignant. A malignant tumor is called cancer and grows into surrounding tissues or spread to distant areas of the body. They function incorrectly when they are at different parts of the body and tend to not die naturally. For instance, a liver cancer cell would secrete protein dissolving enzymes when it is in the kidney, thus causing malfunction. Modern machine learning techniques dramatically improve the diagnosis; from experienced physicians’ accuracy of 79% to the model’s accuracy of 91% - 97%.

The problem has two main components, preprocessing the data matrix to bring out different features or enhance characteristics that are strong indicators of cancer conditions to implement an Ensemble super-learner to automatically classify the subjects into their respective classification segments. The preprocessing is done by feature engineering the salient features of the data matrix, with methods such as normalizing features, min max scaling amongst other techniques. We then will apply concepts taught in CS532 such as SVM, Logistic Regression and Neural Networks to partition, tune hyper-parameters and train the model to create a highly accurate model. The dataset is a numerical dataset from the Breast Cancer Wisconsin (Diagnostic), UCI Machine Learning Repository. Following is the description of the dataset.

Breast Cancer Wisconsin (Diagnostic) Data Set

Attribute Information:

1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	(2 for benign, 4 for malignant)

In order to accomplish the binomial classification of Breast Cancer Wisconsin (Diagnostic) dataset, we will partition the dataset into a training and validation set into a 75-25 split. A function will be written to ensure similar occurrences of diseases in both the training and validation set. Standard data preprocessing such as standardization will be applied to both the training and validation set using Scikit-Learn packages, but the model algorithm will be implemented from learnings in class. Custom functions will be written to import and manipulate the data into NumPy arrays to be workable in python and shaped correctly for the model development. Any categorial variables, if needed would be one-hot encoded.

A Stacked Ensemble would be a super-learner that uses the outputs of each of the models as inputs to a new model that would again predict the output. This technique is famously used in the Netflix challenge, and will be used in this project to show improvement from standalone models.

Approximate Timeline

Until End-October: Understand and process the data in the right format.

Mid November: Feature engineer salient inputs for the model

End November: Model Development and Output Analysis

December: Reporting and Analysis of Output

References:

[SSBD (https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/)] S. Shalev-Shwartz and

S. Ben-David, Understanding Machine Learning, CUP, 2014

[Data (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))]

Breast Cancer Dataset