**Breast Cancer Prediction**

Milestone 2: Data Selection and Project Proposal

Saurabh Biswas and Gourav Verma

DSC630-T301, Fall-2020

Bellevue University, NE

**Introduction**

**Background**

Breast cancer or breast carcinoma is uncontrolled growth of epithelial cells in the breast. The uncontrollable division of one cell results in visible mass named tumor. Tumor can be benign or malignant. By Johns Hopkins Pathology, benign tumors are non-malignant/non-cancerous tumor and malignant tumors are cancerous growths. A cancer is another word for a malignant tumor. Most benign tumors respond well to treatment. But, malignant tumors are often resistant to treatment, may spread to other parts of the body and they sometimes recur after they were removed. Under a study by University of Wisconsin, 569 patients (212 with cancer and 357 with benign masses) provided the data for diagnostic algorithm. Diagnostic 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 in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

**Problem Statement**

Breast cancer is the most common malignancy in women, with over 200,000 being diagnosed in the US every year. 40,000 women will die from it each year. It is 2nd most common cancer in women. But, on rare scenario can also happen to men. It is 2nd leading cause of deaths of women after lung cancer. A fine needle biopsy is an effective tool in evaluating and diagnosing suspect lumps or masses. With early diagnosis of breast cancer, patients can be isolated for early treatment for a better chance of survival.

**Scope**

Common biopsies for breast cancer diagnosis includes fine-needle aspiration (FNA), core needle biopsy, and MRI-guided biopsy. In this analysis, we will be using ten features of tumor cell nuclei extracted from the digital image processing of an FNA of a breast mass to predict breast cancer. The data is collected from UCI Machine learning repository.

**Document Overview**

This proposal will be broken down into five sections: Introduction, Preliminary Requirement, Technical Approach, Expected Results, and Management Approach. Many of these major sections will contain sub-sections that will show the tables, figures, and analysis that went into the predictive model.

**Preliminary Requirement**

**Technical Approach**

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining and has been an established and well documented process since 1996.[3] The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

It is important to note that there are several feedback loops in-between several of these steps as insights are gained through various activities that further reinforce or introduce ideas that benefit the project. So several of the steps will be revisited as the project progresses.

**Data Sources and Variables**

Data can be found on UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

Attribute Information:

1. ID number
2. Diagnosis (M = malignant, B = benign)

3-32) ten real-valued features are computed for each cell nucleus:

* 1. radius (mean of distances from 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 instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

**Analysis**

Our dataset contains 30 different variables that could contribute to our predictive model. We realize this is simply too many, so we have developed a process by which we will select the variables of interest. According to our text, “Applied Predictive Analytics”, many predictive algorithms assume the model variables follow a normal distribution. There are inherent advantages to using normally distributed variables, so our approach will focus on columns that closely follow this distribution.

We will determine which variable are normally distributed by conducting the following analyses:

* Summary statistics on all variables—concentrating on mean and standard deviation values
* Skewness of variables—looking for variables with values less than two and as close to zero as possible

**Requirement Development**

We plan to rely almost exclusively on Jupyter Lab, building a notebook leveraging the Python language. This will allow me to output presentation-quality documentation at the conclusion of the project via Markdown. In addition, if any R coding is required, we can add that into the same notebook via available Python libraries.

**Model Deployment**

We are going to build classification model and evaluate its performance by the training set. With use of different plots we will determine important variables to create different prediction model such as Logistic regression, random forest model, or gradient boosting machine for prediction/classification of benign and malignant tumor.

**Testing and Evaluation**

We will use train-test split of 70%-30% and will test the model with test split. For evaluation confusion matrix, AUC, and F1 score will be used. For some models cross-validation will also be used to decide best model.

**Expected Results**

We are planning to find the nuclei features with more predictive value for the diagnosis.

**Execution and Management of Project**

**Project Plan**

At this point in time, our plan is simply to follow the Milestone requirements and timeline.

* Week 1: Milestone 1 Due (Team Information/Communication Plan)
* Week 2: Milestone 2 Due (Data Selection and Project Proposal) & Peer Review
* Week 4: Peer Review
* Week 5: Milestone 3 Due (Preliminary Analysis)
* Week 6: Peer Review
* Week 9: Milestone 4 Due (Project Presentation & Status) & Peer Review
* Week 10: Peer Review
* Week 12: Milestone 5 Due (Final project paper and presentation) & Peer Review (Due Saturday!)

**Project Risk**

A preliminary viewing of the data reveals many variables are of equal importance. The challenge would be to find the most important features. The other challenging task would be to find the model with best fit.

**Resources**

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995;130:511-516.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792--796, 1995.