PROJECT PROPOSAL

BREAST CANCER WISCONSIN PREDICTION

Ameya Jamgade | BANA 8083 | May 10 2018

1. INTRODUCTION

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. Breast cancer ranks second among cancer deaths in women.

This project aims at analyzing data of women residing in the state of Wisconsin, USA for prediction whether the case of breast cancer is malignant or benign. The project will include application of several data mining and machine learning techniques to classify whether the tumor mass is benign or malignant in women. This will eventually help in understanding the important underlaying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes.

1.1 Dataset and data description

The data for this project is obtained from UCI Machine Learning repository. The link for the data can be found at:-
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Features in the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The dataset contains information of 569 women across 32 different attributes.

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

2) Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)

i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The variables are divided into three parts first is Mean (3-13), Stranded Error (13-23) and Worst (23-32) and each contain 10 parameters (radius, texture, area, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension). Mean is the means of the all cells, standard Error of all cell and worst means the worst cell

2. PROPOSED SOLUTION OVERVIEW

2.1 Data Cleaning and Preparation

This stage would include identifying if the data set contains any missing values or bad data. I would the convert the diagnosis (Y variable) into appropriate format (currently it is classified as M and B for malignant and benign respectively)

2.2 Exploratory Data analysis

The EDA process will help in understanding the nature of the dataset and also help in identification of potential outliers or correlated variables.

2.3 Modelling

Since the dataset contains around 32 different attributes, I would try to incorporate principal component analysis to identify the important attributes.

For modelling, I am planning to use K-nearest neighbor, Random forest and Support Vector Machine algorithms for classification of the diagnosis Y-variable in the data as malignant or benign. Depending on the prediction accuracy and time constraint I might try to include Gradient Boosting Model technique in my analysis.