Breast Cancer Prediction

Introduction

This project is about predicting whether the breast cancer is benign or malignant. Given the diagnostic data set from Wisconsin, we are going to do EDA on this dataset and extract interesting pattern from it. The final goal is to predict the diagnostic results based on the provided characteristics of the cell nuclei presenting in the image and achieve high accuracy in our prediction.

Data Mining Task

1. Understanding the data set. This includes analyzing and visualizing the features and their distribution.
2. Preprocessing the data set. Implementing data cleaning, transformation and reduction to prepare our data set for modeling.
3. Modeling the data set. Applying different model to make diagnostic prediction
4. Evaluation. Based on the results from modeling, we compare performance from different model and determine whether some classifiers are more superior than others statistically.

Data Set

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

This dataset is provided by Kaggle and donated by UCI Machine Learning Repository in 1995-11-01. This dataset has been largely explored for predict the breast cancer.

1. The dataset contains only one single csv file without separate train and test datasets.
2. All of instances are labeled with diagnostic results. There are 357 benign, 212 malignant instances.
3. There are 3 statistical values computed for 10-real valued features : the mean, standard error and "worst" or largest (mean of the three largest values) which results in 30 features in total.

Methods and Models

1. **Data understanding**: data description and feature analysis (statistical table), visualization.
2. **Data preprocessing**: data cleaning (imputation), normalization. Implementing the process of reduction using mutual information, Pearson's Correlation value and PCA if necessary.
3. **Sampling:** K-Fold Cross-Validation. (k = 10)
4. **Classification model:** Perceptron, K-Nearest Neighbors (varying n), Naive Bayes (Gaussian or Multinomial if applicable), Decision Trees, Logistic Regression. And we will also try a few advanced methods such as Assemble (Random Forest, XGBoost) and Artificial Neural Networks.
5. **Evaluation:** We calculate **accuracy score** and **AUROC** for each classifier. To determine which methods are superior to others, we will run each model randomly multiple times to gain average performance and use P-value to evaluate confidence statistically. The successful model should be the one with statistically better performance.

Team Members

Jun Guo, Chenlu Huang, Maoliang Liu

Schedule

Due Date	Tasks to be Completed
Jul 27	Data Understanding & Data Preprocessing
Aug 1	Classification & Regression
Aug 5	Validation & Interpretation & Advanced Topics
Aug 7	Prepare the final deliverable and presentation