Capstone 2 Project Proposal: Breast Cancer in Wisconsin

Tzu-Ying Chen

# Problem

There are numeric different characteristics of tumors that the breast cancer patients would have. The accuracy and efficiency of the diagnosis could significantly impact the patients’ recovery and potentially decrease the complexity of the treatment. Hence, it is important to help doctors to identify the characteristics of the tumors that patients have in a timely manner in order to provide necessary healthcare.

# Goal and Utility

The goal of this project is to analyze the symptoms and tumors characteristics, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension that breast cancer could have at different stage, and provide prediction on whether the tumor is malignant or benign. The analysis and the prediction can quickly help doctor prescribe necessary treatment and healthcare.

# Data

The data set for this project contains 569 observations including patient ID number, and the features of the cell nucleus, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The data set is available on UCI Machine Learning Repository. The page contains information about the data set and its source and relevant scientific publications:

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

# Approach

It is a supervised classification task. The diagnosis (whether the tumor is malignant or benign) will be our target variable, while the characteristics of the cell nucleus will be used as the predictive variables. Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Random Forest Classification are some of the machine learning models that could potentially be used in this project. Data exploratory analysis will provide deeper understandings of the data and address the problem more clearly, and cross-validation could ensure the outcomes of the model is reliable.

# Deliverables

The deliverables for the project will include the source code, the data set, and a paper addressing the purpose, approach, findings and results of the project.