Neha Bhoi
Liam Nguyen

Final Project Proposal

Neha Bhoi, Liam Nguyen

Project objective

To continue the exploration of supervised learning algorithms, **Random Forest** is a natural choice. It is a flexible and easy to use classification and regression algorithm that produces great results in most datasets. It is also one of the most used algorithms, because of its simplicity and diversity. Although scikit-learn has a RandomForestClassifier, we want to implement the algorithm from scratch.

Dataset

The dataset that we plan to use is a public dataset from UCI Machine Learning Archive. Here are the information about the dataset:

- Link: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
- Name: Breast Cancer Wisconsin (Diagnostic) Data Set
- General Information:
    - The attributes are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.
    - There are 10 features for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension.
- Goal: train model to predict the diagnosis (M = malignant or B = benign)
- Number of instances: 569
- Type: Classification, Public Data
- Data format: CSV

Method

Random forest consists of many decision trees. Each decision tree is essentially a flowchart of questions and answers leading to a decision. Since we use a combination of many decision trees in the forest, the algorithm typically provides a very high accuracy, which is important in the diagnosis of breast cancer. The algorithm is also not sensitive to outliers or missing values which might exist

in our dataset. It is also useful to identify variables that are most helpful in prediction since it splits based on the attribute significance. Therefore, we believe Random Forest is the most fit algorithm to predict the type of breast cancer with a given set of features.