456 Project Proposal

Junhyeok Park, Xin Lu

Team Members & Designated Presenter

- Names: Junhyeok Park, Xin Lu
- Designated Presenter: Junhyeok Park

Dataset Description

The Breast Cancer Wisconsin dataset was created by Dr. William H. Wolberg (a physician) from the University of Wisconsin Hospitals in 1990. The dataset contains 699 observations with 11 attributes, including `class`, which indicates whether the clump is benign or malignant. The remaining 10 variables are as follows:

- Sample Code Number: Unique ID number
- Clump Thickness: 1 - 10
- Uniformity of Cell Size: 1 - 10
- Uniformity of Cell Shape: 1 - 10
- Marginal Adhesion: 1 - 10
- Single Epithelial Cell Size: 1 - 10
- Bare Nuclei: 1 - 10
- Bland Chromatin: 1 - 10
- Normal Nucleoli: 1 - 10
- Mitoses: 1 - 10

Each attribute (except `Sample Code Number`) assigns scores from 1 to 10 to indicate the strength of the variable.

Questions & Methods

Our group intends to use all of the observations and variables except Sample Code Number. We aim to answer the following questions: - What is the overall distribution of the variables? Are there any outliers/errors in the dataset that we should address before beginning the analysis? - Which variables are most important in classifying clumps as benign or malignant? - Which variables most capture the variance in the data? - What are the patterns within each class? Are there any characteristics that are clearly distinguishable between the two classes?

We are considering two methods for the analysis. The first option is Principal Component Analysis (PCA), which we could use to reduce the dimension of the dataset and capture the interesting directions in the data with the most variance explained. This analysis will help us identify the components that explain whether the patient's clump is benign or malignant. The second option is Cluster Analysis, which we could use to group the observations after selecting the most important variables. Then, comparing our analysis with the actual class value will help us assess our analysis.

Source

- Breast Cancer Wisconsin (Original) Data Set
- O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.