Week 1 Assignment - MSc AIDA UoM

Professor Nikolaos Samaras

Academic Year: 2021-2022

At the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/index.php`), there is a dataset named *Breast Cancer Wisconsin (Diagnostic) Data Set* (`https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`).

This dataset includes anonymized medical data for breast cancer from 569 patients, of which 212 have malignant cancer and 357 have benign cancer. The features for each patient were calculated from a digitized image of a fine needle aspirate (FNA) of the breast mass.

From the Data Folder, use the files:

- wdbc.data (the dataset's data)
- wdbc.names (dataset's details)

Feature Information:

- ID number
- Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus (columns 3-32):

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (`perimeter^2 / area - 1.0`)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

Tasks:

1. Read the `wdbc.data` file and convert the values of the 'Diagnosis' column from categorical to binary, i.e., M to 1 and B to 0.
2. Calculate for the class variable 'Diagnosis': Mean, Median, Low Mean, High Mean, Standard Deviation of the Sample, and Variance of the Sample.
3. Plot a bar chart for the class variable 'Diagnosis' showing the Mean and Variance.
4. Use the LogisticRegression algorithm to develop a prediction model where the dependent variable is 'Diagnosis' and the independent variables are the remaining columns except the first (ID Number).