Week 2 Assignment - MSc AIDA UoM

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 (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. Apply PCA for dimensionality reduction. Select principal components until the sum of explained variance is at least 75% (or 75%).
2. Choose randomly 85% of the records as a training dataset, ensuring the same ratio between classes as in the entire dataset. Use the remaining 15% as a test dataset.
3. Develop a prediction model using Logistic Regression, where the dependent variable is 'Diagnosis' and the independent variables are those derived from PCA, excluding the first (ID Number).
4. Perform a significance test between the models developed in weeks 1 and 2. Re-run the week-1 model using the same training and test dataset configuration as in this week's task (2). Use an appropriate evaluation technique from Lecture02 slides.