ENSTA-Paris, ROB311 – 16/09/2019

Write a python program to implement K-NN to work with several datasets. Please verify with the 2 datasets below. Calculate also the confusion matrix, the accuracy, and plot the data.

Datasets - UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

1. Breast Cancer Wisconsin (Diagnostic) Data Set

LINK: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

DOWNLOAD: `Data Folder` → *breast-cancer-wisconsin.data*

The data set contains 2 classes. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute information:

1. Sample code number	*ID number (note: **to be ignored**)*
2. Clump Thickness	*1 - 10*
3. Uniformity of Cell Size	*1 - 10*
4. Uniformity of Cell Shape	*1 - 10*
5. Marginal Adhesion	*1 - 10*
6. Single Epithelial Cell Size	*1 - 10*
7. Bare Nuclei	*1 - 10*
8. Bland Chromatin	*1 - 10*
9. Normal Nucleoli	*1 - 10*
10. Mitoses	*1 - 10*
11. Class:	*2 for benign, 4 for malignant*

Class distribution: benign: 699 instances, 458 (65.5%), malignant: 241 (34.5%)

Notes:

- there are 16 instances that contain a single missing attribute value, denoted by "?"
- the sample code number is **not** to be used as an attribute, it is just a unique ID

More details can be found in the *breast-cancer-wisconsin.names* file.

2. Haberman's Survival Data Set

LINK: https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

DOWNLOAD: `Data Folder` → *haberman.data*

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute information:

1. Age of patient at time of operation	*numerical*
2. Patient's year of operation	*year - 1900, numerical*
3. Number of positive axillary nodes detected	*numerical*
4. Survival status (class attribute):	*1 = the patient survived 5 years or longer* *2 = the patient died within 5 year*

Class distribution: 306 patients, 225 patients (73.6%) survived ≥ 5 years, 81 patients (26.4%) died within 5 years

More details can be found in the *haberman.names* file.