1. DEVELOPMENT OF AN INTERPRETABLE MODEL FOR BREAST CANCER PREDICTION

1.1 Filename

— folder: data for case 1
— data.csv: file with a list of ids, labels and features to classify breast cancer

1.2 Columns / attribute information

The following numbers correspond to the respective columns in the file.

1) ID number
2) Diagnosis (Labels: M = malignant, B = benign)
3-32) Ten real-valued features are computed for each cell nucleus:

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

The mean (extension „_mean“), standard error (extension „_se“) and "worst" or largest mean (mean of the three largest values, extension „_worst“) of these features were computed for each image, resulting in 30 features.

All feature values are recoded with four significant digits.
Missing attribute values: none
Class distribution: 357 benign, 212 malignant

1.3 Further information about the dataset

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. The 3-dimensional space is described in K.P. Bennett & O.L. Mangasarian, 1992. The data can also be found on UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Introductory literature:

K. P. Bennett & O. L. Mangasarian (1992). Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimization Methods and Software 1*, 23-34.