hugomcp@di.ubi.pt, 2024/25

Machine Learning

Practical Sheet 5: Dimensionality Reduction

Consider the “breast Cancer” dataset, available at the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

This set contains 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. There are 30 features in this set plus the first two columns in each row, that provide the row IOD and the class information (M=Malign, B=Benign).

1) ID number
2) Diagnosis (M = malignant, B = benign)

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, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

There are a total of 580 instances, with the following class distribution: 357 benign, 212 malignant

1) Create a “Google Colab Python script” that loads the dataset, and divides the normalized data into “Learn”, “Validation” and “Test” subsets.

2) Normalizes the available data according to min-max or z-score criteria.

3) Create a neural network that learns to distinguish between “Malign” and “benign” cases.

4) Use PCA to reduce the dimension of the dataset.

    a. Create different data versions that keep 99%, 95%, 90% of the variability

5) Create neural networks that distinguish between “Malign” and “benign” cases in the compacted spaces.

6) Compare the effectiveness attained by the different neural networks.