Basic k-nearest neighbor rule

Letting the dataset $D^n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ denote a set of n labeled classes and letting $\boldsymbol{x}' \in D^n$ be the class nearest to the test point **x**. Then the *nearest-neighbor rule* for classifying x is to assign it the label associated with $\boldsymbol{x}'$.

In the MATLAB code, we do the following work:

1. Preprocessing the dataset

We use Z-score standardization to preprocess both training dataset and testing dataset, which standardizes the original dataset making mean to be 0 and variance to be 1 for each dimension:

$$\mathrm{z} = \frac{x-\mu}{\sigma} \tag{1}$$

NOTICE:

If σ equals to 0, then we set the whole dimension to 0.

2. Leave-one-out and selecting the k value

When k varies from 1 to 10, we take each sample out in the given training dataset and make classification with the rest samples, then calculate the classify accuracy rates corresponding to different k values. Finally we decide k as the number generates the highest accuracy rate. The table below shows the k values we chose for each dataset.

3. Using the selected k value and classifying the testing dataset.

Euclid distance is used to present the distance between the two samples.

For each testing sample, we calculate the distances between training samples and rank the distances in ascending order. Then we labeled the testing sample as the class that appears most among the k nearest neighbors.