Pattern Recognition Laboratory - Warmup (#1) Nearest neighbour classification

October 2019

Your task is to prepare nearest neighbour classifier recognizing two species of irises. Data sets are in files `iris2.txt` and `iris3.txt`, containing measured features values for species *iris versicolor* and *iris virginica* respectively.

Each line of these files contains:

1. Serial number of the sample
2. Sepal length
3. Sepal width
4. Petal length
5. Petal width

The classifier itself is very simple (since you don't have to optimize it for bigger data, it's implementation will have a few lines of code):

1. Compute distances from the sample to be classified to all samples in the training set.
2. Find the training sample to which this distance is minimal.
3. Set the label of this training sample as the classifier answer.

Real work starts for you by assessing classifier's quality. Because we have only 100 samples at our disposal, traditional partitioning of the set into training and testing parts can have serious impact on classifier's quality (traditional partitioning means here for example 80% training set, 20% testing set). To measure expected error rate you'll use *leave-one-out* (also known as the *jackknife* method), in which test set is as small as possible:

1. Select one sample from the data set – this sample forms *test set*.
2. The rest of the data set will be used as the training set for the 1-NN classifier (there are 99 samples left, so the classifier will be very similar to the final one which uses all 100 samples).
3. Selected in step 1 sample is classified and the result is stored.
4. Steps 1-3 are repeated for all 100 samples in our data set. Number of classification errors divided by the size of our data set is quite good approximation of the error rate of 1-NN classifier.

Assess error rates for classifiers using only one feature and the combination of 2, 3 and 4 features.