

1. Laboratory work №3

"Method of classification of k-nearest neighbors"

Purpose: In laboratory work it is necessary to classify Fisher's irises by the method of k-nearest neighbors. Data on iris flowers are used for classification. The target column corresponds to the type of flower ('Species').

The K-Nearest Neighbor (KNN) algorithm uses "feature similarity" to predict the values of new data points, which also means that a new data point will be assigned a value based on how closely it corresponds to points in the training set.

Task:

1. Select 2 column attributes for classification: length ('PetalLengthCm') and width ('PetalWidthCm') of the flower.
2. Next you need to select the value of k, i.e. the nearest data points. k can be any integer. (recommended k = 5).
3. Divide the sample into training (80%) and test (20%) randomly, because the data is organized.
4. For each point in the test data, do the following:
 - 4.1. Calculate the distance between the object coordinate data and each row of training data using the Euclidean distance method (determine the length of the vector for each neighbor).

Example of calculation

```
math.sqrt((a[0] - b[0])**2 + (a[1] - b[1])**2)
```

$$D_E = \sqrt{\sum_i^n (x_i - y_i)^2},$$

- 4.2. Based on the distance value, sort them in ascending order (choose any data sorting method).
- 4.3. You must select the top k rows from the sorted array.
- 4.4. Assign a class (iris type) to a checkpoint based on the most common class of these strings.

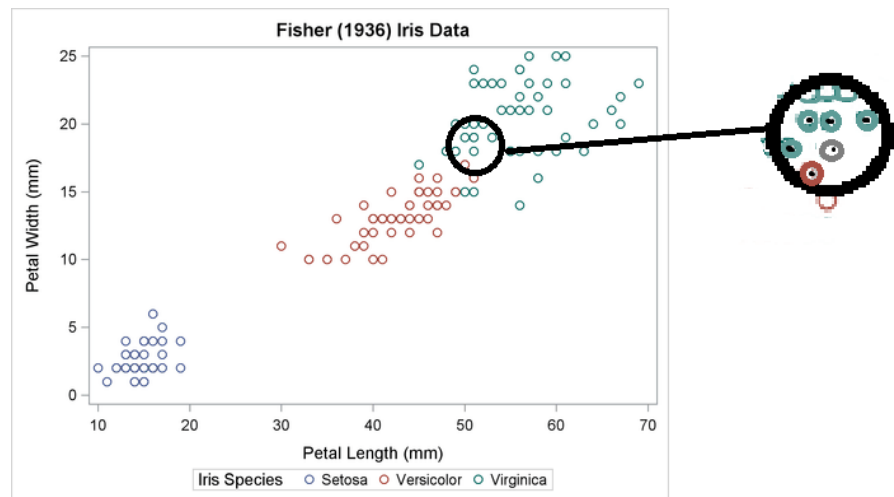


Figure 1. Example of one iteration of the algorithm

Figure 1 graphically shows one iteration of the algorithm. The test object is highlighted in gray, and the nearest neighbors are highlighted in green and red.

To check the correct operation of the algorithm, calculate the error. Evaluate the result of your classification with the original data.