Data Mining Exercise 4: 17.02.2022

Remember to enroll yourself on the examination of the current course no later than 8 days before an examination date!

1. Iris.txt file contains the leave information of three different Iris species. First column is just a running number and can be skipped. Next four columns are the dimensions of leaves and the sixth column contains class label of the Iris in each rows. Select case from fifth row and search its nearest neighbor using Euclidean distance. Does the nearest neighbor belong to the same class as the case from fifth row?

2. Normalize the variables so that the means become zero and the variances become one. Repeat the first task using normalized data.

3. Scale the variables into the interval [0 1] and repeat the first task. Scaling forces the minimum values to zero and maximum values to one.

4. Boxplot figure shows that variable 2 has some outliers. Search for the row numbers of the outliers in the original data.

5. Study the correlation structure of Iris data. What variables have the highest correlation?

6. Reduce the dimension of Iris data by replacing the highly correlated variables using their mean. Repeat the first task using the reduced data.

7.Propose another way to reduce the dimension of Iris data than that given above. Repeat again the first task.

8. Tasks 2 and 3 give us transformed presentations of original data. However, both ways share the common problem that is related to whole population of data and sample from a population. What could this problem be?