COMP 462 ASSIGNMENT1

.

k-Nearest Neighbor Classification

Mert Dogan

MEF University
Department of Computer Engineering
doganme@mef.edu.tr
041701041

Abstract. In this assignment, k-NN classification algorithm were implmented and tested by using the Iris dataset. k-NN classification algorithm implementation has two important input parameters which are number of neighbours and distance metrics.

1 Dataset

Iris dataset contains three flowers, as shown in Figure 1. Each flower is represented by four features: 1=sepal length, 2=sepal width, 3=petal length, and 4=petal width. In this assignment, first and fourth features were used. The iris data imported as a csv file.

Fig. 1. Three flowers in the Iris dataset: Iris Versicolor, Iris Virginica and Iris Setosa.

1.1 Training and Test Sets

In the iris dataset, each flower has 50 samples. In this implementation, 30 samples were put from each flower class into the training set and the rest of the samples were put into the test set.

1.2 Actual Data

In this section, given figure shows a plot which represent the actual data with using matplotlib library.

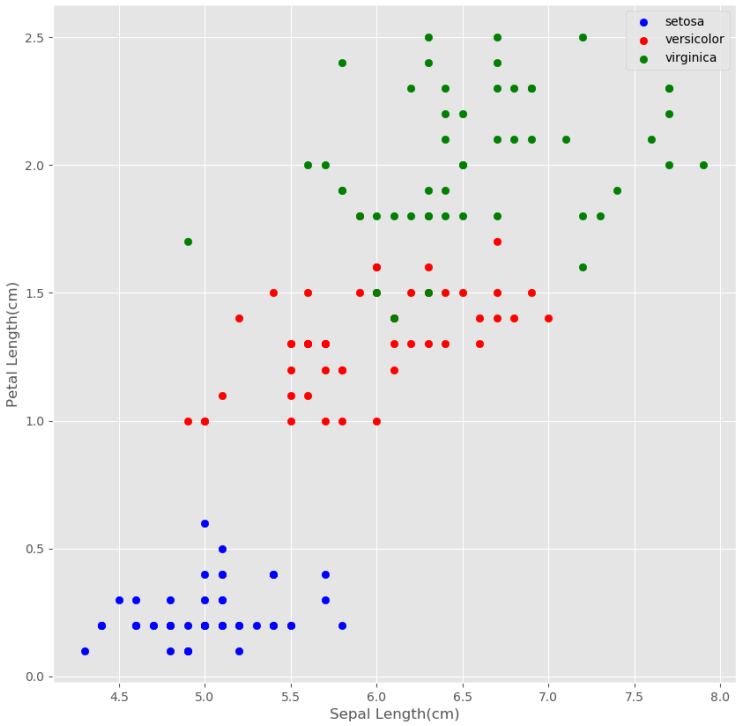

Fig. 2. Actual iris dataset

1.3 Predicted Data

In this section, given figure shows a plot which represent the predicted data with using matplotlib library.

1.4 Classification results

k-NN algorithm were applied to classify test samples. Different k values were tried as shown in Table 1. Three different distance metrics were used: 1) Euclidean distance, 2) Manhattan distance, and 3) Cosine distance.

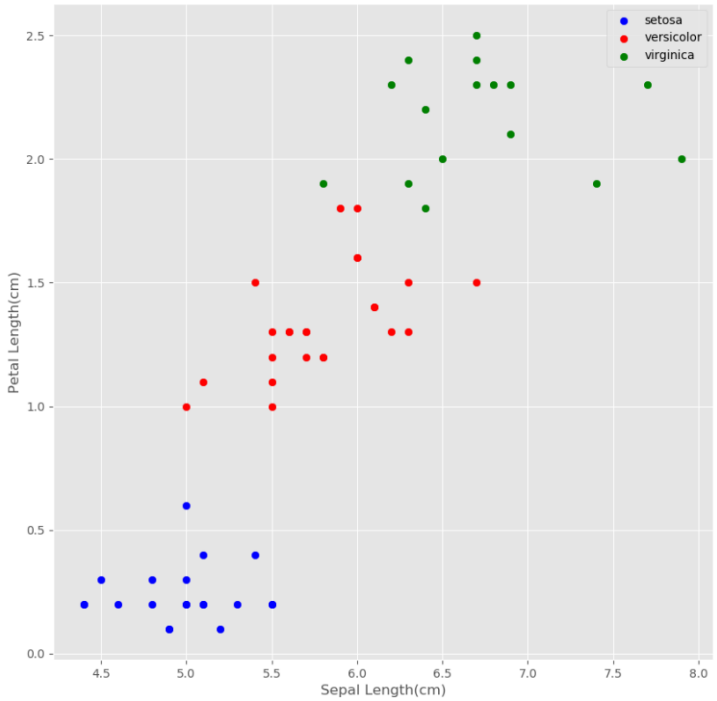

Fig. 3. Predictions.

Table 1.k-NN classification accuracies for different k and distance metrics. For each case, classification accuracy (in percentages) and misclassified test sample counts are given.

Table 1. k-NN classification accuracies for different k and distance metrics.

k-NN	Euclidean Distance	Manhattan Distance	CosineDistance
	Accuracy(%) - Error Count	Accuracy(%) - Error Count	Accuracy(%) - Error Count
k=1	93,33 - 4/60	93,33 - 4/60	86,67 - 8/60
k=3	96,67 - 2/60	96,67 - 2/60	91,67 - 5/60
k=5	96,67 - 2/60	96,67 - 2/60	88,33 - 7/60
k=7	96,67 - 2/60	96,67 - 2/60	88,33 - 7/60
k=9	96,67 - 2/60	96,67 - 2/60	91,67 - 5/60
k=11	96,67 - 2/60	96,67 - 2/60	91,67 - 5/60
k=15	96,67 - 2/60	95,00 - 2/60	88,33 - 7/60

2 Questions

1) Which distance metric is useful for the Iris dataset?

2) Which distance metric is worst for the Iris dataset?

3) In general, which k parameter is good for the Iris dataset?

2.1 Explanations

1) Euclidean Distance is useful for the Iris dataset. Because the accuracy rates clearly indicate, using Euclidean Distance will give the best result.

2) Cosine Distance is not quite useful for the Iris dataset. Because the accuracy rates clearly indicate, using Cosine Distance will give the worse result than the other distance methods.

3) Taking the k value three (k=3) will give the best result for the 3 distance methods.