Statistical Machine Learning – Week 2

Simon Walser

HS 2021/22

1 Task

In this lab you will implement a k-nearest-neighbor classifier and apply it to the following dataset.

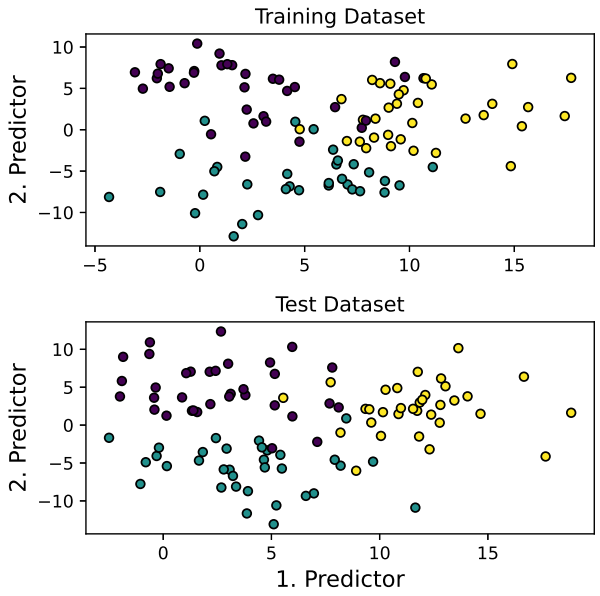

Figure 1: The upper scatter plot shows 100 training data samples and the bottom plot shows 100 test data samples. This is a synthetic dataset which consists of three different classes.

Your task is to complete the implementation of the kNN classifier using the template KNN_.m which is invoked by week2_exercise.m. Next, you evaluate how the hyperparameter k affects the performance of the classifier.

- Open the MATLAB script KNN_.m and complete all TODOs. First, fill the matrix *ed* which contains the euclidean distances between every test data sample and every training data sample (matrix of size 100x100).
- For every test data sample sort the training data samples by distance and store the corresponding training sample indices in *ind* (closest training sample indices are stored in first columns of *ind*).
- Classify all test samples based on a majority vote of the k nearest training samples using *ind*.

- Call the function KNN_ using the MATLAB script week2_exercise.m. Change the value of the hyperparameter k and analyze the effect on the performance of the classifier.
- The script week2_exercise.m plots the decision surface of the classifier for different values of k. Analyze the shape of the decision boundary as a function of k and answer the following questions:
  - How does the hyperparameter k affect the bias/variance trade-off?
  - How big is the training error rate if k is set to one?
- (optional) Try to reproduce figure 2.17 on page 42 in the textbook i.e. plot the training error rate and test error rate as a function of 1/K.

Comments

- The implementation of the KNN was copied from the following source: Mahmoud Afifi (2020). kNN classifier (`https://www.mathworks.com/matlabcentral/fileexchange/63621-knn-classifier`), MATLAB Central File Exchange. Retrieved September 11, 2020.