W3WI DS304 Applied Machine Learning Fundamentals

Exercise Sheet # 5 - k-Nearest Neighbors (kNN)

Question 1 EX 2020

You use a k-nearest neighbors classifier and set $k = n$, where n is the total number of data points in the dataset. Which class is predicted by the classifier?

Question 2 EX 2020

The decision boundary shown in figure 1 was generated by a k-nearest neighbors classifier. How do you rate the performance of the classifier? What might be problems and how could they be mitigated? Can you guess the value of k which was used? *(Explain your answer!)*

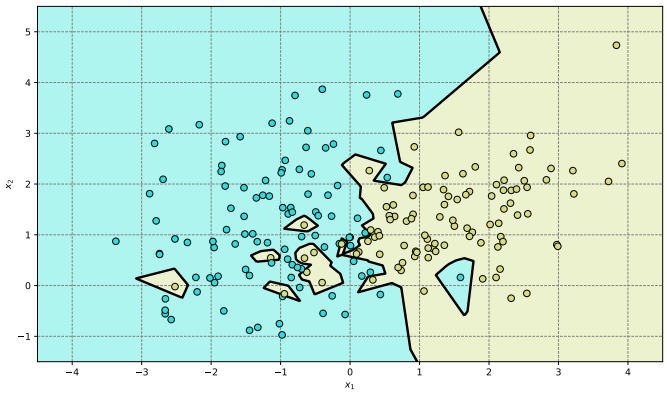


Figure 1: Decision boundary of a k-nearest neighbors classifier.

Question 3 EX 2020 (Non-parametric methods)

You have a dataset consisting of 500,000 data points. Your boss suggests to use a non-parametric method for classification (e. g. a k-nearest neighbors classifier). What does *non-parametric* mean? Do you agree with your boss? *(Please explain your answer.)*

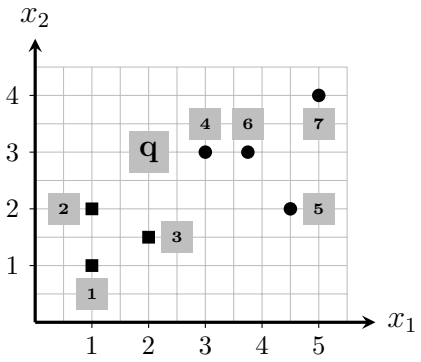
Question 4 EX 2023 (k-nearest neighbors algorithm)

The instances in the following training dataset (see figure 2 below) belong to either of the two classes ■ or ●. Your goal is to classify the unknown data point $q := \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ using the k-nearest neighbors algorithm. You choose $k := 3$.

1. Calculate the prediction a) using the **Manhattan distance**, and b) using the **Euclidean distance**. Do both distance metrics lead to the same result?

2. Suppose you had chosen $k = 7$. Which class would have been predicted? What problem do you see?

3. Illustrate two possible **tie breaking strategies** in case that both classes appear equally often in the neighborhood of q!



Row	x_1	x_2	Class
1	1.00	1.00	■
2	1.00	2.00	■
3	2.00	1.50	■
4	3.00	3.00	●
5	4.50	2.00	●
6	3.75	3.00	●
7	5.00	4.00	●

Figure 2: Illustration of the training data set.

Question 5 EX 2023

Tick the correct statements concerning the k-nearest neighbors algorithm!

☐ The k-nearest neighbors algorithm is model-based.

☐ k can be determined using the validation set.

☐ The choice of k does not have a noteworthy effect on the predictions.

☐ Too large of a k leads to overfitting.

☐ The algorithm is an instance of lazy learning.

☐ The training phase is computationally expensive and time consuming.

☐ The prediction of unseen data points is computationally expensive and time consuming.

☐ k should be termined on the training set.