# K-Nearest neighbors
# Exercises

**Exercise 1.**

1. Describe in your own words the learning and classification process of the kNN-algorithm.
2. Note and explain at least one distance measure for numeric values and for categorical values.

**Exercise 2.**

1. Given the following dataset.

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

a) From looking at the data, describe a classification task that suggests itself.

b) Compute for $k = 4$ the nearest neighbors for the data instance (Sunny, Mild, Normal, Weak) using Overlap as distance measure and classify it accordingly.

c) Compute for $k = 4$ the neighbors of the instance (Sunny, Cool, High, Strong). Which problem does occur? How can it be solved?

**Exercise 3.**

Create a Jupyter Notebook and

1. load the IRIS dataset
2. write a short documentation of the dataset
3. select only features 2 and 3 (counting from 0) as features, thus creating a 2-dimensional feature space
4. use the random seed 1 if you use random
5. create a classification experiment, preprocess the data, split into training (60%) and test data (40%), use random sampling
6. train differently configured instances of the kNN-algorithm and evaluate them on the test data (baseline comparison, comparison between different hyperparameters)
7. think about feature scaling (describe what happens when you use or omit scaling)
8. try splitting the dataset with and without stratification (and random sampling), explain the resulting difference
9. use a diagram to indicate the relationship between certain parameters and the classification quality in terms of accuracy
10. plot the decision regions of the best classifier (use the python module `classification_viz` for that purpose.