

# Machine Learning

## Exercise 5: Dimensionality Reduction with PCA

Prof. Dr. Thomas Kopinski

November 29, 2022

### Abstract

In this exercise you will learn how to use *Principal Component Analysis* (PCA) to extract meaningful information from different datasets by reducing the dimension of the provided data.

### Task 1:

- Please download the Jupyter notebook for this exercise from [here](#) as it contains useful information, code snippets, help and some directions for the following tasks. Additionally please download the "wine.data" dataset from [here](#).
- Work through the examples in the notebook up to the section *Task 2*.
- Additional information about the *PCA* can be found in the course material.

### Task 2: Iris dataset

- In this task you will implement various classifiers to predict the species of Iris flowers, on the original, preprocessed data as well as on the PCA-transformed ones.
- Download the "IRIS.csv" dataset from [here](#).
- Load the dataset into a dataframe and get an overview about its content.
- Visualize different aspects of the dataset (e.g. class distribution, distribution of the individual features)
- Use the seaborn package to plot a correlation matrix of the dataframe (`seaborn.heatmap()`)
- Preprocess the data.
- Choose several suitable classifiers, train the models and compare and visualize the results for untransformed and transformed data.

### Task 3: Credit dataset

- This time you have to deal with the "credit" dataset which can be found [here](#).
- Load the dataset into a dataframe and take a look at it. Which are the feature columns, which column the target? Any correlations?
- Try to deal with the missing data (replace, delete, etc.)
- Encode the variable features.
- Use PCA on the cleaned dataframe. What are your findings? How many components would you choose?
- Train, test, evaluate and plot various combinations of models and PCA-transformations.