Lecture "Digital Signal Processing"

Prof. Dr. Dietrich Klakow, Summer Term 2021

Assignment 9

Submission deadline: 21 June 2021, 23:59

Submission Instructions:

You have one week to solve the assignments.
The code should be well structured and commented. Do not use any Matlab-Toolbox or Python external libraries if it is not mentioned that you can use them.

- You are allowed and encouraged to hand in your solutions in a group of two students.
- There are two parts in this assignment: theoretical part and practical part.
- The practical part that is solved with Python should be submitted as an ipynb file (Jupyter Notebook or Google Colab), where every function is written in a separate block.
- The theoretical part can either be written by hand and scanned, or typed with LaTeX in the text / markdown area of ipynb file.
- Submission of both parts should be done via Microsoft Teams by one of the group members.
- Submission should be named as: Ex09_matriculationnumber1_matriculationnumber2.zip

The submission should contain the following files:

- file "README" that contains an information on all team members:
  name, matriculation number, and email address.
- code files
- file "answers.pdf" which contains answers to the questions appearing in the exercise sheet. *Note: If you use ipynb file, you don't have to submit "answers.pdf". You can embed your scanned copy or write your answers in the text / markdown area.*

1 (5P)PCA

1. Consider the function $f(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{N} \lambda_j}$, where λ_1,....,λ_N are the eigenvalues, k is the number of first Principal Components. Let N be the total number of features, what is the maximum value of $f(k)$ and when does it occur? Explain.
2. What are the three main assumptions in applying PCA?
3. What will happen when eigenvalues are roughly equal?
4. Does PCA work well for the following type of data? Justify.

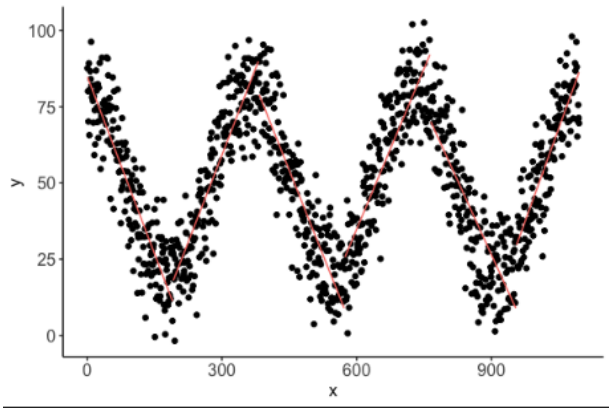

Figure 1

5. Say True or False with a justification:
   PCA will consider the correlation of the features with the label for dimensionality reduction.

2 (5 P)PCA Implementation

The goal of this exercise is to do dimensionality reduction using PCA. Allowed libraries: numpy, matplotlib.pyplot, pandas

2.1 (1P) Subtask

- Read the **Dataset**. You can use pandas to directly download the dataset in the code.
- Standardize the data

2.2 (2P) Subtask

1. Write a function, which takes as input the cumulative sum of explained variances(range = [0:1]) and returns the number of principal components to be selected.
2. Write a function, which takes as input the number of first principal components and returns the cumulative sum of explained variances in percentage.

2.3 (1P) Subtask

1. Use the function of 2.2.2 to obtain the cumulative sum of explained variance when we select 3 principal components.

2. Use the function of 2.2.1 to obtain the first principal components such that it reduces the data dimensionality with cumulative sum of explained variance = 0.5.

2.4 (1P) Subtask

Plot the result of 2.3.2 with different colors for different classes? What are your observations?