University of Stuttgart | ANALYTIC COMPUTING

Foundations of Machine Learning - Exercise (SS 25)

Assignment 2: PCA, Evaluation, K-means

Arvindh Arunbabu
arvindh.arun@ki.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ki.uni-stuttgart.de

Cosimo Gregucci
cosimo.gregucci@ki.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ki.uni-stuttgart.de

Farane Jalali
farane.jalali-farahani@ki.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ki.uni-stuttgart.de

Jiaxin Pan
jiaxin.pan@ki.uni-stuttgart.de

Tim Schneider
tim.schneider@ki.uni-stuttgart.de

Jingcheng Wu
jingcheng.wu@ki.uni-stuttgart.de

Daniel Frank
daniel.frank@ki.uni-stuttgart.de

Submit your theoretical solution in ILIAS as a single PDF file.[1] Make sure to list the full names of all participants, matriculation number, study program, and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g., PPTX files). Submit your programming task in ILIAS as a single Jupyter notebook. If you have any questions, feel free to ask them in the exercise forum in ILIAS.

Submission is open until Monday, 5th of May, 12:00 noon.

[1] Your drawing software probably allows exporting as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

Task 1: Principal Component Analysis

You and your classmates are curious about how students study for exams. You surveyed four friends and recorded how many hours they spent last week doing:

1. Reading textbooks
2. Watching lecture videos
3. Doing practice problems

The table below summarizes the data you collected: You want to find out whether there is a single factor that

Table 1 Weekly study hours by activity for four students

Student	Reading	Videos	Practice
A	2	6	5
B	4	5	3
C	6	8	6
D	5	7	8

best represents how students study. So, you decide to apply **Principal Component Analysis (PCA)**. PCA will help you identify whether the students share a common study pattern and whether it can be captured using just one new feature to summarize their overall study behavior.

1. **Task** In the lecture, you have seen how to find the first axis via the reconstruction error. It was also shown that this first axis refers to the Eigenvector of the covariance matrix. In this task, you are asked to directly compute the eigenvector and their corresponding eigenvalues with respect to the dataset in Table 1. Please do the preparation steps by hand and use NumPy or other libraries to compute the eigenvalues and eigenvectors.
2. **Task** What information do the eigenvalues carry, and how would you choose the projection based on the eigenvalues? Explain your answers.
3. **Task** Select the components such that the projected data explains at least 70% of the variance.
4. **Task** Based on the components you selected in the previous subtask, perform a projection on the first principal component (you can use Python to do the computation). What do these projected values tell you about the students' study patterns?

Task 2: Clustering and Evaluation

In Table 2 you are given a two dimensional dataset. In this task you are asked to cluster the given dataset into two classes by using the K-means algorithm presented in the lecture.

Table 2 Dataset with two features X and Y.

Point	X	Y
A	1	1
B	1	4
C	2	1
D	5	4
E	6	5
F	6	3

1. **Task** Before running the algorithm, how would you initialize the centers?

2. **Task** Using the centers selected in the previous subtask, run the K-means algorithm. If the algorithm does not converge within the first few iterations, stop and select a different initialization. Provide a detailed, step-by-step solution.

3. **Task** Compute the Dunn Index to evaluate the quality of your clustering. The Dunn Index is defined as the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. From a distance perspective, explain what a higher Dunn Index indicates about the clustering. What are some limitations of using the Dunn Index?

Task 3: K-means

Follow the instructions in the Jupyter notebook.