

T1 - Principal component analysis

After a labwork session, you have 7 days to complete the exercises:

- Write a report of at least 2 pages to describe your work (figure and table), discussion, and analysis of labwork.
- Submit in PDF format to Google Classroom (link in Moodle).
- Source code is optional in the report.

Note:

- You can work in a pair to complete all the assignments.
- Remember to provide YOUR student id and full name in the report.

1 Study the dataset

- Choose 2 datasets from any source (UCI Repository, Kaggle, etc.) with more than 3 dimensions.
- For each dataset, determine which feature is discrete or continuous? quantitative or qualitative? numerical or categorical? Explain.
- Is there any label in the datasets? Explain.
- Calculate mean, variance, covariance, correlation of the selected datasets. How do you calculate these measures for categorical features (if any)?
- Find the most correlated couple of features of each dataset. Comment on the results.
- Is there any missing data? How do you handle this issue?

2 PCA

- Apply PCA on the two selected datasets. Describe how you select the principal components. Is there any difference while using principal components with highest and lowest values?
- How much of the new variance in the data is explained by the new principal components?
- Vary the number of used principal components, analyze and comment on the obtained results.
- Visualize data distribution in 2D.