

ITDS342: Class Project

Group Project: 1–3 members

Deadline: Nov 22, 2024, 18.00

"Implementing and Analyzing Principal Component Analysis (PCA) for Dimensionality Reduction"

Objective:

Students will implement PCA from scratch using Python and analyze its effectiveness in reducing the dimensions of a dataset while retaining the most critical information.

Tasks:

- Data Collection: Students choose a dataset with multiple features (e.g., Iris dataset, Wine Quality dataset, or a custom dataset with many features).
- Standardization: Standardize the dataset to have a mean of zero and variance of one.
- Covariance Matrix Calculation: Compute the covariance matrix of the standardized data.
- Eigenvalues and Eigenvectors: Calculate the eigenvalues and eigenvectors of the covariance matrix.
- Principal Components: Select the top k principal components and transform the dataset accordingly.
- Explained Variance Analysis: Plot the explained variance ratio to determine how much variance each principal component captures.
- Visualization: Visualize the dataset in reduced dimensions (2D or 3D) and interpret the results.
- Comparison with Sklearn's PCA: Compare their implementation with the PCA provided by sklearn to validate the results.

Submission

1. Project Report (PDF)

Introduction:

Brief description of Principal Component Analysis (PCA) and its relevance in data science.

Explain the objective and goals of the project.

Mathematical Background: Detailed explanation of the mathematical concepts behind PCA (covariance matrix, eigenvalues, eigenvectors, variance).

Dataset Used

Methodology:

Step-by-step explanation of how PCA was implemented, including preprocessing, covariance matrix calculation, eigen decomposition, and dimensionality reduction.

Results and Analysis:

Explained variance ratio analysis.

Interpretation of the reduced dimension data.

Comparison with the sklearn implementation and insights from the results.

Conclusion:

Summary of findings, challenges faced, and possible improvements.

2. Code Files (Jupyter Notebook or Python Script)