Introduction to Data Science

22KDL

Lab03 - Principal Component Analysis

Deadline: **23h59 - 04/05/2024**
TA contact info: hduc.lee@gmail.com

Objective: In this assignment, you will delve into the practical application of Principal Component Analysis (PCA) for dimensionality reduction and feature extraction. You will use a real-world dataset to implement PCA and gain insights into its effectiveness in reducing the dimensionality of high-dimensional data while preserving its essential characteristics.

Dataset: https://www.kaggle.com/datasets/varunraskar/cancer-regression/data
Requirements: Complete the following tasks with the provided dataset.
Data Preprocessing:

- Perform necessary data preprocessing steps such as handling missing values, scaling numerical features, and encoding categorical variables if applicable.

Implement PCA:

- Utilize scikit-learn to implement PCA on the preprocessed dataset.
- Experiment with different numbers of principal components.
- Visualize the explained variance ratio to understand the amount of variance captured by each principal component.

Dimensionality Reduction:

- Apply PCA to reduce the dimensionality of the dataset and find a suitable number of dimensions to keep the information.
- Compare the performance of machine learning models (e.g., linear regression or others) on both the original and reduced-dimensional datasets using appropriate evaluation metrics for the problem.

Interpretation and Analysis:

- Analyze the results obtained from the PCA transformation and dimensionality reduction.
- Interpret the principal components and their corresponding eigenvectors to understand the underlying structure of the data.
- Discuss any trade-offs observed between dimensionality reduction and model performance.

Report and Conclusion:

- Prepare a concise report inside your notebook summarizing your findings, methodology, and insights gained from the assignment.
- Create visualizations (e.g., scatter plots, bar charts) to illustrate key observations.
- Highlighting the importance of PCA in the context of dimensionality reduction and its implications for real-world data analysis.

Encouragement for Experimentation:

- Encourage collaboration and discussion (BUT NOT **Plagiarism**) among students to share insights and learn from each other's approaches.
- Encourage students to leverage platforms like Google Colab/ Kaggle Notebook to experiment with the implementation. Google Colab provides free access to resources (CPU/GPU), facilitating faster experimentation.
- Encourage students to seek help from teaching assistants during lab sessions if they encounter difficulties.

Plagiarism Warning:

- Students are strictly prohibited from copying or reproducing the solution code from their peers. Each submission must be the individual work of the student. **Any instances of plagiarism or copying will result in a grade of 0 points for the assignment.**

Submission Guidelines:

- Jupyter Notebook containing:
  - Python code
  - Analysis
  - Visualizations
  - Report summarizing your findings and insights.
- Please send me your work before the due date.
- You can download the jupyter-notebook file (*.ipynb) by the following steps:
  - *File -> Download -> Download .ipynb*
- Name your notebook by the following pattern (same for Google Colab notebook's title): DS2024_Lab<LabID>_<StudentID>_<StudentName>.ipynb.
  - Example: DS2024_Lab01_21280075_NguyenVanA.ipynb
- The code results have to be printed out in the notebook.
- Include comments explaining key parts of the code if possible.
- Submit the notebook at: https://forms.gle/DwcBKS2a9n85LNX69

There is **NO** acceptance for **cheating** or **copying**.