CS F320 Foundations of Data Science

Assignment-2

Submission Time & Date: 13:00hrs on 30th Nov 2023

Instructions

- This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.
- This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib. You are not supposed to use libraries like scikit-learn for the regression models. Jupyter Notebook/Google Colab can be used. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.
- All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A1_<id-of-first-member>_ <id-of-second- member>_ <id-of-third-member> before submission.
- Submit the zip file on CMS on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held later which shall be conveyed to you. All group members are expected to be present during the demo.

Assignment 2-A

Implementing PCA from Scratch and Applying it to Car Data

The objective of this assignment is to gain a deeper understanding of Principal Component Analysis (PCA) by implementing it using NumPy and Pandas libraries and applying it to the 'Car_data' dataset to reduce dimensionality and visualize principal components.

Steps:

1. **Data Understanding and Representation:**
   - Import the 'Car_data' dataset and understand the features present.
   - Represent the features in matrix format, where each row represents an observation (car) and each column represents a feature.

2. **Implementing PCA using Covariance Matrices:**
   - Calculate the mean of each feature in the dataset.
   - Center the dataset by subtracting the mean from each feature.
   - Compute the covariance matrix of the centered dataset.

3. **Eigenvalue-Eigenvector Equation:**
   - Formulate and solve the eigenvalue-eigenvector equation using NumPy's eigenvalue and eigenvector functions on the covariance matrix obtained in the previous step.

4. **Solving for Principal Components:**
   - Implement a method to find the solutions to the eigenvalue-eigenvector equation using NumPy functions.
   - Select the top k eigenvectors corresponding to the largest k eigenvalues to represent the principal components.

5. **Sequential Variance Increase:**
   - Calculate the total variance covered by the principal components.
   - Observe and analyze the sequential cumulative increase in total variance explained as more principal components are considered.

6. **Visualization using Pair Plots:**
   - Plot pair plots of the original features.
   - Project the principal components onto these pair plots and visualize them as vectors, showing their directions and importance.

7. **Conclusion and Interpretation:**
   - Interpret the results obtained from the PCA analysis.
   - Discuss the significance of principal components in capturing variance and reducing dimensionality.
   - Analyze the effectiveness of dimensionality reduction and the insights gained from the visualizations.

8. **Documentation and Presentation:**
   - Provide a well-documented report detailing the steps performed, mathematical derivations, and code implementation.
   - Include visual representations, such as pair plots and principal component projections, to support the analysis.
   - Prepare a clear and concise presentation summarizing the key findings for presentation purposes.

Data Set: Car.csv

Assignment 2-B

PCA Analysis and Determining Optimal Number of Components

The objective of this assignment is to conduct Principal Component Analysis (PCA) on the 'Hitters.csv' dataset, determine the optimal number of components for efficient prediction using Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), and test the most efficient model.

Steps:

1. **Exploratory Data Analysis (EDA):**
   - Load the 'Hitters.csv' dataset and perform EDA to understand its structure, features, and relationships.
   - Handle NULL values and eliminate any unwanted columns or data inconsistencies.

2. **PCA Analysis:**
   - Apply PCA on the cleaned dataset to reduce dimensionality.
   - Determine the number of principal components required for efficient prediction. Try a range of component numbers.

3. **Model Training and MSE/RMSE Calculation:**
   - Split the dataset into training and testing sets.
   - For each number of principal components considered, build a regression model using those components.
   - Calculate the MSE or RMSE for each model on the test set to assess prediction efficiency.

4. **Plotting Number of Components vs RMSE:**
   - Plot a graph of the number of components against RMSE to visualize the relationship.
   - Identify the point where RMSE reaches a minimum or starts stabilizing, indicating an efficient number of components.

5. **Testing the Most Efficient Model:**
 - Select the model with the optimal number of components based on the graph.
   - Test the selected model by predicting a specific point and providing its predicted y value (y_pred).

6. **Conclusion and Analysis:**
   - Interpret the graph of the number of components vs RMSE to identify the most efficient model.
   - Discuss the significance of selecting an appropriate number of components for prediction efficiency.
   - Analyze the predicted value (y_pred) from the chosen model and its significance.

7. **Documentation and Presentation:**
   - Present a comprehensive report detailing the steps taken, model implementations, results, and analysis.
   - Include the graph illustrating the number of components vs RMSE and the interpretation of the findings.
   - Prepare a clear and concise presentation summarizing the key findings for a broader audience.

Dataset: Hitters.csv