Project 2: Exploring Dimensionality Reduction Techniques with PCA, Kernel PCA, Sparse PCA, and PCR

Due: Sep. 26, 2024, 11:59 pm

Datasets:

We will use four datasets for this project, each focusing on different data structures:

1. Wine Dataset (for PCA): This dataset contains chemical analysis results of wines grown in the same region in Italy but derived from three different cultivars. Download the data from the Canvas.

2. Digits Dataset (for Kernel PCA): This dataset contains images of handwritten digits (0-9) and is useful for exploring kernel-based dimensionality reduction methods. Available directly from sklearn via `load_digits()`.

3. Breast Cancer Dataset (for Sparse PCA): This dataset contains features computed from digitized images of fine needle aspirate (FNA) of breast mass tissue. Download the data from the Canvas.

4. Boston Housing Dataset (for PCR): This dataset contains housing data in Boston. Download it from the Canvas.

Part 1: Principal Component Analysis (PCA) on the Wine Dataset

Task 1.1: Data Preparation

• Question 1: Load the Wine dataset and normalize the features. Why is feature normalization important when applying PCA?

Task 1.2: Apply PCA

• Question 2: Apply PCA to reduce the dimensionality of the Wine dataset. Keep enough principal components to explain 90% of the variance. How many principal components are required to capture 90% of the variance?

• Question 3: Plot the explained variance ratio and the cumulative explained variance. What does this tell you about the dataset?

Task 1.3: Visualize and Interpret Results

• Question 4: Plot the data in the first two principal component spaces. Can you see clear separation between the three wine cultivars?

• Question 5: Reconstruct the original dataset from the reduced PCA components. What information is lost when reducing the dimensionality?

Part 2: Kernel PCA on the Digits Dataset

Task 2.1: Data Preparation

• Question 1: Load the Digits dataset using sklearn.datasets.load_digits(). Visualize some of the digits using matplotlib.

Task 2.2: Apply Kernel PCA with RBF Kernel

• Question 2: Apply Kernel PCA using an RBF kernel to reduce the dimensionality of the dataset.

• Question 3: Visualize the data in the 2D space of the first two kernel principal components. What do you observe?

Task 2.3: Compare with Linear PCA

• Question 4: Apply standard PCA to reduce the data to 2 components. Plot the results and compare them with Kernel PCA results.

Task 2.4: Interpretation and Evaluation

• Question 5: Reconstruct the original data from the reduced components. Compare the reconstruction errors for both methods.

Part 3: Sparse PCA on the Breast Cancer Dataset

Task 3.1: Data Preparation

• Question 1: Load and normalize the Breast Cancer dataset. Why might Sparse PCA be useful in this context?

Task 3.2: Apply Sparse PCA

• Question 2: Apply Sparse PCA to reduce the dimensionality of the dataset to 10 components.

Task 3.3: Visualize and Interpret Results

• Question 3: Visualize the first two sparse principal components. Are the components still interpretable?

• Question 4: Interpret the sparse components. Which features contribute the most to each component?

Task 3.4: Reconstruction and Comparison

• Question 5: Reconstruct the original dataset from the sparse principal components. Compare the reconstruction error with that of regular PCA.

Part 4: General Comparison of PCA, Kernel PCA, and Sparse PCA

Task 4.1: Summary Comparison

• Question 1: Compare the reconstruction errors for PCA, Kernel PCA, and Sparse PCA across the three datasets.

Task 4.2: Visualization Comparison

• Question 2: For each dataset, visualize the first two principal components (or kernel components) obtained through PCA, Kernel PCA, and Sparse PCA.

Task 4.3: Usefulness of Dimensionality Reduction

• Question 3: Based on your experiments, summarize the advantages and disadvantages of PCA, Kernel PCA, and Sparse PCA.

Part5: Principal Component Regression (PCR) on Boston Housing Dataset

Task 5.1: Data Preparation

Question 1: Load the Boston Housing dataset, normalize the features, and apply PCA. How many components are needed to explain 95% of the variance in the dataset?

Task 5.2: Apply PCR

Question 2: Use the principal components obtained from PCA to fit a linear regression model. Compare the performance of PCR with a regular linear regression model using the same dataset. Report the Mean Squared Error (MSE) and R^2 score for both models.

Try to interpret the linear model of PCR and the model of regular linear regression model.

The **Boston Housing dataset** contains data about housing in the Boston area, and it is often used for regression tasks.

The goal is to predict the **median value of owner-occupied homes (MEDV)** using features like per capita crime rate, pupil-teacher ratio, and property tax rate. MEDV is the label.

Features:

1. **CRIM**: per capita crime rate by town
2. **ZN**: proportion of residential land zoned for lots over 25,000 sq. ft.
3. **INDUS**: proportion of non-retail business acres per town
4. **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **NOX**: nitrogen oxides concentration (parts per 10 million)
6. **RM**: average number of rooms per dwelling
7. **AGE**: proportion of owner-occupied units built prior to 1940
8. **DIS**: weighted distances to five Boston employment centers
9. **RAD**: index of accessibility to radial highways
10. **TAX**: full-value property tax rate per $10,000
11. **PTRATIO**: pupil-teacher ratio by town
12. **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town
13. **LSTAT**: percentage of lower status of the population
14. **MEDV**: median value of owner-occupied homes in $1000's