CMPE 302 - Problem Set: Principal Components Analysis

Prepared by Anıl Demirel

Problem 1: PCA with Synthetic Dataset

You are given a synthetic dataset with 2 features and 10 data points:

$$X = \begin{bmatrix} 2.5 & 0.5 & 2.2 & 1.9 & 3.1 & 2.3 & 2.0 & 1.0 & 1.5 & 1.1 \\ 2.4 & 0.7 & 2.9 & 2.2 & 3.0 & 2.7 & 1.6 & 1.1 & 1.6 & 0.9 \end{bmatrix}$$

(a) Standardize the data by subtracting the mean of each feature.

(b) Compute the covariance matrix of the standardized data.

(c) Find the eigenvalues and eigenvectors of the covariance matrix.

(d) Project the data onto the eigenvector associated with the largest eigenvalue.

Use Python and NumPy to perform these steps. Include your code and the final projected dataset.

Hint: Use `np.cov`, `np.linalg.eig` and matrix multiplication.

Problem 2: PCA Visualization on the Iris Dataset

The Iris dataset contains 4 features: sepal length, sepal width, petal length, and petal width.

(a) Load the Iris dataset using `sklearn.datasets.load_iris()`.

(b) Perform PCA and reduce the dataset to 2 dimensions.

(c) Visualize the result using a 2D scatter plot colored by class label.

(d) Calculate and print the explained variance ratio for each principal component.

```
from sklearn.datasets import load_iris
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Load dataset
data = load_iris()
X = data.data
y = data.target

# PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Plot
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y)
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('PCA - Iris Dataset')
plt.show()

# Explained variance
print("Explained variance ratio:", pca.explained_variance_ratio_)
```