CISC 435 Machine Learning Fundamentals Assignment # 5
Mina Gabriel.

Problem 1. *PCA by Hand*

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving as much of the original data's variance as possible. This can help in simplifying the data and extracting its most important features.

Suppose we have a dataset with two features, X_1 and X_2, with three data points:

Data Point	x_1	x_2
X	2	3
Y	4	5
Z	6	7

Table 1: Dataset

reduce it to a single dimension using PCA.

Steps:

1. *Mean-Center the Data*
2. *Calculate the Covariance Matrix*

$$\text{Cov} = \begin{bmatrix} \text{Var}\left(X_1\right) & \text{Cov}\left(X_1, X_2\right) \\ \text{Cov}\left(X_2, X_1\right) & \text{Var}\left(X_2\right) \end{bmatrix}$$

3. *Calculate Eigenvectors and Eigenvalues*

   Solve the characteristic equation $\det(\text{Cov} - \lambda I) = 0$ *to find the eigenvalues* λ

   Next, for each eigenvalue, we solve the system of linear equations $(\text{Cov} - \lambda I)v = 0$ *to find the corresponding eigenvectors:*

4. *Project Data onto the Eigenvector*

Problem 2. *SVM by Hand*

Consider a binary classification problem with two features, X_1 and X_2, and a dataset consisting of the table below.

You are using a linear Support Vector Machine (SVM) with the objective of finding the optimal hyperplane that separates the classes while maximizing the margin. The equation of the hyperplane is

$$w_1 \cdot X_1 + w_2 \cdot X_2 + b = 0$$

.

Data Point	X_1	X_2	Class
A	2	3	-1
B	4	6	1
C	6	1	-1
D	7	3	1

Table 2: Dataset

After training the SVM, you obtain the following weights and bias:

$$w_1 = -0.5, w_2 = 0.5, b = -1$$

- ***Question A:*** *Calculate the margin of the hyperplane.*
- ***Question B:*** *Determine the equation of the decision boundary in slope-intercept form ($y = mx + c$).*

Problem 3. *PCA & SVM Coding Question*

In this question, you will be performing face classification using PCA in Python. Use the well-known "Labeled Faces in the Wild" (LFW) dataset available in the Scikit-Learn library, which contains labeled images of faces for classification.

Here is your code header, which should help you understand what you need to use:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import fetch_lfw_people
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import classification_report, confusion_matrix
```

Listing 1: Header

Steps:

1. *Load the LFW dataset and split it into training and testing sets*
2. *Perform Principal Component Analysis (PCA) for dimensionality reduction.*
3. *Train a Support Vector Machine (SVM) classifier on the reduced data.*
4. *Predict the test set and evaluate the classification using a classification report and confusion matrix.*
5. *Visualize a few predicted faces along with their titles. In Figure 1, you can see an example.*
6. *Provide a brief overview of the context and purpose of parameter tuning. Explain why parameter tuning is crucial in optimizing the performance of your trained model. Also, in detail explain your confusion matrix.*

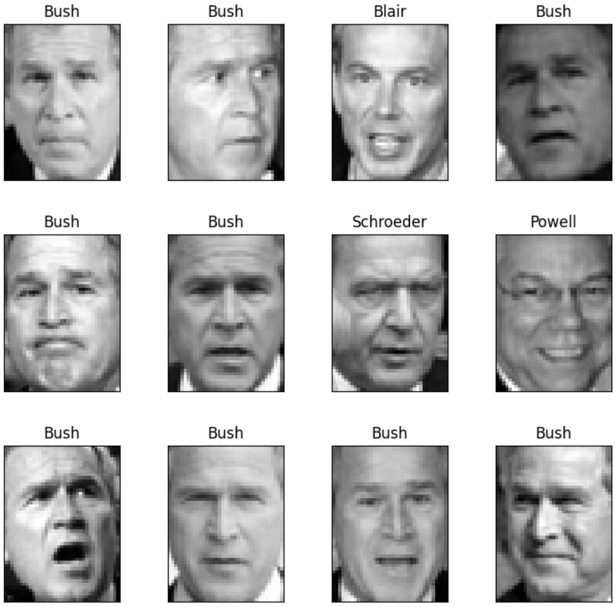

Figure 1: Predicated Faces Example.