Assignment 3: Machine Learning Pipeline and Hyperparameter Tuning

Purposes

- Downloading and importing datasets from online sources.
- Creating machine learning pipelines in `scikit-learn`.
- Dimensionality reduction using *Principal Component Analysis (PCA)* and *Linear Discriminant Analysis (LDA)*.
- Hyperparameter tuning using grid search.

We are going to build classifiers for the images in the MNIST and Fashion-MNIST datasets, and evaluate their performance.

MNIST Dataset. It is one of the most basic datasets for image processing and pattern recognition. The dataset of MNIST consists of totally 70,000 images of handwriting digits from 0 to 9. Each sample is a 28×28 grayscale image. 60,000 of them are used for training and 10,000 are used for testing. There are several ways for you to import the MNIST dataset using Python. One way is to first install the `tensorflow` and `keras` libraries and then import the images using their API. However, it is not necessary to do so. The dataset can be downloaded in its original format (IDX files), the files are attached to this assignment. You may use the package `idx2numpy` to import those files to NumPy arrays:

```
import idx2numpy
train_images = idx2numpy.convert_from_file(train_image_file)
train_labels = idx2numpy.convert_from_file(train_label_file)
```

such that `train_image_file` and `train_label_file` are the files that store the images and labels respectively for training. **Make sure that you include the correct paths for reaching those files.** Then the translated data are represented as NumPy arrays. For example, the object referenced by `train_images` is an array of $60000 \times 28 \times 28$, i.e., 60000 images of the size 28×28. You may also use the function `imshow` in `matplotlib` to plot an image:

```
import matplotlib.pyplot as plt
plt.imshow(train_images[0])  # The first image is plotted
plt.show()
```

We may use the above method to import all 60000 training samples and 10000 test samples.

Fashion-MNIST Dataset. It is a dataset of Zalando's article images and is more challenging to classify. The set also has 60000 training images and 10000 test images which have 10 labels. More details are given in the repository where the IDX files for the images are available.

Implementation Tasks

1. (2 points) Download and import the training and test images from MNIST and Fashion MNIST. The imported data are required to be kept in the NumPy array format. Then, perform the rest of the tasks for both datasets.
2. (2 points) Perform a data format transformation by flattening each image to a 1-D NumPy array. You may use the NumPy function `reshape`.
3. (11 points) Make a machine learning pipeline using `scikit-learn` to integrate the following components:
   - 3.1 (1 point) Standardize the (flattened) samples. **Notice that the preprocess mapping is computed based on the training data, but it will also be applied to the test data.**
   - 3.2 (4 points) Dimensionality reduction on the data. You are required to use two ways: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), to reduce the number of features of the data. The original dimensionality is 784, you are required to consider the reduced dimensionalities: 50, 100 and 200. Similar to the previous task, the reduction projections should be derived from the training set but are also used for compressing the test data.

3.3 (6 points) Build a Support Vector Classifier (SVC) with a kernel for classifying the compressed data. You should use the scikit-learn SVC class. You are required to consider the three kernels along with their hyperparameters:

- `'linear'` - Linear kernel, the only hyperparameter is `C`.
- `'rbf'` - Radial basis function kernel, the hyperparameters are `C` and `gamma`.
- `'poly'` - Polynomial kernel, the hyperparameters are `C`, `gamma` and `degree`.

You are required to define the search space (at least 8 values) for each hyperparameter and find the best setting using grid search for each kernel.

4. (5 points) Evaluate the performance of the best ML models (3 for each dataset regarding to the 3 scales in dimensionality reduction) you found and compare them using confusion matrices. From the comparisons, you are required to give at least 2 insightful observations for the performance of different kernels, dimensions, and 2 insightful observations for the classifiability of the two datasets along with discussions. Create tables to compare all 48 experimental results. The details are given in the report requirement.

Report

You are required to write a report for this assignment. It should include the following parts:

- A brief description of your implementation.
- The best settings you found using grid search and explain how the search space is specified for the hyperparameters.
- A summary of your experiments. This is a part of Task 4. You are required to show the confusion matrices, the observations and the discussions. Notice that your experimental results (accuracy) should cover: (1) two datasets individually, (2) the comparison between PCA and LDA for the 3 scales, (3) the comparison of the 3 kernels with their best settings. As a composition of them, you need to include $2 \times (2 \times 3) \times 3 = 36$ results. You are required to create two tables (for MNIST and Fashion-MNIST respectively) each of which compares the 18 results.