MP4

November 5, 2025

1 AI 221: Machine Exercise 4 - Handwritten Digits Analysis

This exercise focuses on dimensionality reduction and classification techniques applied to the 8x8 handwritten digits dataset from sklearn.

2 Dataset Preparation

This section contains the loading the handwritten digits and showing random samples.

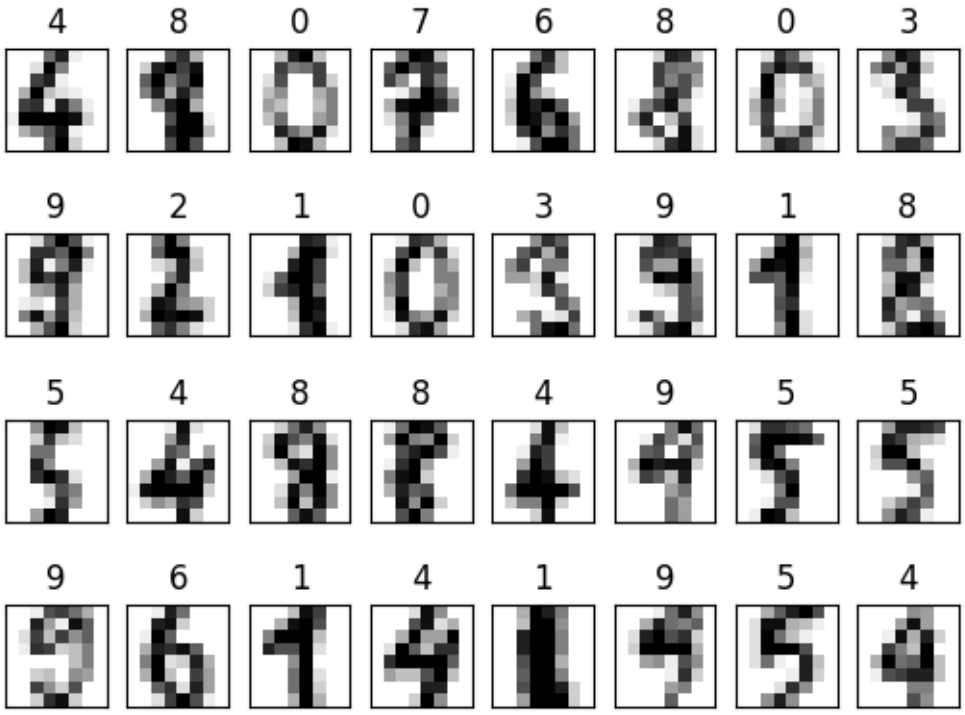
[1]:
```
from sklearn.datasets import load_digits
import matplotlib.pyplot as plt
import numpy as np
import random
X, y = load_digits(return_X_y=True)
print(X.shape)
random.seed(0)

rows, cols = 4, 8
fig, ax = plt.subplots(rows, cols, sharex='col', sharey='row',
                       subplot_kw=dict(xticks=[], yticks=[]))

for row in range(rows):
    for col in range(cols):
        n = np.random.randint(1796)+1    # show random samples
        im = ax[row, col].imshow(X[n].reshape((8,8)), cmap=plt.cm.binary)
        ax[row, col].set_title(y[n])
        im.set_clim(0, 16)
        #print(X[n])

plt.show()
```

```
(1797, 64)
```



3 Exercises

This section contains the core implementations of Machine Exercise 4 and is subdivided in to 3 subsections;

- Part A: Dimensionality Reduction Visualization
- Part B: Variance Analysis
- Part C: Classification Comparison

3.1 Part A: Dimensionality Reduction Visualization

3.1.1 Tasks:

1. Load and preprocess data:
   - Import digits dataset from sklearn.datasets
   - Apply StandardScaler normalization
2. Implement 6 dimensionality reduction techniques:
   - Local Linear Embedding (n_neighbors=200, random_state=0)
   - t-SNE (perplexity=50, random_state=0)
   - Isomap (n_neighbors=200)
   - Laplacian Eigenmap (n_neighbors=200)
   - Kernel PCA (kernel='rbf', gamma=0.01)
   - PCA
3. Visualization requirements:

- Project data to 2D for each method
- Color points by digit labels
- Compare clustering quality

3.1.2 Local Linear Embedding (n_neigbors = 200, random_state=0)

3.1.3 t-SNE (perplexity = 50, random_state = 0)

3.1.4 Isomap (n_neighbors=200)

3.1.5 Laplacian Eigenmap (n_neighbors=200)

3.1.6 Kernel PCA (kernel='rbf', gamma=0.01)

3.1.7 PCA

3.1.8 Results

Which of the methods produced clear clusters of data points?

3.2 Part B: Variance Analysis [10 pts]

3.2.1 Task Breakdown:

1. Generate CPV (Cumulative Percent Variance) plots for:
   - Kernel PCA
   - Standard PCA
2. Analysis requirements:
   - Determine number of components for 95% variance retention
   - Compare results between KPCA and PCA

3.2.2 Kernel PCA

3.2.3 Standard PCA

3.2.4 Results

This section contains relevant inf

3.3 Part C: Classification Comparison [60 pts]

3.3.1 Data Preparation:

- Split data: 70% training, 30% testing
- Ensure stratified sampling by class label

3.3.2 Classification Methods:

1. Method 1: Kernel PCA Pipeline [20 pts]
   - StandardScaler
   - Kernel PCA (kernel='sigmoid', n_components=40)
   - SVC (default parameters)
   - Report accuracy and F1-score
2. Method 2: LDA Pipeline [20 pts]

- StandardScaler
- LDA (n_components=9)
- SVC (default parameters)
- Report accuracy and F1-score
- Explain LDA component limitation

3. Method 3: Baseline SVC [20 pts]
    - StandardScaler
    - SVC (default parameters)
    - No dimensionality reduction
    - Report accuracy and F1-score

3.3.3 Analysis Requirements:

- Compare performance metrics
- Explain best performing method
- Discuss trade-offs

3.3.4 Kernel PCA Pipeline

- StandardScaler
- Kernel PCA (kernel='sigmoid', n_components=40)
- SVC (default parameters)
- Report accuracy and F1-score

3.3.5 LDA Pipeline [20 pts]

- StandardScaler
- LDA (n_components=9)
- SVC (default parameters)
- Report accuracy and F1-score
- Explain LDA component limitation

3.3.6 Baseline SVC [20 pts]

- StandardScaler
- SVC (default parameters)
- No dimensionality reduction
- Report accuracy and F1-score

3.4 Submission Guidelines

1. Format Requirements:
    - Submit as PDF file
    - Export from Jupyter Notebook
    - Highlight final answers clearly
2. Code Organization:
    - Clear section headers
    - Well-documented implementations
    - Proper visualization labels

3. Analysis Documentation:
    - Detailed explanations for each part
    - Comparative analysis between methods
    - Clear conclusions and insights
4. Submit through UVLE platform