**PCA and Kernel PCA Analysis on MNIST and Swiss Roll Datasets**

**Abstract**

This report presents a comprehensive analysis of dimensionality reduction techniques applied to the MNIST digit dataset and the Swiss Roll dataset. The methods utilized include Principal Component Analysis (PCA), Incremental PCA (IPCA), and Kernel PCA (kPCA) with various kernels. The results are explored through visualizations and performance evaluations using Logistic Regression classifiers.

**Introduction**

Dimensionality reduction is a crucial task in data processing and machine learning, particularly when dealing with high-dimensional datasets. Techniques like PCA, IPCA, and kPCA help to reduce dimensional complexity while preserving essential data characteristics.

This report will detail:

1. The loading and visualization of the MNIST dataset.

2. Application of PCA followed by Incremental PCA to compress the dataset dimensions.

3. Generation and analysis of the Swiss Roll dataset using kPCA with different kernels.

4. Application of Logistic Regression to classify the transformed dataset and tuning using GridSearchCV

**Experimentation**

**1. MNIST Dataset**

1.1 Data Loading and Preprocessing - The MNIST dataset, comprising 70,000 instances of handwritten digits (0-9), was loaded using the `arff` library and split into training (60,000 instances) and test sets (10,000 instances).

1.2 Visualization of Digits- A grid displaying the original 28x28 pixel images of the first 100 digits was created to confirm the dataset integrity.

1.3 PCA Implementation- PCA was applied on the training dataset to retrieve the first two principal components. The explained variance ratio for these components was printed:

Output- Explained variance ratio of the first two principal components: [0.09704664 0.07095924]

This indicates that these components capture approximately 15.82% of the overall variance.

1.4 Projections onto Hyperplanes- Projections of the first two principal components onto both 1D and 2D hyperplanes were plotted. The 2D scatter plot effectively separated the digits into clusters based on the extracted principal components.

1.5 Incremental PCA Reduction- IPCA was utilized to reduce the MNIST dataset down to 154 dimensions. Original and compressed images were displayed in a subplot format, showing the effectiveness of the compression.

**2. Swiss Roll Dataset**

2.1 Data Generation-A Swiss Roll dataset was generated with 3,000 samples, and target variable labels were prepared.

2.2 3D Visualization- A 3D plot of the Swiss Roll dataset illustrated the inherent structure of the data.

2.3 Kernel PCA- kPCA was evaluated using three kernels: Linear, RBF, and Sigmoid.

- \*\*Linear Kernel Output:\*\*

A colorful spiral with numbers and letters

Description automatically generated with medium confidence

- \*\*RBF Kernel Output:\*\*

A colorful circle with numbers

Description automatically generated with medium confidence

- \*\*Sigmoid Kernel Output:\*\*

A colorful circle with lines and numbers

Description automatically generated with medium confidence

Each plot depicted different separation capabilities, with the RBF kernel providing the best separation.

2.4 Logistic Regression Classification

Pipeline creation facilitated the application of kPCA followed by Logistic Regression. GridSearchCV was

employed to find optimal parameters:

output- Best parameters found by GridSearchCV: {'kpca\_\_gamma': 0.01, 'kpca\_\_kernel': 'rbf'}

This signifies the best-performing combination found during the hyperparameter search.

2.5 Best kPCA Transformation

The transformed dataset using the best kernel choice was visualized, showcasing the improved classification

potential enabled by effective dimensionality reduction.

**Conclusion**

The analysis demonstrated effective application of PCA and kernel-based methods in reducing dimensionality and enhancing data structure visibility. The conclusions drawn from the experiments indicate the importance of selecting appropriate dimensionality reduction techniques based on dataset characteristics and intended analysis goals.Future work may include further tuning of hyperparameters, exploring other classification algorithms, or investigating alternative kernels in kPCA for improved accuracy and robustness.

**References**

- scikit-learn documentation

- "Pattern Recognition and Machine Learning" by Christopher M. Bishop

- "An Introduction to Machine Learning" by Ethem Alpaydin