Dimensionality Reduction

Note: Project 2

Xiangsheng Chai
ECE Department
UFID: 44036693

***Abstract*—In this project, I applied different kinds of dimensionality reduction methods on the custom handwritten characters dataset, and evaluated the performances of each method. At the beginning, Recursive Feature Elimination (RFE) which is one of the feature selection's methods was applied and the pixels it selected was identified. The Principal Component Analysis (PCA) was used. By visualizing its performance, we can see the characteristic of PCA. Then, supervised dimensionality reduction method Fisher's Linear Discriminant Analysis (LDA) and t-SNE were applied to reduce the dataset into 2-dimensions and visualize the performance. Finally, three kinds of manifold learning algorithm were used. By comparing and observing the performance, we got some conclusion for them.**

Keywords—Custom handwritten characters dataset, Recursive Feature Elimination, Principal Component Analysis, Linear Discriminant Analysis, Manifold learning.

I. Introduction

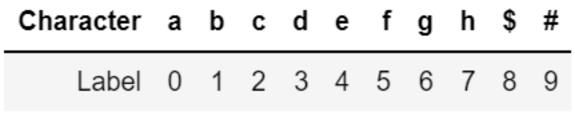
Dimensionality reduction is essential for today's machine learning works. Analysis with a large number of variables generally requires a large amount of memory and computation power. And the curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. In order to avoid these things, people use feature selection and feature extraction to implement the dimensionality.

In feature selection, there are three main approaches which includes filters, wrappers and embedded. Filters rank the original features according to their relationship with the problem (labels) and just select the top of them. Correlation and information gain are the most widespread criteria to rank the features. A wrapper evaluates a specific model sequentially using different potential subsets of features to get the subset that best works in the end. They are highly costly and have a high chance of overfitting, but also a high chance of success, on the other hand. RFE is a backward feature elimination wrapper. And embedded things like L1 and L2 regularization is made up of all the Machine Learning techniques that include feature selection during their training stage.

In feature extraction, starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. PCA and manifold learning are both belong to the feature extraction.

In this project, the dataset we used is custom handwritten characters dataset. It includes 10 handwritten characters (a, b, c, d, e, f, g, h, $, #).

TABLE I. 10 CLASSES AND ITS LABEL ENCODING



Character	a	b	c	d	e	f	g	h	$	#
Label	0	1	2	3	4	5	6	7	8	9

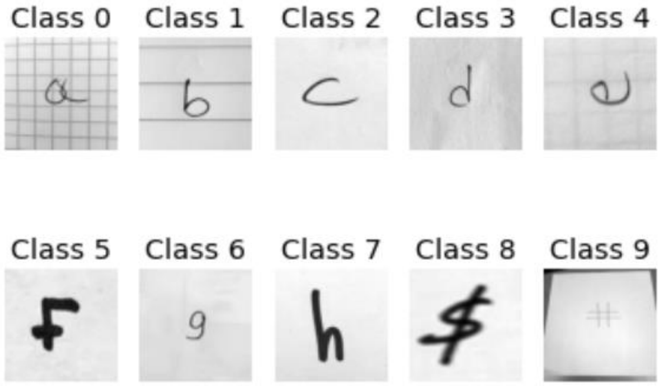


Fig. 1. Samples of 10 classes

This dataset is used to do the dimensionality reduction, and there are totally 6 questions for us to solve. The questions are listed below:

1. Implement Recursive Feature Elimination (RFE) to select the subset of features. Experiment with at least 2 different estimators.

2. Implement Principal Component Analysis (PCA) to select the number of components that explain at least 90% of the explained variance. Train a classifier on the original dataset and the reduced dataset.
3. Use Fisher's Linear Discriminant Analysis (LDA) and t-SNE to reduce the dataset to 2-dimensions and visualize it.
4. Implement at least 3 manifold learning algorithms for reducing the dimensionality of the feature space. Utilize the new lower-dimensional feature space to build a classifier.

II. Implementations

A. Preprocessing

In this project, the dataset we used contains 6720 pictures and 90000(300×300) pixels for each picture. 90000 dimensions is too large to do the training task, so the first thing we should do is to resample the original dataset from 90000 pixels to 2500(50×50) pixels. For dataset of pictures, minmax scaler is a better choice. After resampling, the minmax scaler was applied. Then the dataset is ready for training.

B. Feature selection task

In feature selection task, our goal is to use RFE with 2 different estimators. The classifier I choose is Logistic Regression with L2 regularization and Support Vector Machine (SVM) with linear kernel.

RFE is a classic sequential feature selection algorithm, which aims to reduce the dimensionality of the initial feature subspace with a minimum decay in performance of the classifier to improve upon computational efficiency. In certain cases, RFE can even improve the predictive power of the model if a model suffers from overfitting. The RFE algorithm can be outlined the in 4 simple steps:

- Initialize the algorithm with k=d, where d is the dimensionality of the full feature space $\boldsymbol{X}_d$.
- Determine the feature x^- that maximizes the criterion
- Remove the feature x^- from the feature set.
- Terminate if k equals the number of desired features, if not, go to step 2.

RFE with Logistic Regression and with SVM are used to build the model. In test, we visualized the pixels RFE selected and displayed its mask examples.

C. Feature extraction task

In feature extraction task, I will apply PCA, LDA, t-SNE and three kinds of manifold learning to reduce the dimensionality.

PCA is the most common strategy of dimensionality reduction. It is an unsupervised method. PCA uses a linear transformation to minimize the redundancy of the resulting transformed data. And it finds the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one. The basic steps of PCA are shown below. Consider the data X with N data points defined in a D-dimensional space, that is, X is a D×N matrix.

- Subtract the mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Compute the covariance matrix R_X (by definition, the covariance already subtracts the data's mean). This matrix is of size $D \times D$
- Compute eigenvectors and eigenvalues of the matrix R_X , and store the sorted eigenvectors (e_i) in decreasing eigenvalue (λ_i) order.
- Build the modal matrix $U = \left[e_1 \middle| e_2 \middle| \ldots \middle| e_D \right]$, where all the (unit-length) eigenvectors are stacked in columns, sorted by their respective eigenvalues.
- Apply the linear transformation: $\mathbf{y} = \mathbf{U}^T \mathbf{X}$. Here **y** is a matrix of size $M \times N$, where $M \leqslant D$.

Manifold Learning is Nonlinear dimensionality reduction. There are three main kinds of manifold learning------ Multi-Dimensional Scaling (MDS), Isometric Mapping (ISOMAP) and Locally Linear Embedding (LLE).

MDS finds a low-dimensional projection of the data such as to preserve the pairwise distances between data points, and involves finding the eigenvectors of the distance matrix. It is the generalization of PCA and ISOMAP. It preserves the global intrinsic relationship.

The goal of ISOMAP is to project the data to a lower-dimensional space using MDS, but where the distance or dissimilarities are defined in terms of the geodesic distances measured along the manifold.

LLE first computes the set of coefficients that best reconstructs each data point from its neighbors. These coefficients are arranged to be invariant to rotations, translations, and scaling of that data point and its neighbors, and hence they characterize the local geometrical properties of the neighborhood.

III. Experiment

A. Question 1

Logistic Regression classifier and SVM with linear kernel were applied to do dimensionality reduction by RFE. The reason of choosing linear kernel is that only linear kernel of SVM has attribute coefficient in scikit-learn model, and only classifier with coefficient attribute can be applied in RFE model.

The hyperparameters are set as default, and it took about 15 minutes to train. The selected pixels are shown below:

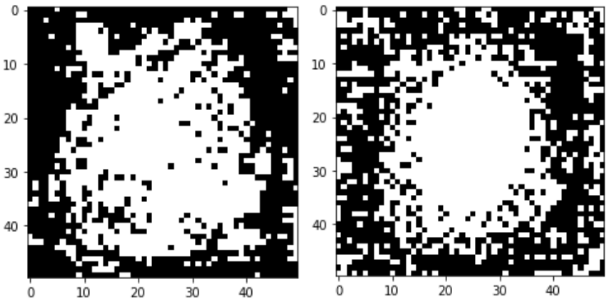


Fig. 2. Selected pixels of RFE with LR classifier (left) and RFE with SVM classifier (right).

We can observe that most of the reserved pixels are in the center of the picture. It is because central pixels contain more information about the handwritten characters, and the surrounding pixels of the picture are always blank and contain no information.

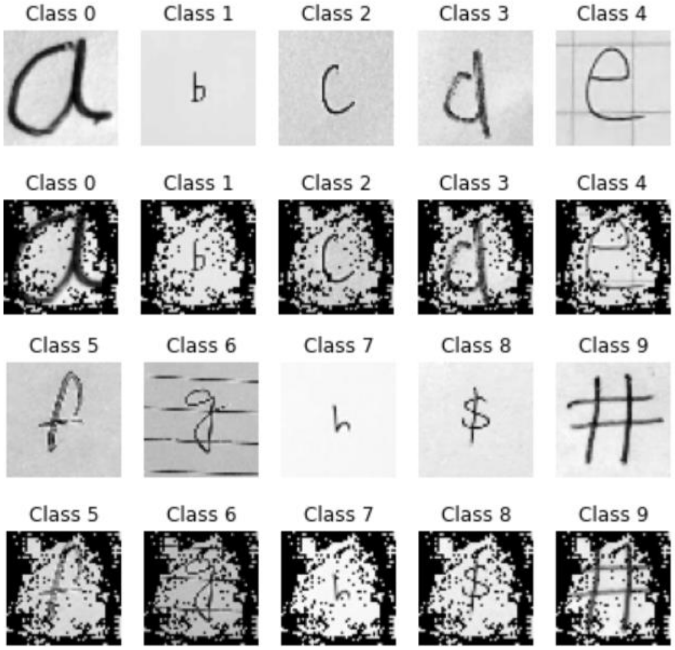


Fig. 3. Mask figures of each class。

B. *Question 2*

There are two classifiers applied in question 2. The classifier with PCA has 182 dimensions which keep 90% of the explained variance. In the following table, it shows the performance of two classifiers.

TABLE II. PERFORMANCE IN TRAINING SET

MODEL	**Accuracy**	**95% CI**
SVM WITH PCA	0.61	[0.452, 0.479]
SVM WITHOUT PCA	0.66	[0.436, 0.460]

TABLE III. PERFORMANCE IN TEST SET

MODEL	**Accuracy**	**95% CI**
SVM WITH PCA	0.47	[0.365, 0.402]
SVM WITHOUT PCA	0.47	[0.350, 0.392]

From tables above, we can see SVM without PCA has the higher accuracy in training dataset but the same accuracy as the SVM with PCA in test dataset. It means SVM without PCA is overfitting in training dataset. Compared the training time of two classifiers, SVM with PCA used 10.12 secs and SVM without PCA used 104.32 secs. In this case, we can see the advantage of dimensionality reduction ------ substantially decreases the time for training.

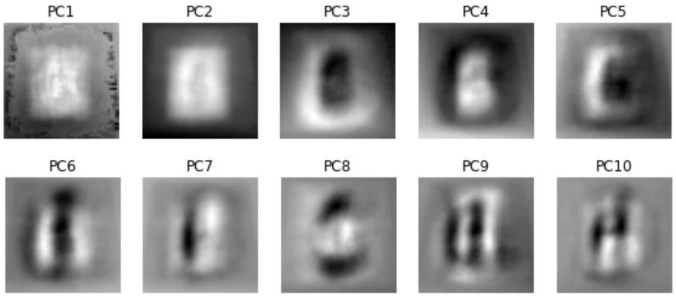


Fig. 4. Top ten eigenvectors

Top ten eigenvectors are shown above. They represent the 10 maximum variance direction in the data. They are also the outline handwritten characters.

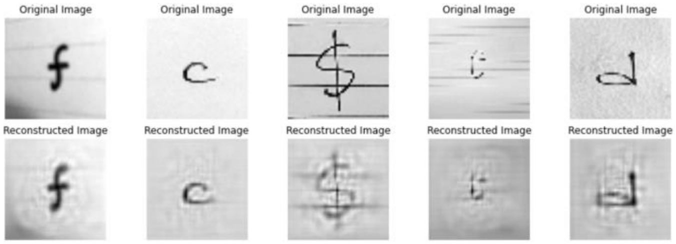


Fig. 5. Reconstructed Image from PCA projection

We can see it can successfully reconstruct the image from 182 dimensions PCA projection.

C. *Question 3*

In question 3, Fisher's Linear Discriminant Analysis (LDA) and t-SNE were applied to reduce the dataset to 2-dimensions and visualize it. The figures below are the visualization of training and test dataset.

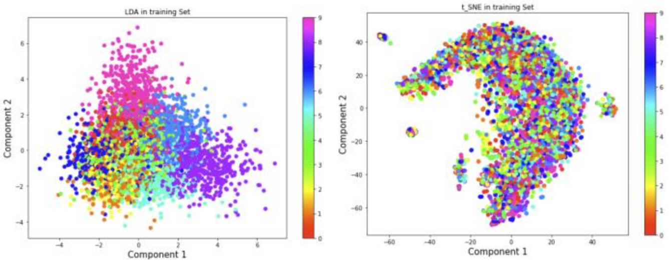


Fig. 6. Visualization in 2-dimensions of LDA (left) and t-SNE (right) in training dataset

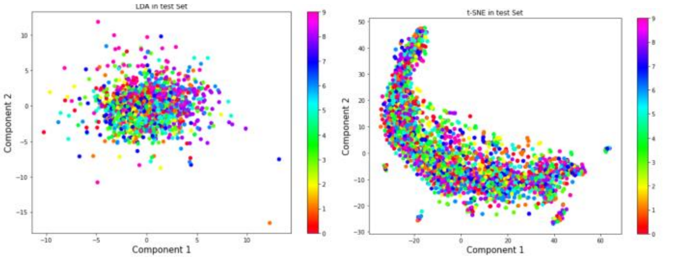


Fig. 7. Visualization in 2-dimensions of LDA (left) and t-SNE (right) in test dataset

In order to compare with the PCA, I also trained the 2-dimensional projection with PCA. The visualization is shown below:

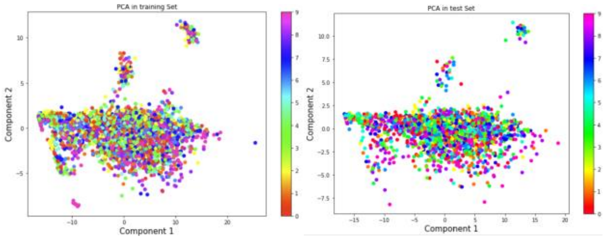


Fig. 8. Visualization in 2-dimensions of PCA in training (left) and test (right)

From the observation, we can see LDA which is a supervised dimensionality reduction strategy has the best effect to distinguish each class. Although t-SNE is the dimensionality reduction strategy which focuses on providing better visualization, it has the worse visualization than LDA because of unsupervised.

D. Question 4

In question 4, the classifier I chose is SVM with RBF kernel. And MDS, ISOMAP and LLE were applied to reduce the dimensionality respectively. The accuracy score of each manifold learning method is shown below.

TABLE IV. ACCURACY SCORE OF MANIFOLD LEARNING ALGORITHM

ALGORITHM	Number of Componen ts	ACCURACY IN TRAIN	ACCURACY IN TEST
MDS	60	0.105	0.096
ISOMAP	80	0.372	0.256
LLE	75	0.288	0.205

I tried different number of components in train set. And as the number of components increases, the accuracy score in test set doesn't increase a lot. Finally, I chose a relatively low dimension range between 50 to 80 with step 5 to do the grid-search. The number in table IV is the best number got by cross validation.

The training in MDS algorithm took nearly 9 hours, and the accuracy score is low. I think MDS with euclidean dissimilarity is not well for the handwritten characters dataset. Compared the three algorithm, ISOMAP has the best accuracy score in both training set and test set, and it took relatively shorter time on training. Hence, I will select ISOMAP as my dimensionality reduction algorithm.

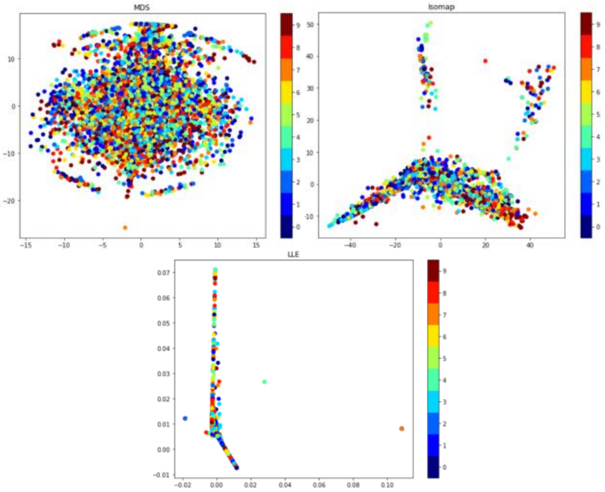


Fig. 9. Visualization in 2-dimensions of 3 manifold learning algorithm

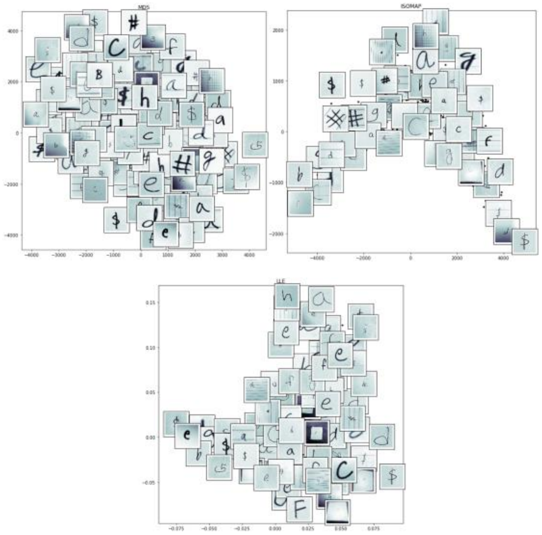


Fig. 9. Represent images for each algorithm

From the figure above, in ISOMAP, we can see the thickness of the handwritten characters become thick from the bottom to the top, and the direction of each character is changed from left to right. So, the thickness and direction of the characters are the first two components in ISOMAP. In LLE, thickness is also one of the components, but I don't see the other component clearly. In MDS, because of its low accuracy and the figure, the 2 dimensions should represent nothing.

IV. CONCLUNSION

From the results of four questions, we can see both feature selection and feature extraction have a good performance on dataset. The accuracy for each classifier is not well, this is because the classifiers I choose are not powerful enough to do the handwritten characters classification. According to the result, dimensionality reduction helps a lot on saving time and avoiding the influence of curse of dimensionality.