Dimension Reduction Visualization and Classification on Hand Writing Data

CHENG Wei
Department of Electronic and Computer Engineering

Objectives

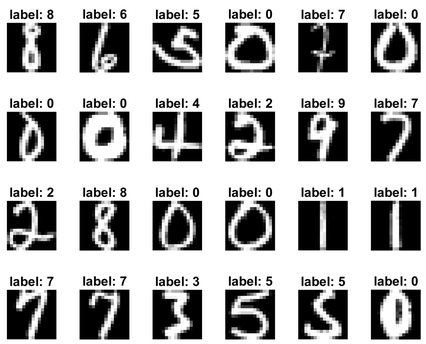
Dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables, which is key topic in computer vision, graphic and image processing. Modern researchers started with Principal Component Analysis (PCA) and Multi-Dimension Scaling (MDS) using Euclidean representations. To overcome the curse of dimensionality, an assumption that data concentrate around a low dimensional manifold in high dimensional spaces, leads to manifold learning or nonlinear dimensionality reduction, for example, ISOMAP and Lcoally Linear Embedding (LLE) methods.

In this project, we want to investigate the following critical points on Hand-Written Digits dataset

- How to perform dimension reduction and visualize the data distribution on principle components?
- How to classify and predict the Hand-Written Digits dataset using dimensionality reduction.
- What's the relationship between classification accuracy and dimension?

Dataset

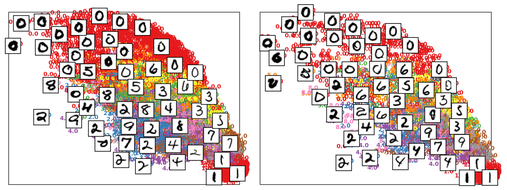
This Hand-Written Digits dataset consists of 7291 handwritten numbers from 0 to 9 offered by Prof. Tibshirani. Each digit is stored by row and is reshaped into a 16 by 16 gray matrix, with the first tuple represent the correct type of the digit. Proportion of Each type of digit is balanced. Some of the samples in Hand-Written Digits are displayed in the following figure.



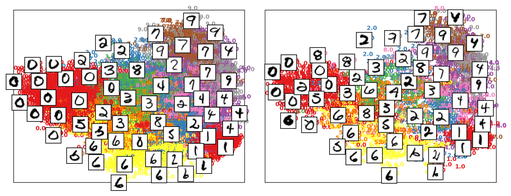
Visualization

We first perform the dimension reduction using following methods, PCA, ISOMAP, LLE, Modified LLE, LSTA, t-SNE and Diffusion map. Then, we visualize the sample distribution on the first two components.

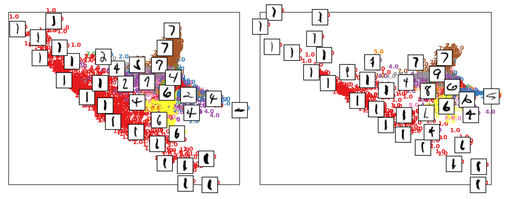
- PCA (train, test)



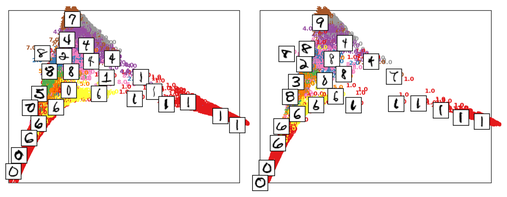
- ISOMAP (train, test)



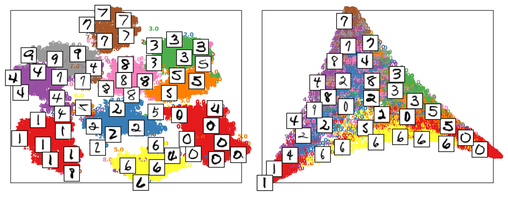
- LLE (train, test)



- Modified LLE (train, test)



- t-SNE, Diffusion map (both train)



Classification

- Classification with different embeddings and classifiers

In training process, we use the embedding models to fit to the training data, and using the transformed embedding on training data, then we use the training embedding and labels to train classifier. In regression phase, testing data are transformed using the same embedding, then these features are used to predict the classification label.

Noted that the ISOMAP, LLE, Modified LLE, LSTA are highly nonlinear, so SVM is not suitable with these kind of features.

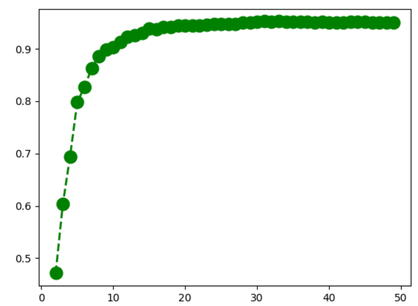
Method	0	1	2	3	4	5	6	7	8	8	Avg
PCA+SVM	0.99	0.95	0.93	0.90	0.93	0.94	0.94	0.94	0.91	0.95	0.94
PCA+kNN	0.98	0.96	0.85	0.89	0.84	0.84	0.93	0.90	0.84	0.90	0.90
ISO+kNN	0.98	0.97	0.78	0.91	0.75	0.81	0.92	0.91	0.80	0.89	0.88
LLE+kNN	0.96	0.97	0.80	0.90	0.83	0.57	0.90	0.84	0.68	0.81	0.85
MLLE+kNN	0.98	0.95	0.91	0.92	0.82	0.85	0.92	0.93	0.87	0.90	0.91
LSTA+kNN	0.97	0.95	0.88	0.88	0.85	0.84	0.91	0.85	0.83	0.86	0.89

Note that all the parameter in embedding method and regression method are the same, in my implementation, I used 20 dimensions and 60 nearest neighbours on building the embeddings and training the classifiers.

We can conclude that the PCA+SVM method acheived best averaging performance, while in manifold learning methods, modified LLE achieved the high averaging accuracy.

- Dimension on classification (PCA)

Finally, we investigated the relationship between dimension (number of features) and classification accuracy, we use the PCA+SVM method to dig this relationship plotted in the following figure. We can conclude that the accuracy increasing nearly stops when dimension is greater than 20.



Conclusion

Aming on solving the three objective problems, this project has the following contribution

- Visualized the dataset distribution on low dimension embeddings, we can easily distinguish the existence of certain clusters of hand-written digits, digits with the same type stay closer with each other.
- Estimated the classification accuracy using different embedding methods and its suitable classifiers, for example, linear classifier SVM is not suitable for manifold learning methods. The combinations all achieved acceptable classification accuracy, while among them PCA+SVM achieved the best performance in all combination and MLLE+kNN won the race within manifold learning methods.
- We also explored the relationship between dimension reduction with accuracy, we took PCA+SVM as an example and found that the accuarcy converges when dimension is greater than a certain level.

Project Page

This slide and the source code of implementation can be found in the following link

https://github.com/wchengad/course_2019_spring/tree/master/CSIC5011/Project1

References

- Hastie, Zip code digits datasets of "The Elements of Statistical Learning";
- Y. Yao, A Mathematical Introduction to Data Science, Chapter6 - Chapter7, 2017;
- J. Vanderplas, Comparison of Manifold Learning methods, replicated at 13, October, 2017;