

2017

# DIGIT RECOGNIZER PROPOSAL

Advisor:  
Professor Ray-Bing Chen

MACHINE LEARNING

UNDERGRADUATE STUDENT: STAT 107 吳竹祐 STAT 107 詹皓雯  
STAT 107 陳育婷 TCM 107 王軒至

NATIONAL CHENG KUNG UNIVERSITY

# **Table of Contents**

## **I Introduction**

- 1 Data Description**
- 2 The task we want to perform**

## **II Method Introduction**

- 1 Introducing PCA**
- 2 Introducing KNN**

## **III Combining PCA and KNN**

## **IV A Deeper Look**

- 1 R-Shiny**
- 2 14\*14 resolution**

## **V Conclusion**

- 1 Conclusion on original data**
- 2 Conclusion on degrade resolution**

## **VI Reference**

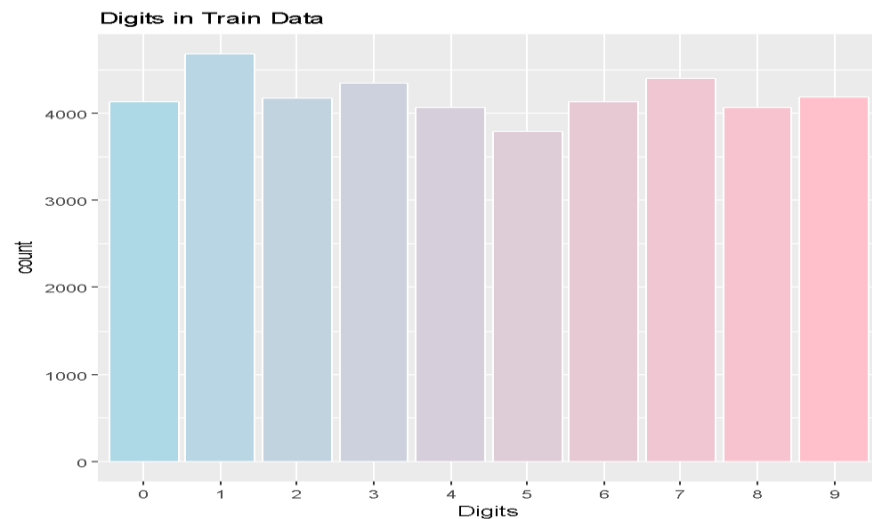
# I. Introduction

## 1. Data Description

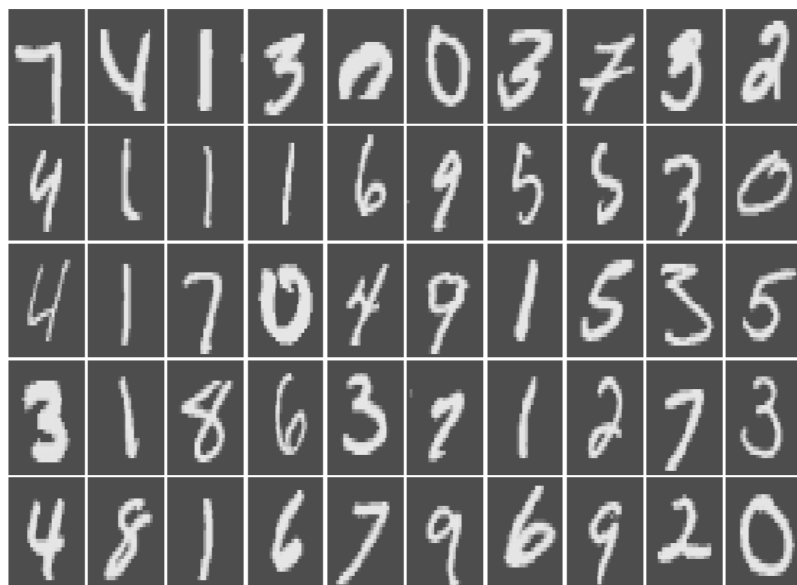
- (1) Training data: 785 variables, 42000 observations

	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5
41995	4	0	0	0	0	0	0
41996	0	0	0	0	0	0	0
41997	1	0	0	0	0	0	0
41998	7	0	0	0	0	0	0
41999	6	0	0	0	0	0	0
42000	9	0	0	0	0	0	0

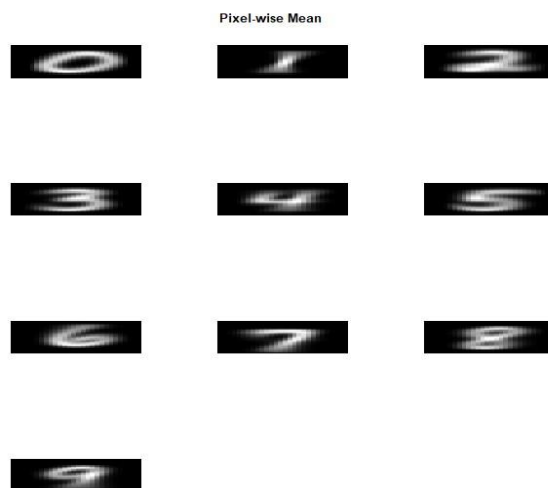
- (2) Numbers of observations of each digit: each digit have almost equal observations.



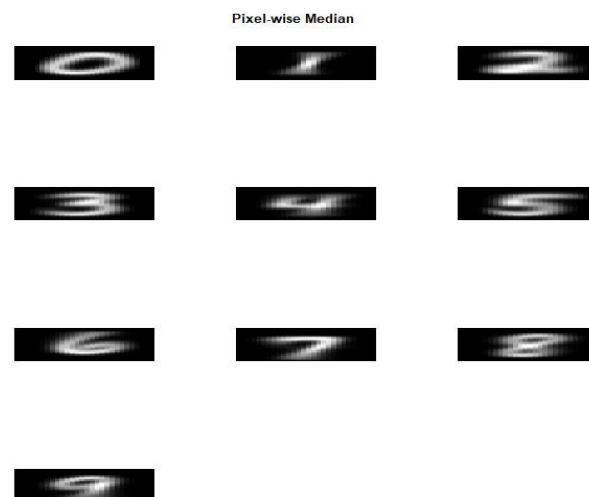
- (3) Visualizing a part of the data(50 digits)



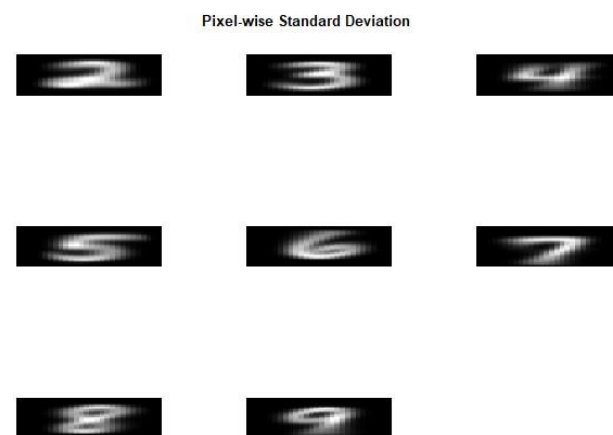
(4) Mean of each digit



(5) Median of each digit



(6) Standard Deviation of each digit, this will give us an idea of how each digit varies in their respective samples.



## 2. The Task We Want To Perform


- (1) To correctly identify digits from a dataset of tens of thousands of handwritten images.
- (2) Try to find the algorithm which works best and enhance the accuracy rate.

## II. Method Introduction

### 1. Introducing PCA

- (1) Principal Components Analysis
  - a. Commonly used to reduce the number of predictive variables while maintaining information as much as possible.
  - b. Computes new variables called principal components which are obtained as linear combinations of the original variables.
- (2) First 20 principal components already cover 95% of the variability of data. However; to enhance our accuracy rate, we can observe that using first 60 PCs capture over 99.5% of the variation. Adding ten principal components based on 60 pcs only explains less than 0.05% variation. Therefore, we then decide to extract 60 PCs for use in the KNN classifier. The following is the cumulative sum table.

number of PCs	Cumulative Proportion
10	0.8711
20	0.9523
30	0.9776
40	0.9887
50	0.9928
60	0.9968
70	0.9971
80	0.9978
90	0.9984
100	0.9988

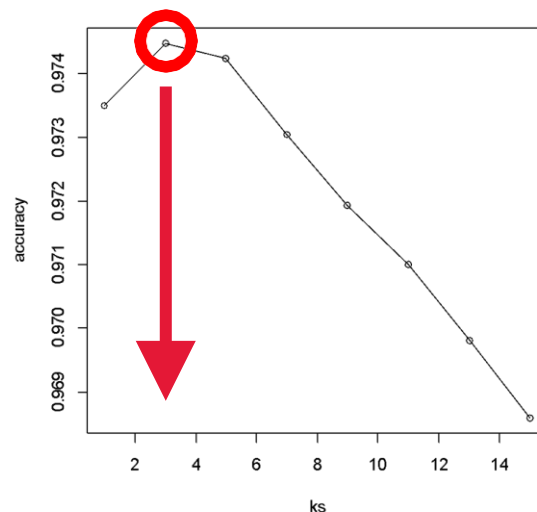


- (3) Advantages
  - a. Dimensionality reduction.
  - b. Completely nonparametric- do not be restricted by any distribution's assumption.
- (4) Disadvantages

- a. Cannot handle non-linear data-but there are alternatives which can deal with non-linear data such as kernel PCA. In our case, we do not have to worry about this problem.
- b. The requirement that the direction of each new component must be spread perpendicular to the previous one is overly stringent.

## 2. Introducing KNN

- (1) K-nearest neighbors algorithm
  - a. KNN is considered as a lazy learning algorithm that classifies data sets based on their similarity with neighbors.
  - b. “k” stands for number of data set items that are considered for the classification.
- (2) use cv to decide best k for KNN. (1/10(n=4200) for testing, 9/10(n=37800) for training)



We can achieve the highest accuracy rate when k=3.

- (3) Advantages
  - a. Robust to noisy training data
  - b. Effective if the training data is large
  - c. Nonparametric approach
  - d. Simple, easy to understand but very powerful
- (4) Disadvantages
  - a. Need to determine the value of parameter k
  - b. It is not clear which type of distance to use. In our case, we use the statistical distance.
  - c. Computation cost is quite high

### III. Combining PCA and KNN - two stage analyzation

#### 1. Procedure

ORIGINAL DATA → PCA → KNN → GOAL

#### 2. Advantage of PCA based KNN (PC-KNN)

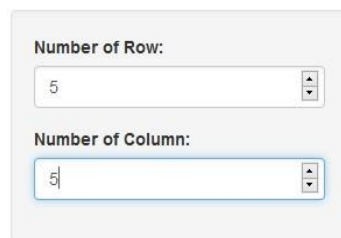
- (1) They are free of distribution assumption.
- (2) When dealing with high dimensional data, PC-KNN often performs better than only using KNN.

### IV.A Deeper Look

#### 1. R-Shiny

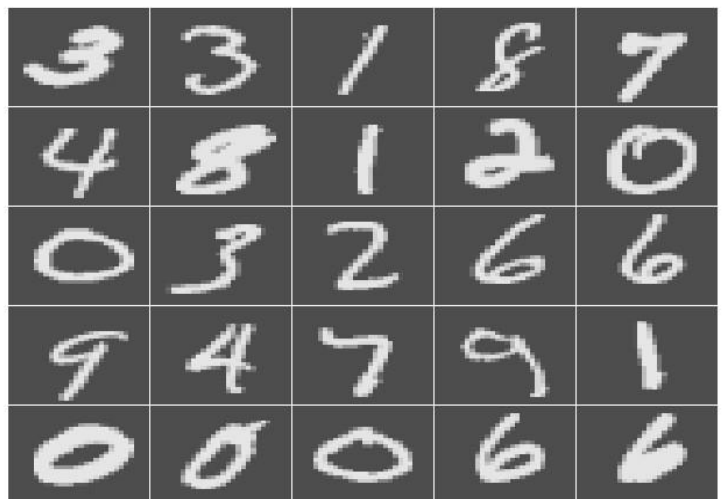
In R-shiny, we establish numeric function, user can just click the button to input the number of row and column to visualize digits.

Number of Digits to view



Number of Row:  
5

Number of Column:  
5



#### 2. 14\*14 resolution

We use the concept of the moving average to cut the data and use the same method which is the combination of PCA and KNN.

The following illustration shows how we cut the data with the concept of the moving average.

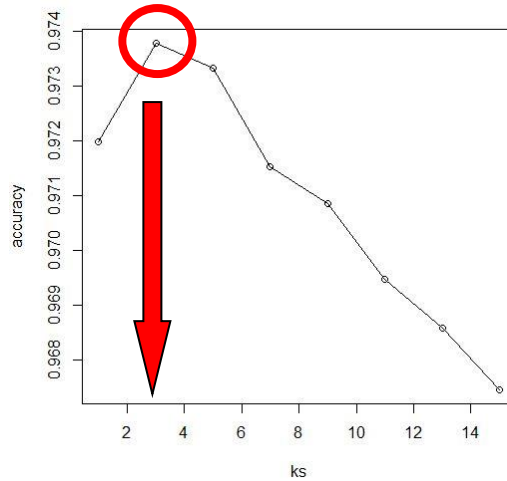


The first stage is PCA. First 20 principal components already cover 95% of the variability of data. However; to enhance our accuracy rate, we can observe that using first 50 PCs capture over 99.5% of the variation. Adding ten principal components based on 50 pcs only explains less than 0.26% variation. Therefore, we then decide to extract 50 PCs for use in the KNN classifier. The following is the cumulative sum table.

number of PCs	Cumulative Proportion
10	0.8688757
20	0.9511046
30	0.9770509
40	0.9876378
50	0.9927819
60	0.9954043
70	0.9969294
80	0.9978551
90	0.9984302
100	0.9988228

The second stage is kNN . We use cv to decide best k for KNN. (1/10(n=4200) for testing, 9/10(n=37800) for training)





We can achieve the highest accuracy rate when  $k=3$ .

## V. Conclusion

### 1. Conclusion for original data

- (1) Optimal result exists when  $PC = 60$ ,  $k = 3$
- (2) Achieve the accuracy rate of 98%
- (3) confusion matrix

True/Prediction	0	1	2	3	4	5	6	7	8	9	sum
0	422	0	0	0	0	0	2	1	0	0	425
1	0	452	2	0	1	1	1	1	1	1	460
2	0	0	448	0	0	0	1	5	0	0	454
3	0	0	3	376	0	2	0	1	3	1	386
4	0	0	0	0	399	0	1	1	1	7	409
5	0	0	0	4	0	385	4	1	1	3	398
6	1	0	0	0	1	0	434	0	0	0	436
7	0	3	1	0	1	0	0	420	0	2	427
8	0	0	2	1	1	3	3	0	394	0	404
9	0	0	0	1	6	3	0	4	1	386	401
Accuracy rate : 98.0%											

- (4) We submit our prediction to kaggle website, and received our accuracy rate 97.245%, which is close to 98%. Our ranking is approximately in the middle of all competitors.
- (5) In our process of selecting our tuning parameters, it's more reasonable utilizing cross validation to choose both numbers of principal components and  $k$ . Since this is a two stage analyzation, each step is related and will influence each other. However, in practice, it

takes such a long time running this procedure. Also, the selection of tuning parameters does not change if we choose to select PCs by the cumulative proportion of variation explained. Therefore; we still choose 60 principal components as our basis of the following analyzation.

## 2. Conclusion for degrade resolution

- (1) Optimal result exist when PC=50, k=3
- (2) Achieve the accuracy rate of 96.86%
- (3) In practice, it is possible to convert to the low resolution data if we have the higher resolution data.
- (4) Confusion matrix

True/Prediction	0	1	2	3	4	5	6	7	8	9	sum
0	419	0	0	0	0	0	5	0	0	1	425
1	0	454	2	0	1	0	0	1	2	0	460
2	0	1	442	0	0	1	0	7	2	1	454
3	0	1	4	372	0	3	0	2	3	1	386
4	1	1	0	0	391	0	2	1	0	13	409
5	0	0	0	6	0	384	5	1	0	2	398
6	2	0	0	0	1	1	432	0	0	0	436
7	0	2	1	1	2	0	0	416	0	5	427
8	1	1	4	2	4	5	5	0	380	2	404
9	0	3	0	5	4	3	1	6	1	378	401
Accuracy rate : 96.86%											

## VI. Reference

1. Jonathon,S. 2014. A Tutorial on Principal Component Analysis  
<https://arxiv.org/pdf/1404.1100.pdf>
2. Wikipedia  
[https://en.wikipedia.org/wiki/Principal\\_component\\_analysis](https://en.wikipedia.org/wiki/Principal_component_analysis)  
[https://en.wikipedia.org/wiki/Knearest\\_neighbors\\_algorithm#Validation\\_of\\_results](https://en.wikipedia.org/wiki/Knearest_neighbors_algorithm#Validation_of_results)
3. Revoledu - Strength and Weakness of K-Nearest Neighbor Algorithm  
<http://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm>
4. Gradient-Based Learning Applied to Document Recognition