

Digits Recognizer

Group 2

G.X. Liu, D.W. Chen and K.I. Chung

May 3, 2017

Introduction

- Data Resource
- Information
- Image Examples

Summary Statistics

- Means
- 20% Trimmed Means
- Principal Component Analysis

Future Work

- Models
- Predict Instantly Online

The Handwritten Digits

They are originally from the MNIST database.

- ▶ 60000 training data
- ▶ 10000 test data

Now, it is a competition on the Kaggle.

- ▶ 42000 training data
- ▶ 28000 test data

Information

- ▶ Images contain 28×28 pixels.
- ▶ There are 784 independent variables.
- ▶ The response labels the real values of images.

Table: Frequency of Digits

digit	0	1	2	3	4
freq.	4132	4684	4177	4351	4072
digit	5	6	7	8	9
freq.	3795	4137	4401	4063	4188

Representation

The function, f transfers an observation with 784 variables into an 28×28 image matrix.

$$f : \mathbb{R}^{784} \rightarrow M_{28 \times 28}$$

$$M = \begin{bmatrix} \text{pixel}_{000} & \text{pixel}_{001} & \cdots & \text{pixel}_{027} \\ \text{pixel}_{028} & \text{pixel}_{029} & \cdots & \text{pixel}_{055} \\ \vdots & \vdots & \ddots & \vdots \\ \text{pixel}_{756} & \text{pixel}_{757} & \cdots & \text{pixel}_{783} \end{bmatrix}_{28 \times 28}$$

Images from the Training Data

A handwritten digit '0' in blue ink on a light gray background.A handwritten digit '1' in blue ink on a light gray background.A handwritten digit '2' in blue ink on a light gray background.A handwritten digit '3' in blue ink on a light gray background.A handwritten digit '4' in blue ink on a light gray background.A handwritten digit '5' in blue ink on a light gray background.A handwritten digit '6' in blue ink on a light gray background.A handwritten digit '7' in blue ink on a light gray background.A handwritten digit '8' in blue ink on a light gray background.A handwritten digit '9' in blue ink on a light gray background.

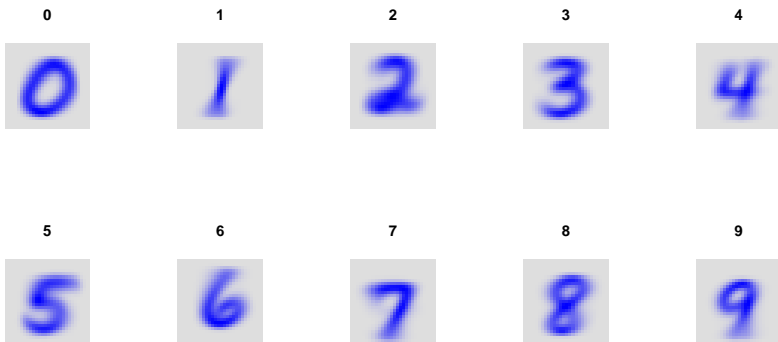
Images from the Test Data

A handwritten digit '6' in black ink on a light gray background.A handwritten digit '9' in black ink on a light gray background.A handwritten digit '3' in black ink on a light gray background.A handwritten digit '1' in black ink on a light gray background.A handwritten digit '8' in black ink on a light gray background.A handwritten digit '9' in black ink on a light gray background.A handwritten digit '2' in black ink on a light gray background.A handwritten digit '4' in black ink on a light gray background.A handwritten digit '3' in black ink on a light gray background.A handwritten digit '4' in black ink on a light gray background.

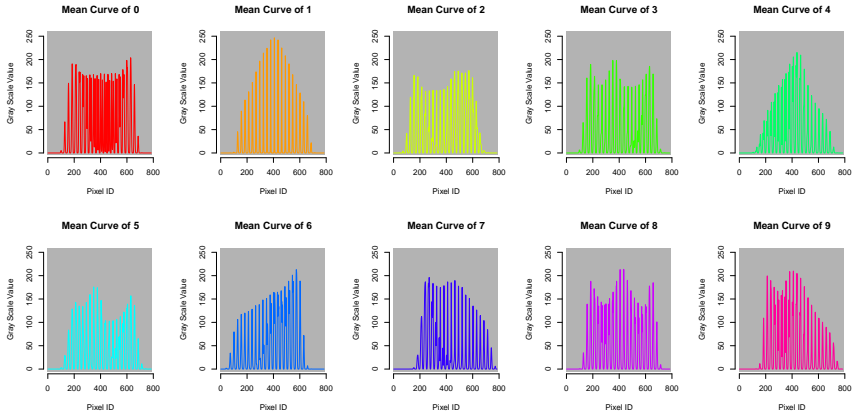
Some Naked-eyed Unrecognizable Images



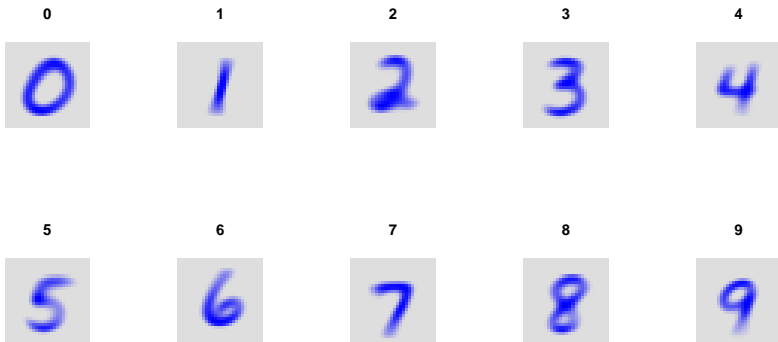
The Mean Images of the Different Digits



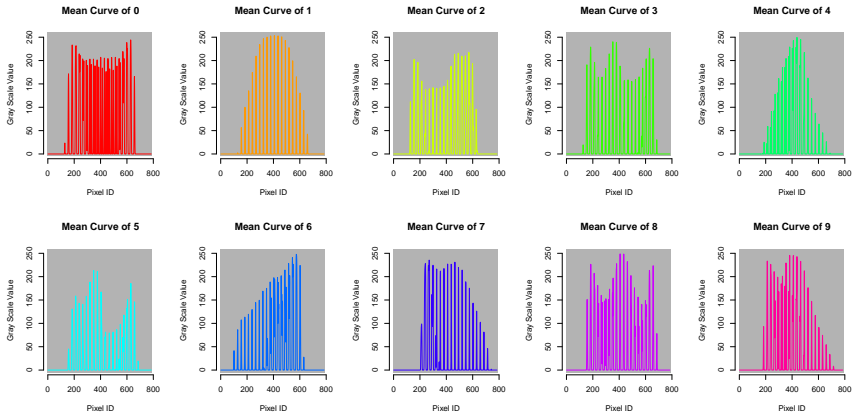
The Mean Curves of the Different Digits



The 20% Trimmed Mean Images of the Different Digits



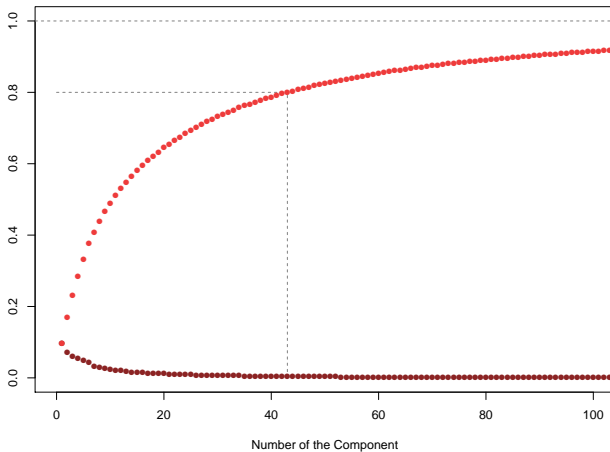
The 20% Trimmed Mean Curves of the Different Digits



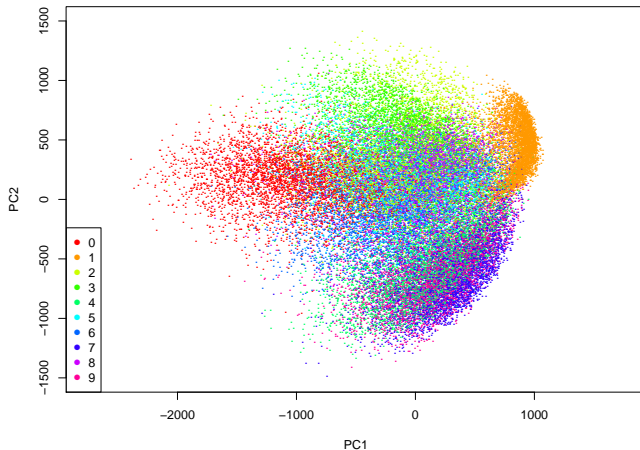
Coefficients for the Principal Component

	PC1	PC2	...	PC43	...	PC784
$pixel_0$	0.000	0.000	...	0.000	...	0.000
$pixel_1$	0.000	0.000	...	0.000	...	-0.072
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
$pixel_{462}$	0.075	-0.013	...	0.059	...	0.000
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
$pixel_{783}$	0.000	0.000	...	0.000	...	0.000
Variance	5.149	3.781	...	0.221	...	0.000
Cumulative Ratio of Total Variance	0.098	0.169	...	0.800	...	1.000

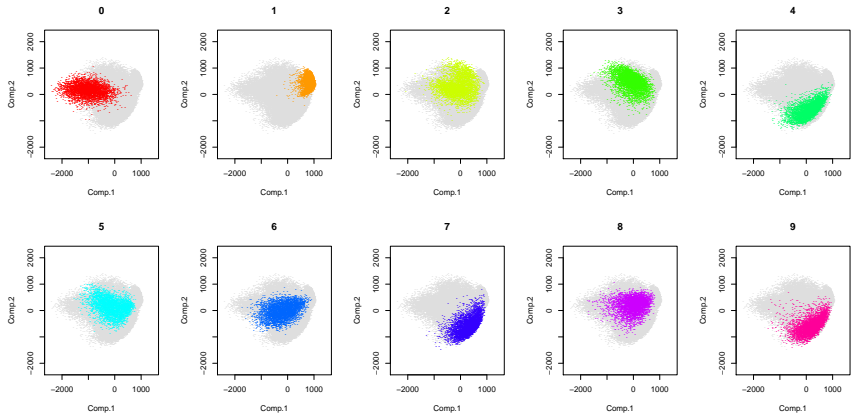
Cumulative Ratio of Total Variance



Scatter Plot for the First Two Components



Separated Scatter Plot for the First Two Components



Models

- ▶ One-Versus-All Logistic Regression
- ▶ PC One-Versus-All Logistic Regression
- ▶ K-Nearest Neighbors Classifier (KNN)
- ▶ Random Forest
- ▶ Support Vector Machine (SVM)
- ▶ Convolutional Neural Network (CNN)

Example

- ▶ <http://myselfph.de/neuralNet.html>
- ▶ Neural Network Model
- ▶ Trained with Matlab and Displayed in JavaScript
- ▶ 3 msec
- ▶ 1.92% error rate

Plans

Plan A Calculate on JavaScript (Less Time Complexity)

Plan B Front-End + LAMP + R (More Time Complexity)

Plan C Front-End + PHP + Cloudinary + R

Thank you for listening.