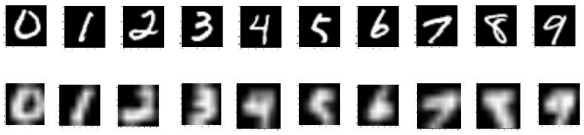
K-nearest-neighbor and GDA for digit recognition

Anh T. Huynh

Digit recognition – Reducing image size



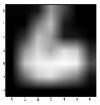
Training set consists of:
•trainingImagesA.npy, trainingImagesB.npy, each consists of 20000 of 28x28 grayscale images of digits.
•trainingLabelsA.npy, trainingLabelsB.npy, each consists of 20000 labels for the images above.

In order to avoid the curse of dimensionality, the images are reduced to 7x7 grayscale images by taking the averages of overlapping subsquares. The dimension is reduced by a factor of 16.

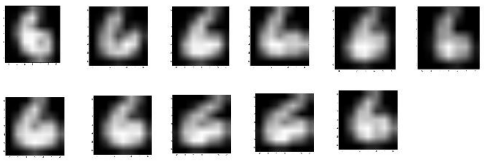
K-nearest-neighbor Model

- Assign Euclidean distances between images.
- For each image, find the k nearest neighbors of it.
- Each of the k neighbors give a vote for its own label
- The label with most votes wins, and is used to predict the label of the image.

An image with label 6



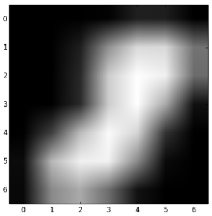
Its 11 nearest neighbors, all of which have label 6



Gaussian Discriminant Analysis Model

- Assume that images for each digit are distributed according to the multivariate normal distribution.
- For each digit find the Mean and Covariance matrix using training images
- To predict, compute P(label = i|image)~P(image|label=i) and choose the mode.

P(image | label = 0) = 2.82 e-094
P(image | label = 1) = 8.69 e-176
P(image | label = 2) = 2.34 e-083
P(image | label = 3) = 2.46 e-079
P(image | label = 4) = 4.28 e-137
P(image | label = 5) = 3.58 e-077
P(image | label = 6) = 2.47 e-234
P(image | label = 7) = 3.18 e-278
P(image | label = 8) = 1.47 e-051
P(image | label = 9) = 0
Therefore, predict the label to be 8

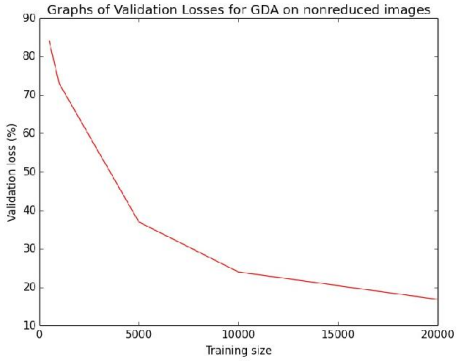


Knn: reduced vs nonreduced performance

training size ($k = 3$)	nonreduced images	reduced images
500	85.76%	17.36%
1000	81.10%	14.36%
5000	77.20%	9.86%
10000	77.14%	7.50%

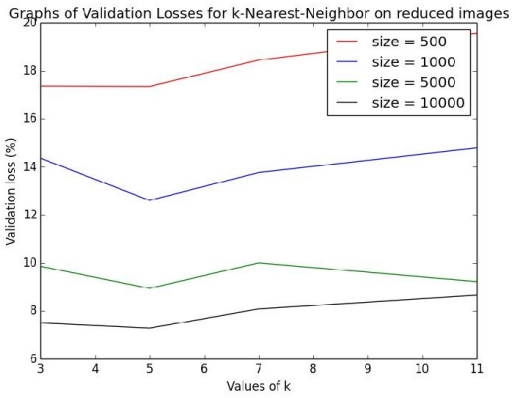
- Knn on nonreduced images performs barely better than random guess.
- Improvement in increasing training size is modest for knn on nonreduced images.
- Knn on reduced images performs reasonably well, but prediction time using 10000 images as training set is very slow: 800 seconds for 5000 test images

Training size for GDA



- Prediction is better with increase of training size.
- Best validation loss is at 16.82%, with 20000 training images.
- Very fast: 20 seconds for 5000 test images

Parameters for k-nearest-neighbor



- k = 5 consistently beats k = 3, and is more stable
- Best validation loss is 7.28% for k = 500, trained on 10000 images