MATLAB Project

Introduction to Linear Algebra **Fall, 2020**

1. MNIST database classification

   The goal of this problem is to develop a classifier that categorizes handwritten digits, using least squares and PCA. MNIST database is a large database of handwritten digits$(0, 1, \ldots, 9)$ that is commonly used for training various image processing systems, which consists of $60,000$ training images and $10,000$ test images of size 28×28.

Figure 1: MNIST database

In this problem, we will use $50,000$ training images(each digits has $5,000$ images) to train the classifier, and $10,000$ test images to compute its accuracy. Use the code `load('MNIST_DATA.mat')` to get the (modified) MNIST database.

variables	size	meaning
DX	5,000×784×10	(number of data)× (image size)× (class)
DY	5,000×10	(number of data)×(class)
TX01	1,000×784	(number of image)×(image size)
TY01	1,000×1	(class)
TX	10,000×784	(number of image)×(image size)
TY	10,000×1	(class)

For example, $DX(5, :, 3)$ is 5^{th} image of a digit 2 and its label 2 can be obtained by $DY(5, 3)$. One can check the image by using the following code :

```
imshow( reshape(DX(5, :, 3), [28, 28])' )
```

Here, D, T, X and Y denote training data, test data, image and label, respectively. For example, DX is training image data and TY is test label data. $TX01$ and $TY01$ are test image data and test label data for 1-(a).

(a) Binary classification using least squares

The goal of this subproblem is to develop a binary classifier that categorizes digits 0 and 1, using least squares. In specific, we will construct a binary classifier

$$\text{sign}([x \mid 1]a) = \text{sign}\left(\sum_{i=1}^{784} x_i a_i + a_{785}\right),$$

which outputs 1 if a row vector x corresponds to an image of digits 0, and -1 if x is an image of digit 1. Note that the dimension of a is 785, where a_{785} is a bias term. Construct a user-defined file

$$\text{function [a] = MNIST\_LS}(X_1, X_2),$$

which solves the following least square problem

$$a^* = \arg\min_{a \in \mathbb{R}^{785}} \|Xa - y\|_2^2 \quad \text{where } X = \begin{bmatrix} X_1 \; 1 \\ -- \\ X_2 \; 1 \end{bmatrix}_{10{,}000 \times 785} \quad \text{and } y = \begin{bmatrix} 1 \\ - \\ -1 \end{bmatrix}_{10{,}000 \times 1}.$$

Here, X_1 is a $5,000 \times 784$ matrix of training images for digit 0, and X_2 is a $5,000 \times 784$ matrix of training images for digit 1. Draw the image of a^* by using the code :

$$\text{imshow(reshape}(a(1:784), [28, 28])', [-1\ 1]\ ).$$

Compute the accuracy of classification for the test data TX01. Considering that the value of $[x \mid 1]a^*$ implies the confidence of the classifier, find two images with digit 0 that have the largest and the smallest values of $[x \mid 1]a^*$, and provide brief explanation.

(b) Multi-class classification using least squares

The goal of this subproblem is to extend 1-(a), and develop a 10-class classifier that categorizes all handwritten digits from 0 to 9. In specific, we will first construct ten binary classifiers a_k^*, $k = 1, \ldots, 10$ that each classifies each digit $k-1$ from all others $(0, \ldots, k-2, k, \ldots, 9)$. Then, the 10-class classifier

$$\arg\max_{k=1,\ldots,10} [x \mid 1]a_k^*$$

will classify any image x of handwritten digits. Explain how you constructed a_k^*, and draw the images of a_k^* for $k = 1, \ldots, 10$, as in 1-(a). Compute the accuracy of 10-class classification for the test data TX. Note that MATLAB index is 1 to 10 and label Y is 0 to 9.

(c) Multi-class classification using PCA

The goal of this subproblem is to construct a 10-class classifier using PCA. In specific, find the first d principal components $W_k^d = (w_{k1}, w_{k2}, \cdots, w_{kd})$ of training images $DX(:, :, k) \in \mathbb{R}^{5,000 \times 784}$ for each digit k using PCA. Then, to classify any handwritten image x, project it onto a span of the first d principal components, i.e., $\text{span}(W_k^d)$, for each class k, and use the norm of the projected image as the confidence score for classification. Plot the accuracy for the test data TX, versus dimension d from 1 to 40. Find the dimension d that gives the highest accuracy, and determine the smallest dimension d that has an accuracy higher than that of 1-(b).

2. Randomized Linear Algebra

The goal of this problem is to have experience on the randomized matrix multiplication of A and B:

$$AB \approx CR = \sum_{j=1}^{s} \frac{A_{k(j)} B_{k(j)}}{s p_{k(j)}}$$

using two different probability distributions p_k for sampling. Use the code `load('RLA_DATA.mat')` to get A and B. We choose $s = 10$ and simulate this $N = 200$ times to check the mean and variance. Plot two probability distribution p_k for uniform sampling and norm-squared sampling, versus index k. Plot the histograms of the number of sampling throughout all simulations for each index k. Plot the relative errors for each $N = 200$ simulations:

$$\text{relative errors} = \frac{\|CR - E[CR]\|_F}{\|E[CR]\|_F}$$

Report the sample variances for two relative error, and discuss whether or not such empirical results match the theory.

Submission guide

- Submit `Report` as pdf file and corresponding `m-files` to *Homework box for MATLAB Project* on the KLMS.
  [To validate and grading your code, we will run the code.]

Report guide

- Do not exceed 2 pages for each problems. (Totally 8 pages)
- You can write in Korean.
- You can write in 2 column format.
- There are keywords for each problem

1-a Draw a^*//plot the confidence for the test data $TX01$//accuracy of $TX01$//explain how you solve the least-square problem(refer to built-in function)

1-b Draw a_k^* in subplot(2, 5)//accuracy of TX

1-c plot the accuracy versus dimension and 1-(b) accuracy in same axis//find the smallest dimension that has an accuracy higher than that of 1-(b)//draw the 5 principal components for each class k. i.e., draw (W_k^5)//Compare 1-(b) and 1-(c)

2 plot the two probability distributions//plot the histograms of the number of sampling throughout all simulations for each index k//plot the two relative errors//sample variances for two relative error//explain how you code the sampling process

Due date : Dec 15 (Tue) 15:00
Late submission will not be allowed.