AMATH 482/582: HOME WORK 2

RAPHAEL LIU

Department of Appliede Mathematics, University of Washington, Seattle, WA
raph651@uw.edu

Abstract. In this assignment, a train dataset of 2000 images and a test dataset of 500 images are given to us. The images are handwritten digits of 0-9. We are going to develop a supervised ML model to classify the digits and predict future images. The task is to select two digits and develop a linear regression model that tells the two digits apart from each other. The methods will be used include PCA(Principal Component Analysis), Ridge regression, and MSE(Mean Sqaured Error).

1. Introduction and Overview

Our goal in this homework is to train a classifier to distinguish images of handwritten digits from the famous MNIST data set. The training set contains 2000 instances of handwritten digits, the "features" are 16 ×16 black and white images while the "labels" are the corresponding digit (note the images are shaped as vectors of size 256 and need to be reshaped for visualization). The test set has the same attributes except that there are only 500 instances. Below is a list of tasks to complete.

(1) Use **PCA** to investigate the dimensionality of X_{train} and plot the first 16 PCA modes as 16 ×16 images.

(2) How many PCA modes do we need to keep in order to approximate X_{train} up to 60%, 80% and 90% in the **Frobenius norm**? Do we need the entire 16×16 image for each data point?

(3) Train a classifier to distinguish the digits 1 and 8 using **Ridge regression** and report the corresponding **MSE** of the classifier.

(4) Same procedures to train a classifier for the pairs of digits **(3,8) and (2,7)** and report and train and test MSE's. Explain the performance variation if applicable.

2. Theoretical Background

2.1. Singular Value Decomposition (SVD).

The singular value decomposition of a matrix is usually referred to as the SVD. This is the final and best factorization of a matrix:

$$A = U\Sigma V^T$$

where U is orthogonal, Σ is diagonal, and V is orthogonal. In the decomoposition $A = U\Sigma V^T$, A can be any matrix. We know that if A is symmetric positive definite its eigenvectors are orthogonal and we can write $A = Q\Sigma Q^T$. This is a special case of a SVD, with $U = V = Q$. For more general A, the SVD requires two different matrices U and V.

Date: Feb 11 2022.

2.2. **Supervised Learning.**

The function model assumes there exists a function $f^+ : X \to y$ so that $y_i = f^+(x_j) + \epsilon_j$, where ϵ_j are some noise that may be in the output or our observation of the $f^+(x_j)$. By far the most common assumption is the Gaussian noise,

$$\epsilon_j \approx \mathcal{N}(0, \sigma^2)$$

This implies $y_i|x_j \approx \mathcal{N}(f^+(x_j), \sigma^2)$, and that $\Pi(y_j|x_j) \propto exp(-\frac{1}{2\sigma^2}|f^+(x_j) - y_j|^2)$, where Π is the PDF of y for fixed x_j. At this moment, it is useless without a model, since there are many solutions. One of the most simple model is **linear regression**.

$$f_{MLE} \equiv \beta_{MLE} = \underset{B\in\mathbb{R}^d}{\arg\min} \frac{1}{2\sigma^2}||A\beta - Y||^2$$

Therefore, MLE is nothing but a least square solution to the problem. Typically, the system is over-determined. Solution is given by solving the normal equations,

$$\frac{\partial}{\partial\beta}(\frac{1}{2\sigma^2}(A\beta - y)^T(A\beta - y)) = \frac{1}{\sigma^2}A^T(A\beta - y) = 0$$

$$\implies A^T(A\beta - y) = 0$$
$$\beta = (A^TA)^{-1}A^Ty$$

2.3. **Regularization and MSE.**

We consider an extra term in our linear regression model so that becomes

$$\hat{\beta} = \underset{B\in\mathbb{R}^d}{\arg\min} \frac{1}{2\sigma^2}||A\beta - y||^2 + \frac{\lambda}{2}||\beta||_p^p$$

$\lambda \geq 0$ is called the regularization/penalty parameter & $p \geq 1$ denotes the choice. $p = 2$ for Ridge regression.

$$\beta = (\frac{1}{\sigma^2}A^TA + \lambda I)^{-1}A^Ty$$

So doing SVD of A,

$$\frac{1}{\sigma^2}A^TA + \lambda I = V(\frac{1}{\sigma^2}\Sigma^2 + \lambda I)V^T$$

, the diagonals are non-negative, eliminating zeros so that A matrix can be invertible.

Again, the choice of λ is important for stability and accuracy. The Mean Sqaured Error(MSE) reflects how effective our model is. It is given in the Euclidean norm form:

$$MSE := \frac{1}{\text{length of y}}||A\beta - y||_2^2$$

3. Algorithm Implementation and Development

The procedures for solving the tasks are given below, separately:

(1) Prof.Hosseini gives a valuable helper notebook for splitting the train and test data into features matrix X and labels array y. The 16×16 image is flattened to a 256 array, and 2000 different images give a 2000×256 matrix X.

(2) The PCA part is done by using Python package sklearn's PCA library. First, fit the X and find the first 100 principal components. Analyze each component's singular value and reconstruct the first 16 PCA modes in the image form.

(3) The Frobenius norm is calculated by summing the first n squared singular values and takes square root. Find the least n for keeping a specific level of Frobenius norm. This is the number of modes needed.

(4) Set PCA components number to be 16. Extract the train data relevant to digits 1,8 by filtering its label and normalizing them to -1 and 1, respectively. From sklearn's linear regression model, build two Ridge regression models with different λ values, which are also the alpha values. Fit the extracted train data and use the coefficient of regression model β to calculate MSE for train data.

(5) Extract the test data relevant to selected digits and compute MSE for test data.

(6) Repeat the procedures for different pairs of digits. Record the MSE and note the performance of variation.

4. Computational Results

We start by plotting the first 64 images features in our train dataset for visualization of the handwritten digits. The images show no pattern on occurrence of digits and are visible to tell digits 0-9 apart by human eyes.

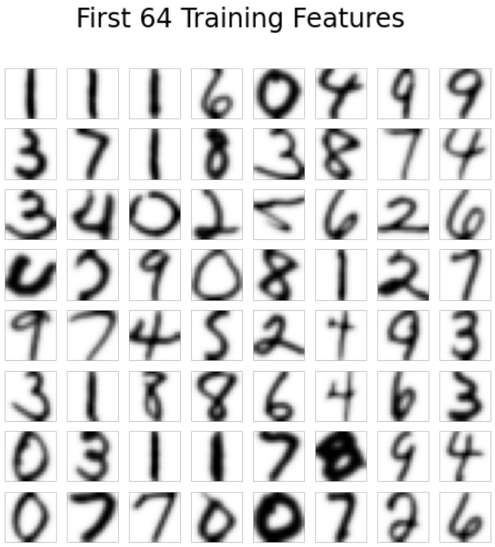
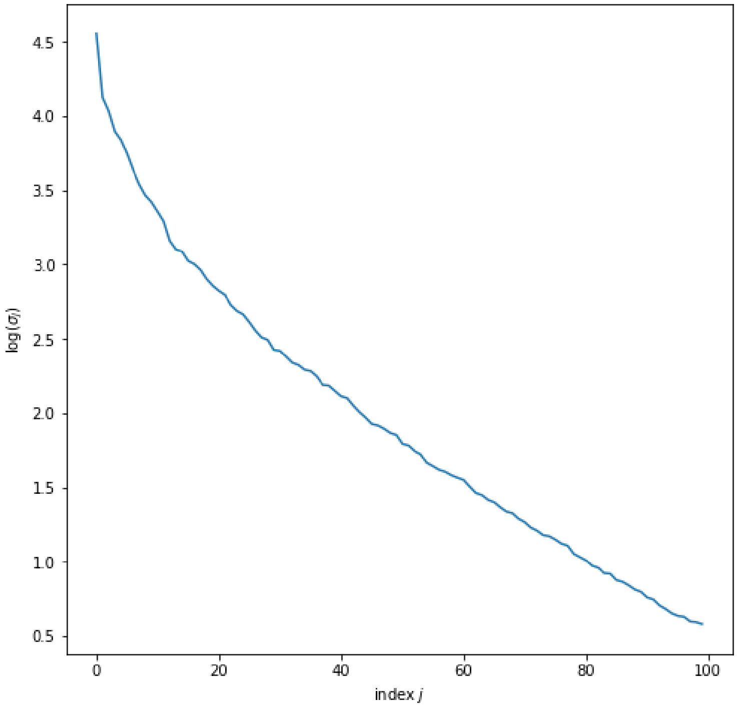


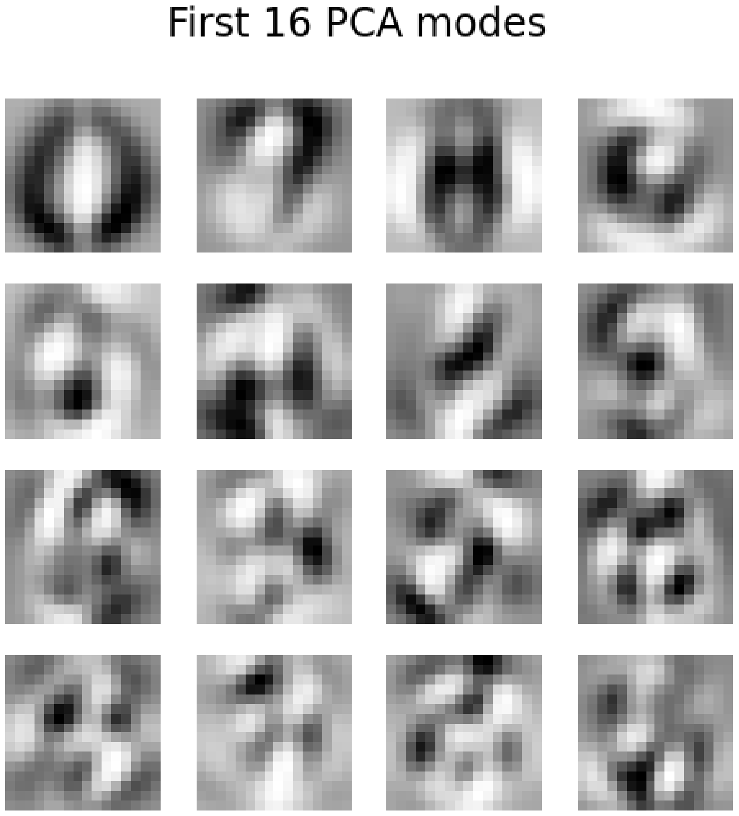
Figure 1. First 64 features.

PCA explains the covariance and direction of maximum singular values in our dataset. The sklearn PCA method finds the principal singular values for us. Below is a plot for log of the first 100 PCA coefficients.2a We see that the value of coefficients drops rapidly as index increases, which means the first several PCA modes dominate the data decomposition. In fact, to keep the 60%, 80%, and 90% Frobenius norm in our low-rank approximation we only need the first 3,7, and 14 PVA modes. Hence, we don't need the entire 16×16 image for each data point.

The 16 PCA modes are plotted for visualization below. 2b



(A) The first 100 PCA coefficients



(B) The first 16 PCA modes

After extracting the digits 1,8 from out train & test data, we project them on the first 16 PCA modes using Ridge regression with two different λ values. One $\lambda = 1$ and the other with value 0.02. The train and test MSE are calculated. We see that two different alphas give results that are very close. The smaller alpha =0.02 gives a slightly smaller MSE. So for the next part, we use alpha =0.02. See table below.

MSE	$\alpha = 1$	$\alpha = 0.02$
train	0.074613	0.074599
test	0.13818	0.13508

Table 1. MSE for digits 1,8 classifier with two different α values

Repeat the same procedures for pairs of digits 3,8 and 2,7. The below table shows corresponding MSE when $\alpha = 0.02$. We see that for the model's MSE is higher for telling 3 and 8 apart from the other. This makes sense because the number '3' and number '8' look more alike, both from human eyes and from machines.

MSE	digits (3,8)	digits (2,7)
train	0.18040	0.091779
test	0.57001	0.37424

Table 2. MSE for digits 1,8 classifier with two different α values

5. Summary and Conclusions

In order to classify the digits, there are a few steps most importantly to obtain the results. These steps include performing PCA on train data, finding the number of modes needed for preserving Frobenius norm, building regression model case-wise, and computing MSE for validation. There is a noticeable MSE difference between classifiers when classifying different pairs of digits. Specifically, ('3' and '8') look much more alike than ('1' and '8') or ('2' and '7'), which makes its MSE greater than the other two. From this homework, we practice a powerful technique and algorithm for developing supervised learning models.

Acknowledgements

The author is thankful to Prof. Hosseini for useful discussion about the topics in supervised learning and the visualization helper he provides.