10-601: Homework 8 Experiment

Due: 24 November 2014 11:59pm (Autolab)

TAs: Kuo Liu, Yipei Wang

Name: _____

Andrew ID: _____

Please answer to the point, and do not spend time/space giving irrelevant details. Please state any additional assumptions you make while answering the questions. For Questions 1 to 5, 6(b) and 6(c), you need to submit your answers in a single PDF file on autolab, either a scanned handwritten version or a LaTeXpdf file. Please make sure you write legibly for grading. For Question 6(a), submit your m-files on autolab.

You can work in groups. However, no written notes can be shared, or taken during group discussions. You may ask clarifying questions on Piazza. However, under no circumstances should you reveal any part of the answer publicly on Piazza or any other public website. The intention of this policy is to facilitate learning, not circumvent it. Any incidents of plagiarism will be handled in accordance with CMU's Policy on Academic Integrity.

⋆: Code of Conduct Declaration

- Did you receive any help whatsoever from anyone in solving this assignment? Yes / No.
- If you answered *yes*, give full details: _____ (e.g. *Jane explained to me what is asked in Question 3.4*)
- Did you give any help whatsoever to anyone in solving this assignment? Yes / No.
- If you answered *yes*, give full details: _____ (e.g. *I pointed Joe to section 2.3 to help him with Question 2*).

Experiment: Dimensionality Reduction using PCA

I. Introduction

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA can be used for feature dimension reduction.

In this problem set, you are going to do two experiments. In the first experiment, you are going to train pca on image data and visualize the principle components. You will also calculate the reconstruction error and compare the reconstructed image with the original one. In the second experiment, you are going to compare the performance of classifiers using different feature representation, including the original feature and several lower dimension features by applying PCA.

II. Dataset

Please find the dataset and corresponding description from this link http://www.cs.cmu.edu/~tom/faces.html. The face data set can be downloaded through this link: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-8/faceimages/faces.tar.Z

The training list and testing lists can be found here: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-8/faceimages/trainset/ Please download all_train.list, all_test1.list, all_test2.list. The images listed in all_train.list will be used as training data. The images listed in all_test1.list and all_test2.list will be used as two different testing datasets. You might have to change the directories in these lists since they are the directories on the server.

Experiment I: Visualize the Eigen Face and Image Reconstruction

(a) Train PCA using all the data (all_train, all_test1 and all_test2) and visualize the first 5 principle components. Please choose SVD for PCA algorithm and show your visualized results below.

Hints:
1. By default, the pca function in matlab centers the data and uses the singular value decomposition (SVD) algorithm.

2. Each image should be treated as one instance. So you need to vectorize the matrix. The following matlab code help you to load the image data and form the matrix for training PCA.
$A = imread(imagefile,'pgm');$
$Data = [Data; double(A(:)')];$

3. To visualize the principle component, you need to recover to the original space (if the data is centered when training PCA) and reshape the vector into matrix. Check the 'reshape' function in matlab. Also, you have to rescale the matrix value to 0-255 and store as image format. Check the 'uint8' function in matlab.

4. This link might help you understand the problem: http://en.wikipedia.org/wiki/Eigenface

[*10 points*]

(b) Please use the first image in training data (kawamura_straight_happy_open_4.pgm) for this question. reconstruct this image using the first 50 princpical components and calculate the reconstruction error. Please show the reconstructed image.
Please calculate the reconstructed error using the first n principle components. n=100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600.

Hint
You can refer to the 24th page of the lecture slides the definition of reconstruction error.

[*10 points*]

Experiment II: Feature Dimension Reduction in Classification

You are also going to compare the classification accuracy of 'sunglasses' using the feature vectors before and after doing dimensionality reduction. The label of sunglasses is included in the name of the image. You can also refer to the named rules for the image in the data description.

Here we use LibSVM tool as classifier(http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The classifier is trained on data in all_train.list and evaluated on two testing datasets (all_test1.list and all_test2.list) separately.

The PCA reduced feature are the scores of projecting the original feature vector into the space constructed by a subset of the principle components. The K dimension PCA feature means you take the scores by projecting into the first K principle components.

Hint

1. In training SVM classifier, you might need to rescale the feature value to 0-1 to avoid domination error. The testing data need to be rescaled using the same parameters. The following Matlab code can help you.

$upper = max(X);$
$lower = min(X);$
$N = size(X, 1);$
$scaled_X = (X - repmat(lower, N, 1))./repmat(upper - lower, N, 1);$

You can also use **svm_scale** in LibSVM.

2.In training SVM, you need to tune the parameter of the penalty and the kernel. Please use the following parameter setting.

$original\ data,\ -c\ 100\ -g\ 0.01$
$PCA\ 50\ dimension,\ -c\ 500\ -g\ 0.07$
$PCA\ 150\ dimension,\ -c\ 100\ -g\ 0.16$
$PCA\ 200\ dimension,\ -c\ 100\ -g\ 0.07$

You can use grid search by yourself to find parameters to achieve higher accuracy. LibSVM provides the functionality for automatic grid search.

(a) What is the classification accuracy using the original feature?

[*10 points*]

(b) What is the classification accuracy using 50 dimensions PCA feature? How about the classification accuracy using 150 dimensions and 200 dimensions?

[*10 points*]

(c) Please compare the performance of these classifiers and explain the advantage of using PCA feature.

[*5 points*]

Submission

You only need to submit the report in a pdf file (all the images required to show and the calculation results need to be included). You should also submit your code with clear comment in a separate folder.