CS303B Artificial Intelligence

CS303B.2 Assignment Details

CS303B.2 Assignment: Classification and Clustering

DUE: CS303B.2 – Friday, 11:55 pm, Week 16 (29 Dec 2022)

SUGGESTED EFFORT: 35 hours

WEIGHTING: 30% + 5% of Final Mark

Description:

This assignment is based on the lectures and practical labs about classification and clustering, dimensionality reduction in this semester.

The Dataset

The MNIST handwritten digit database is a commonly used dataset in data mining and machine learning. The purpose of MNIST dataset is providing data on which to compare methods to classify handwritten digit images into 10 digits (0-9). In this assignment, the dataset used is a sub-set of the MNIST dataset, containing images from three handwritten digits (1, 5, 8). Each digit class contains 100 images of size 28x28 pixels. Below is a snap shot of 20 images from the MNIST dataset.

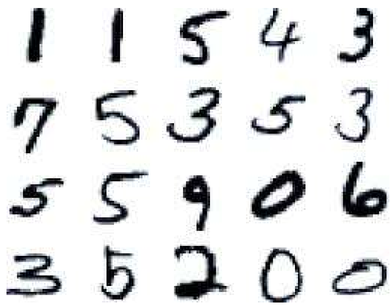

Fig. 1, a snap shot of 20 sample images from the MNIST dataset.

The handwritten digit dataset used in this assignment is available from Blackboard under the assignment menu item.

Before you start, the first step is to create a feature vector for each image. Here we will use the simplest form of concatenating each column of an image forming one vector to represent the image, e.g., in Matlab, if we use a matrix A to represent an image, F=A(:) will convert the image matrix into a vector (a 784-dim feature vector for each image in our dataset). The images in the dataset provided are already vectorised. To get you started, try the following:

```
% Load dataset in Matlab
load  ac50001_assignment_data.mat
```

% you will then see three different variables: *digit_one*, *digit_five*, *digit_eight*, containing 100 images for each of the digits '1', '5', and '8' respectively.

% To visualise one image , you can simply try:
im = reshape(digit_one(:,1), [28, 28]); % the first image of the digit '1'
imshow(im,[]);

Now you can proceed to the following questions:

Questions

1) Use PCA to reduce the dimensions of each image descriptor to two using the first two principal components, and cluster those data points in the 2-D space into 3 clusters using one of the clustering methods that we have learned in lectures (hierarchical clustering, k-means, GMM etc), and plot a scatter plot, that shows the cluster labels. Discuss whether the resulting clusters match well the actual ground truth partition of classes, i.e., are images from the same digit are clustered into one region?

2) Use LDA instead of PCA for dimensionality reduction and repeat Question 1. Plot the scatter plot after applying LDA. Compare the results from Question 1, and discuss.

3) Now consider separating the images of digit '5' from the rest (the images of '1' and '8'). Note this is now a two-class classification problem. Use SVM with a RBF kernel, SVM with a linear kernel, and a neural network classifier with one hidden layer to classify the dataset in a 5-fold cross validation setting. Compare and discuss the results using the obtained validation accuracy. Create and plot the ROC curves of the results using the three different classifiers, and compare their performance using the area under the ROC curve (AUC). Pick one parameter (e.g., a penalty parameter, or a parameter from a kernel) from SVM, and show that how you can properly tune a parameter on this dataset.

Assessment:

In assessing this piece of work, we will be looking for the following:

- An understanding of classification, dimensionality reduction, and clustering.
- An ability to use Matlab or C to program with existing tools for data mining and classification.
- An ability to discuss machine learning methods and compare performances.
- Note you should not directly call the build-in function PCA and LDA as a black box. Instead LDA and PCA should be implemented by following the procedure in the lectures.

Submission:

The submission will include:

- A written report containing the answers to each question together results and graph, as well as a brief description of how you implemented your software system.

- Code for the use of PCA, LDA, one clustering method, and constructing the scatter plots.
- Code for the use of SVM, and neural network for classification.
- Code for 5-fold cross validation, and generating ROC curves.
- Instructions on how to run the code.

The full submission of this assignment must be submitted in **both paper and electronic** formats by **11:55 pm. on the Friday of Week 16 (29 Dec 2022):**

- The paper submission must be submitted to TAs;
- The electronic copy must be zipped into a file and submitted to blackboard.