

**University of Waterloo**  
**ECE 657A: Data and Knowledge Modeling and Analysis**  
**Winter 2017**

**Assignment 1: Data Cleaning and Dimensionality Reduction**

**Due:** February 10th, 2016 11:59pm

## Overview

**Assignment Type:** done in groups of up to three students.

**Notification of Group:** by Wednesday Jan 25, **one member** of the group should email mcrowley@uwaterloo.ca and the TA Rasoul (r26mohammadinasiri@uwaterloo.ca) with the names of everyone in the group with *[first name, last name, student number, email]* for everyone in the group.

**Hand in:** One report (PDF) per group, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. (If you don't know  $\text{\LaTeX}$  you should try to use it, it's good practice and it will make the project report easier)

**Objective:** To study how to apply some of the methods discussed in class on three datasets. The emphasis is on analysis and presentation of results not on code implemented or used. You can use libraries available in MATLAB, python, R or any other programs available to you. You need to mention explicitly the source with proper references.

## Data sets

Available on LEARN, if you aren't registered yet they should be emailed to you.

**Dataset A :** This is a time-series dataset which is collected from a set of motion sensors for wearable activity recognition. The data is given in time order, with 19,000 samples and 81 features. Some missing values ('NaN') and outliers are present. (note: The negative values are not outliers) This data is used to illustrate the data cleaning and preprocessing techniques. (File: [DataA.mat](#))

**Dataset B :** Handwritten digits of 0, 1, 2, 3, and 4 (5 classes). This dataset contains 2066 samples with 784 features corresponding to a 28 x 28 gray-scale (0-255) image of the digit, arranged in column-wise. This data is used to illustrate the difference between feature extraction methods. (File: [DataB.mat](#))

**Dataset C:** This data contains measurements of heart cardiocograms. The goal is to classify an observation to be one of the three categories: normal(1) / suspect(2) / pathologic(3) given as the ground truth level **gnd**. It includes a sample-feature matrix **fea** with 2100 samples with 21 features and 3 classes. Features represent measurements of heart rate and uterine contraction features. Each sample is a separate row. This data is used to illustrate the difference between feature selection methods. (File: [DataC.mat](#))

## Questions

### I. Data Cleaning and Preprocessing (for dataset A)

1. Detect any problems that need to be fixed in dataset A. Report such problems.
2. Fix the detected problems using some of the methods discussed in class.
3. Normalize the data using min-max and z-score normalization. Plot histograms of feature 9 and 24; compare and comment on the differences before and after normalization. For both features, plot auto-correlation before and after normalizations and report and discuss observations.

### II. Feature Extraction (for dataset B)

1. Use PCA as a dimensionality reduction technique to the data, compute the eigenvectors and eigenvalues.
2. Plot a 2 dimensional representation of the data points based on the first and second principal components. Explain the results versus the known classes (display data points of each class with a different color).
3. Repeat step 2 for the 5th and 6st components. Comment on the result.
4. Use the Naive Bayes classifier to classify 8 sets of dimensionality reduced data (using the first 2, 4, 10, 30, 60, 200, 500, and all 784 PCA components). Plot the classification error for the 8 sets against the retained variance ( $r_m$  from lect3:slide22) of each case.
5. As the class labels are already known, you can use the Linear Discriminant Analysis (LDA) to reduce the dimensionality, plot the data points using the first 2 LDA components (display data points of each class with a different color). Explain the results obtained in terms of the known classes. Compare with the results obtained by using PCA.

### III. Nonlinear Dimensionality Reduction (for dataset B)

Apply the nonlinear dimensionality reduction methods Locally Linear Embedding (LLE) and ISOMAP to the dataset B, set the number of nearest neighbours to be 5, the projected low dimension to be 4.

1. Apply LLE to the images of digit '3' only. Visualize the original images by plotting the images corresponding to those instances on 2-D representations of the data based on the first and second components of LLE. Use the given Matlab function `plotImages.m` to do this, see Figure for an example of what this looks like for random location of images on of the number 1-3. Describe qualitatively what kind of variations is captured.
2. Repeat step 1 using the ISOMAP method. Comment on the result. Does ISOMAP do better in some way? Are the patterns being found globally based or locally based?

3. Use the Naive Bayes classifier to classify the dataset based on the projected 4-dimension representations of the LLE and ISOMAP. Train your classifier by randomly selected 70% of data, and test with remained 30%. Retrain for multiple iterations (using different random partitions of the data) and use the average accuracy of multiple runs for your analysis. Justify why your number of iterations was sufficient. Based on the average accuracies compare their performance with PCA and LDA. Discuss the result.

## IV. Feature Selection (for dataset C)

Reduce the number of features by filter-based and wrapper-based feature selection methods. Data need to be normalized (min-max) before further processing. Experiment and report on the following tasks:

1. Using Sequential Forward Selection (SFS) strategy and the sum of squared Euclidean distances as an objective function, implement a filter feature selection to select 8 features. Report the selected feature subset.
2. Using the Naive Bayes classifier as the objective function, realize a wrapper based feature selection with SFS search strategy (to select 8 features). Report the selected feature subset.
3. Use the same objective function as (2) and implement Sequential Backward Selection (SBS) strategy to select 8 features. Report the selected feature subset.
4. Using the Naive Bayes classifier to classify the data set using the selected feature subsets obtained above in (1), (2), (3), and the case that uses all 21 features. Follow the same policy as indicated in section III.3 to do your experiment and divide data in train and test. Report the average accuracy and run time in each case and discuss the results.

## Deliverables

For submitting your assignment please consider the following notes:

- Submit all of your work as one compressed file (.zip, .rar) named as Gx.zip or Gx.rar where “x” indicate your group number. (You will be able to see your group number on LEARN after you submit your group members to mcrowley@uwaterloo.ca and r26mohammadinasiri@uwaterloo.ca by Jan 25).
- Your compressed file should have all code, images, etc in addition to your report’s document.
- Write a technical document as your report and submit its PDF format included it in your compressed file.
- Your report (.pdf file) should have the name and student number of all members of your group at the beginning and separated sections for the answer of each part of each question .
- Late submissions (up to 3 days) are accepted with penalty of 10% per day.
- All code should be clearly written and commented and be runnable on another system with just the data set files beside the code in the same folder.

- Do not upload the data set files.
- One member of each group should upload the report to your group's dropbox on Learn. Each member does not need to submit same version. The last version submitted will be the one which is graded.

## Some Helpful Info:

1. Naive Bayes classifier: `PredictClass = classify(Xtest,Xtrain,Ytrain,'diaglinear');`
2. Random data split: `p = randperm(n,k)`
3. Plot images on defined coordinates: Use [plotImages.m](#)  
Example of use:

```
% X: n x d, n  number of samples
% xy_coord: n x 2
digitsImages = reshape(X', height, width, size(X,1));
scale = 0.02;
skip = 1;
plotImages(digitsImages, xy_coord(:,1:2), scale, skip);
```

Sample code and result plot is shown in Fig. 1.

4. Libraries to load and use:

**LLE:** <http://www.cs.nyu.edu/~roweis/lle/code.html>

**ISOMAP:** <http://web.mit.edu/cocosci/isomap/isomap.html>

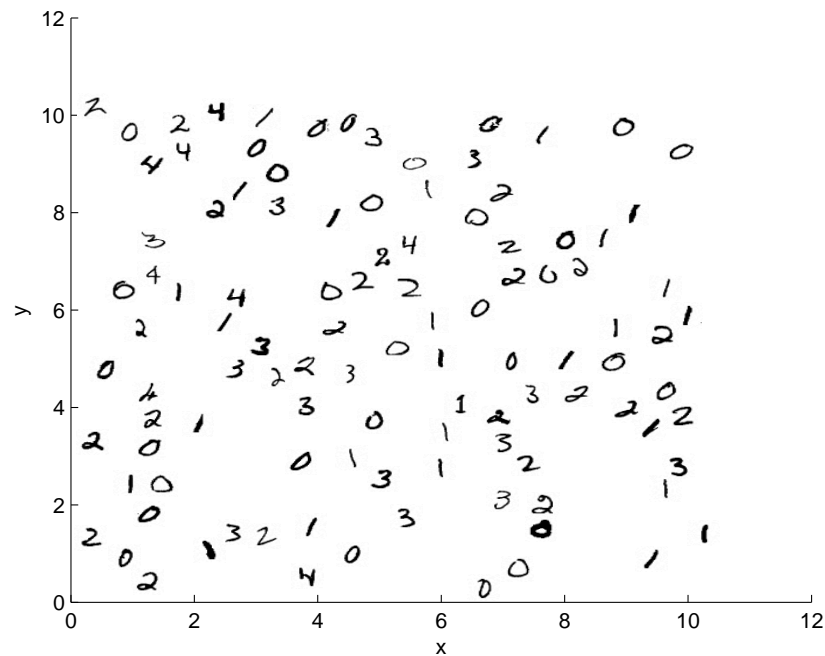
**LDA-dimensionality reduction:** <http://lvdmaaten.github.io/drtoolbox/>

```

load('DataB.mat'); % load data set
fea = fea(1:10:end,:); % select a subset just for test
nsample = size(fea,1);
width = 28; % width of images
height = 28; % height of images
xy_coord = 10 * rand(nsample,2); % some random locations to put images
digitsImages = reshape(fea', height, width, size(X,1)); % create array images
scale = 0.05; % scale images on plot
skip = 1;
plotImages(digitsImages, xy_coord(:,1:2), scale, skip);

```

(a)



(b)

Figure 1: Using plotImages to plot pictures in a sample 2D space; a) A sample code, b) Output of the sample code.