University of Waterloo
ECE 657A: Data and Knowledge Modeling and Analysis
Winter 2024
Assignment 1: Data Cleaning and Dimensionality Reduction
Due: Feb 16th, 2024, 11:59pm

Overview

Assignment Type: done in groups of up to three students. The same group which is made on LEARN.

Hand in:

One report (PDF) or python notebook per group, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. (If you are submitting by PDF, if you don't know 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 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 will be placed in the owncloud dropbox. The datasets are in csv format (.csv), if you are using python you can load them using the Scipy method loadmat.

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 are denoted by Not Available (NA) and also some outliers are present. (note: The negative values are not outliers) This data is used to illustrate the data cleaning and preprocessing techniques. (File: DataA.csv)

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.csv)

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.*

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 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.*
6. *Prove that the PCA is the best linear method for transformation (with orthonormal bases)*

III. Parameter Estimation

1. Identify whether the following parameter is biased or unbiased?

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$where\ \mu = mean\ ; n = Total\ number\ of\ samples$$

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, if you have any question please contact Josh Sun (q84sun@uwaterloo.ca))
- 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.
- There are no accepted Late submissions.
- 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.