

COMP4702/COMP7703 - Machine Learning

Prac 4 – Dimensionality Reduction using Principal Component Analysis

Aims:

- To complement lecture material in understanding the operation of PCA.
- To gain experience with simulating and implementing these techniques in software.
- To produce some assessable work for this subject.

Procedure:

Principal Component Analysis: PCA can be implemented very simply in Matlab. Given a dataset (as a matrix X), the covariance matrix can be found using the Matlab covariance function ($\text{cov}(X)$). Then, the eigenvectors and eigenvalues of this covariance matrix are the principal component (vectors) and principal values respectively. The eigenvalues reflect the amount of variance accounted for by each principal component and are ordered. To perform dimensionality reduction (e.g. down to 2 dimensions), we need to multiply X by the two eigenvectors with the largest corresponding eigenvalues.

- **Q1:** Write a matlab function implementing PCA. You can simply have your function output the eigenvectors and eigenvalues, or it could take as an argument the number of principal components to be used for dimensionality reduction, perform the projection and output the resulting (reduced) dataset (as well as the eigenvectors and eigenvalues).
- **Q2:** Load the iris dataset, and run your PCA function on the data. What percentage of the data variance is accounted for by the first two principal components? Produce a plot of the data in the space spanned by the first two principal components.
- **Q3:** Run PCA on the diabetes dataset. From the results, produce a Scree graph similar to that shown in Fig 6.2 of the text (slide 10 of the notes for Chap. 6). Also, produce a plot of the data in the space spanned by the first two principal components.

A well-known database for handwritten digits in machine learning is known as the MNIST database (see course webpage for link). A pre-processed version of this data is available from (<http://cis.jhu.edu/~sachin/digit/digit.html>). You can find this data on the course webpage.

- **Q4:** Run your PCA function on the MNIST data and produce:
 - (a) A Scree graph
 - (b) A plot of the data in the space spanned by the first three principal components (using the matlab `plot3()` function)

→ (c) Pictures of the first 8 eigenimages of the data.

Some hints for Question 4:

- The webpage mentioned above describes how to read the data into matlab and display an image. Based on this, here are some example commands:

```
% open the file corresponding to digit 8
fid=fopen('data8','r');
% read in first data point and store in 28x28 matrix t1
t1=fread(fid,[28 28]);
% read in first data point and store in 28x28 matrix t2
t2=fread(fid,[28 28]);
% Set colourmap for grayscale
colormap(gray(256))
% Display first image - need to transpose the matrix t1
image(t1')
```

- I suggest writing a matlab script file to read in all the data files and to get the data into one matrix so that you can then call your pca matlab function.
- You may find the matlab reshape() function useful, to arrange data into a matrix. See the matlab Help for more information.

Appendix

The diabetes dataset is from the UCI repository – see link on course webpage. Some details on this dataset:

1. Diabetes Data set

Input features = 8

No. of classes = 2

No. of samples= 768

Training samples = 691 (Class 1 samples=449, Class 2 samples=242)

Test samples = 77 (Class 1 samples= 51, Class 2 samples= 26)

More information about the data set:

- Sources:
 - (a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases
 - (b) Donor of database: Vincent Sigillito (vgs@aplcn.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory, The Johns Hopkins University
Johns Hopkins Road, Laurel, MD 20707, (301) 953-6231
 - (c) Date received: 9 May 1990
- Relevant Information:
Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
- For Each Attribute: (all numeric-valued)
 1. Number of times pregnant
 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
 3. Diastolic blood pressure (mm Hg)
 4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (μ U/ml)
 6. Body mass index (weight in kg/(height in m)²)
 7. Diabetes pedigree function
 8. Age (years)
 9. Class variable (1 or 2)
- Missing Attribute Values: None