CS423 Data Mining: Assignment 2

Due date: 11 Oct 2016

Instruction

In this assignment, you will empirically study the characteristics of one of the most widely used dimensionality reduction technique: Principle Component Analysis (PCA). This is a computer-based assignment. You will need a computer and software of your choice (Scilab is preferred) to complete this assignment.

1 Computational complexity

You will study the running time of PCA given by
$\mathcal{O}(nm^2) + \mathcal{O}(time\ to\ find\ eigenvectors)$, where n is the number of data points and m is the dimensionality.

Question 1.1

Generate data randomly by fixing $n = 1024$ and vary the dimensionality $m = 2^i$ where $i = 1$ to 10. Measure the time to find all the principle components of these data. Plot and discuss your result.

Question 1.2

Generate data randomly by fixing $m = 2^5$ and vary the dimensionality $n = 2^i$ where $i = 10$ to 15. Measure the time to find all the principle components of these data. Plot and discuss your result.

Useful function: `cov()`, `spec()`, `rand()` and `tic(),toc()` for measuring the elapsed time.

2 Principle component selection

One of the main application of PCA is for preprocessing the data before performing a classification, we will now study the classification accuracy of a k-NN classifier on a biomedical dataset (`colon-cancer`, see the website). The data consists of 2 variables x and y, where x is an $n \times m$ data matrix and y is a column vector representing labels of the data.

Question 2.1

Use the given k-NN code to classify the given high-dimensional dataset.

- First, perform the classification task without using the dimensionality reduction technique, record the classification error.
- Secondly, use the rule,

$$\frac{\sum_{i=k+1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

  to find a good subspace in which you will project the data into. Report the dimensionality of the subspace selected and the classification error of k-NN in such space.
- Discuss your results.

Tips: you only need to call the function `myknn(x,y)` where the first parameter is the data matrix (or the reduced data matrix) and y is a set of labels.

3 Submission

Send the report and the codes to `jakramate.b@cmu.ac.th`