CS 383 - Machine Learning

Assignment 1 - Dimensionality Reduction
Winter 2017

Introduction

In this assignment you'll work on visualizing data and reducing its dimensionality.

You may not use any function from the Matlab ML library in your code. Look at the *Matlab Functions* section on Blackboard for a list of functions that are ok to use.

In particular for this assignment you **MAY NOT** use Matlab functions like:

- pca
- entropy

But you **MAY** use basic statistical functions like:

- std
- mean
- cov
- eig

As a reminder, make sure to clear out old variables prior to running your script.

Grading

Although all assignments will be weighed equally in computing your homework grade, below is the grading rubric we will use for this assignment:

Part 1 (Theory)	23pts
Part 2 (PCA)	30pts
Report	5pts
TOTAL	58pts

Table 1: Grading Rubric

DataSets

Pima Indians Diabetes Data Set In this dataset of 768 instances of testing Pima Indians for diabetes each row has the following information

1. Class Label (-1=negative,+1=positive)
2. Number of times pregnant
3. Plasma glucose concentration
4. Diastolic blood pressure (mm Hg)
5. Triceps skin fold thinkness (mm)
6. Insulin (mu U/ml)
7. Body mass index (kg/m^2)
8. Diabetes pedigree function
9. Age (yrs)

Data obtained from: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

1 Theory Questions

1. Consider the following data:

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

   (a) Find the principle components of the data (you must show the math, including how you compute the eivenvectors and eigenvalues). Make sure you standardize the data first and that your principle components are normalized to be unit length. As for the amount of detail needed in your work imagine that you were working on paper with a basic calculator. Show me whatever you would be writing on that paper. (5pts).

   (b) Project the data onto the principal component corresponding to the largest eigenvalue found in the previous part (3pts).

2. Consider the following data:

$$\text{Class 1} = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix}, \text{Class 2} = \begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

   (a) Compute the information gain for each feature. You could standardize the data overall, although it won't make a difference. (5pts).

   (b) Which feature is more discriminating based on results in part a (1pt)?

   (c) Using LDA, find the direction of projection (you must show the math, however for this one you don't have to show the computation for finding the eigenvalues and eigenvectors). Normalize this vector to be unit length (5pts).

   (d) Project the data onto the principal component found in the previous part (3pts).

   (e) Does the projection you performed in the previous part seem to provide good class separation? Why or why not (1pt)?

2 Dimensionality Reduction via PCA

Download the dataset *diabetes.csv* from Blackboard. This dataset has eight features ($D = 8$) and 768 samples ($N = 78$). The first column is the class label {-1, 1}. **However** your script should be able to work on any dataset that lacks a header row and then has an arbitrary number of data observations, N, one per row, in the format:

$$(y_i, x_{i,1}, x_{i,2}, ..., x_{i,D})$$

where $x_{i,j}$ are real valued numbers, $r_i \in \{-1, 1\}$, and D is the number of features.

Write a script that:

1. Reads in the data
2. Standardizes the data (except for the first column of course)
3. Reduces data (except for the first column of course) to 2D using PCA
4. Graphs the data for visualization
   (a) Even though we're not using class labels to do the dimensionality reduction, plot the -1 data in blue and the +1 data in red

Your graph should end up looking similar to Figure 1.

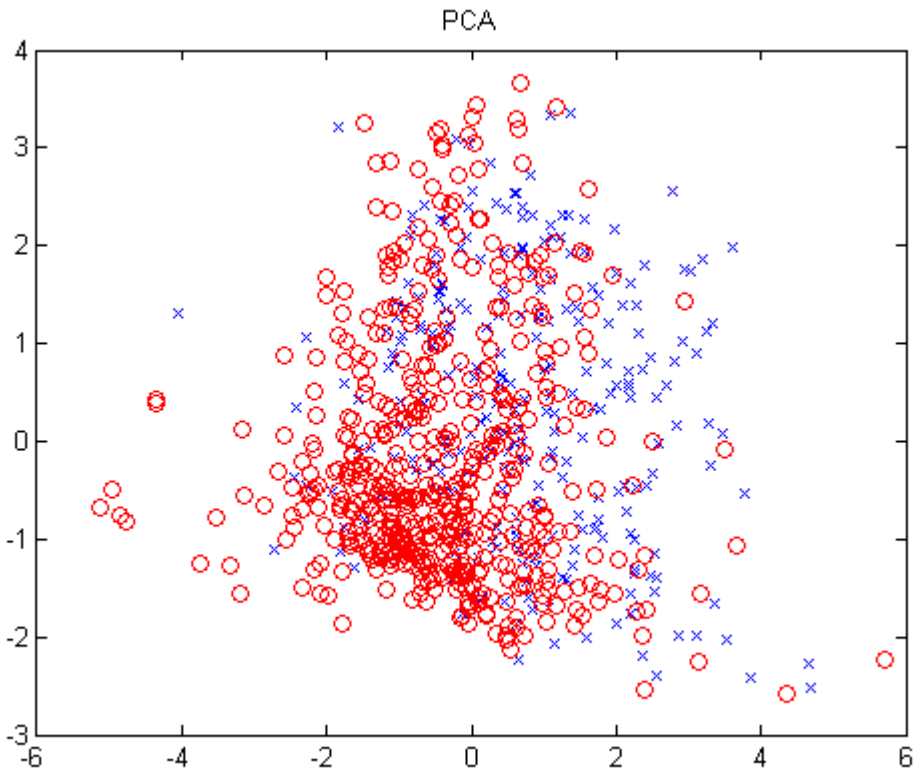

Figure 1: 2D PCA Projection of data

Submission

For your submission, upload to Blackboard a single zip file containing:

1. PDF Writeup
2. Source Code
3. readme.txt file

The readme.txt file should contain information on how to run your code to reproduce results for each part of the assignment.

The PDF document should contain the following:

1. Part 1: Your answers to the theory questions.
2. Part 2: The visualization of the PCA result