Homework 3: Principal Component Analysis

About

Due

Monday 2/11/19, 11:59 PM CST

Goal

The goal of this homework is to use PCA to smooth the noise in the provided data.

Submission

Submission will be through gradescope (https://www.gradescope.com):

Code and External Libraries

The assignment can be done using any programming language.

You may use a PCA package if you so choose but remember you need to understand what comes out of the package to get the homework right!

Problems

Total points: 100

At here (./hw3-data.zip), you will find a five noisy versions of the Iris dataset, and a noiseless version. For each of the 5 noisy data sets, you should compute the principal components in two ways. In the first, you will use the mean and covariance matrix of the noiseless dataset. In the second, you will use the mean and covariance of the respective noisy datasets. Based on these components, you should compute the mean squared error between the noiseless version of the dataset and each of a PCA representation using 0 (i.e. every data item is represented by the mean), 1, 2, 3, and 4 principal components. The mean squared error here should compute the sum of the squared errors over the features and compute the mean of this over the rows. For example, if the noiseless version has two rows [1,2,3,4] and [0,0,0,0] and the reconstructed version is [1,2,3,0] and [1,1,1,1] the MSE would be (16 + 4) / 2 = 10

. You should produce:

- A csv file showing your numbers filled in a table set out as below, where "N" columns represents the components calculated via the noiseless dataset and the "c" columns of the noisy datasets.
    - Example: The entry corresponding to Dataset I and 2N should contain the mean squared error between the noiseless version of the dataset and the PCA representation of Dataset I, using 2 principal components computed from the mean and covariance matrix of the noiseless dataset.
    - *Update (for clarity of instructions): In all cases you compare the reconstruction with the noiseless dataset to get the MSE.*
      The first part, with "N" columns asks to reconstruct the noisy datasets using the PCs of the noiseless dataset.
      The second part, with "c" columns asks to reconstruct the noisy datasets using the PCs of the noisy dataset.
- A csv file containing your reconstruction of Dataset I ("dataI.csv"), expanded onto 2 principal components, where mean and principal components are computed from Dataset I.

Number of PCs	**0N**	**1N**	**2N**	**3N**	**4N**	**0c**	**1c**	**2c**	**3c**	**4c**
Dataset I										
Dataset II										
Dataset III										
Dataset IV										
Dataset V										

Submission

Submit to Gradescope:

1. **(50 points)** A CSV file containing your numbers for the table. The CSV file should be named "{yournetid}-numbers.csv" (e.g., taw-numbers.csv). The first line should read "0N, 1N, 2N, 3N, 4N, 0c, 1c, 2c, 3c, 4c" (see the previous section for what it should look like). The following lines should be the rows of the table, in order, and contain only numbers. You should provide your numbers to at least three digits.
2. **(45 points)** The reconstruction of Dataset I as a CSV file in the same format as the datasets. The CSV file should be named "{yournetid}-recon.csv" (e.g., taw-recon.csv). Each row is your reconstructed version of the data item. The first line should be a header reading: "Sepal.Length","Sepal.Width","Petal.Length","Petal.Width" Each following line is your reconstruction of a data item, in order (so first data item first, etc).
3. **(5 points)** Your source code. It can be a .R file if you use R, or a .py file if you use Python.