Homework 3

Sunday, April 30, 2017 8:33 PM

PCA

Logistics

- Submit your solutions on Canvas, this applies to each and all sections:
  - Code in one or more Matlab (.m) files
  - Report and figures as a single MS Word or pdf file
  - Put everything into a single ZIP file
- You must use Octave or Matlab code/instructions to obtain the results
  - Submit any scripts you used
  - You can use the code provided in the PCA in-class exercise
- Besides just finding the results, you must discuss your findings, explain the why's and hows, and give reasonable interpretations.
- Don't forget to include references in your report!
- List any assumptions when needed to disambiguate

Goal

Practice PCA and interpret results. In this homework we will use PCA to investigate potential clustering in the data.

Dataset

The following dataset is also commonly used to understand and explore the potential and application of PCA.

wine_beer_...

This is a csv file, but you can open it like a txt file.

The "Wine Beer" dataset, contains the following information regarding alcohol consumption in variuos countries and a couple of health-concerning metrics:

	Liquor consumpton liter/year	wine consumption liter/year	beer consumption liter/year	life expectancy years	hear disease rate cases/10e5/year
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58	25.1	78	94.1
Switzerland	1.7	46	65	78	106.4
Australia	1.2	15.7	102.1	78	173
Great Britan	1.5	12.2	100	77	199.7
United States	2	8.9	87.8	76	176
Russa	3.8	2.7	17.1	69	373.6
Czech Republic	1	1.7	140	73	283.7
Japan	2.1	1	55	79	34.7

Mexico	0.8	0.2	50.4	73	36.4

There is only one clas in this da taset. Each country is an observation, the columns are the features. We will use PCA to see if the data clusters in any way.

Part 1 Initial Statistics

As usual, obtain relevant statistics: means and standard deviations for each of the features.
Obtain covariance matrix
Do a summary of 2D scatter plots like those in Homework 2 Part 0 using the original da ta points (observations)

Show and discuss results in your report.

Part 2 Preprocessing

- Preprocess your data like in Homework 2 (mean center and scale).
- Discuss whether there are any "useless" or "noisy" features

Part 3 SVD

Perform SVD on the preprocessed data. Do not use the built-in PCA function, use th e SVD function

Obtain the Scree plots and discuss the r esults: identify relevant principal components (dimensions)
Obtain and plot loading v ectors, discuss results: what features are positively/negatively correlated and why? Are there any redundant features and why? What features are relevant and which ones seem irr elevant and why?
Based on the relevant principal components (new dimensions) iden tified with the help of the scree plot, create 2D scatter plots and discuss clusterings and loc ation of the da ta points in the original and transformed spaces

Rubric

Criteria	Points	Notes
Code	30p	Code should be well commented and structured Must include all the following implementations: • Center and scale: 2pt • Svd: 2:pt • Scree plots • Loading Vectors • Scatter: before and after transformation NOT: Must NOT use Matlab's pre-built PCA function
Analysis	Outstanding: 70pt Sufficient: 65pt Poor: 40pt Missing: 0pt	Include and discuss statistics and plots. • The discussion must include **interpretations** for each of the following: • Scree Plots: e.g., discussion of how many components are useful • Loading Vectors: should have four plots (one per vector), discussion about relevant /irrelevant features, correlations, etc. • Scatter: 2D plots in summary form, for data before and after svd Discuss your results, putting them in context of the results (plots and statistics) Focus on the WHYs not just the What's. Remember, your interpretation of the results is what matters the most!

		Points will be deducted for lousy redaction, spelling/grammar mistakes, typos -- professionalism is expected.