Analytics for Observational Data (IT142IU)

Lab 3-4: PCA

## Objectives

* Calculate covariances and correlations given datasets.
* Apply PCA to select features.
* Dataset sources:
  + Mid-term.dataset.csv (provided on the Blackboard)
  + <https://archive.ics.uci.edu/ml/datasets/climate+model+simulation+crashes>
* Programming languages: Python/Java

## Analyzing the data

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | Mid-term.dataset.csv |
| Correlation before PCA |  |
| Covariance before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Remarks | After PCA correlation between data is reduce to approximately 0 |

***Part 2.***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | climate+model+simulation+crashes  using columns 3-20 |
| Correlation matrix before PCA |  |
| Covariance matrix before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA using **5** components |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| New data after PCA using **10** components |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Plot the scores for the first two components. What do you notice? Investigate the outliers, and the raw data for each of these unusual observations. What do you conclude about those observations? | At first, the two data have nothing in correlation, one column is an outlier of another columns, after PCA 2 data seem fits to each others. |
| Exclude the unusual observations and refit the PCA model. |  |
| Remarks | PCA transformation helps us to maximize the variance thus reducing dimension of data, from n columns to 2 columns, speed up machine learning process |

***Part 3***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | climate+model+simulation+crashes  using columns 3-20 |
| Use sklearn.decomposition  to build a PCA model on all the data |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Compare with the above results and evaluate | There are several difference in data between PCA steps by steps and PCA with sklearn library. Column 0 and 2 in PCA data using library show the opposite value when compare with data using PCA step by step. This was due to eigenvalues and eigenvector in linalg().eig() has go through normalization for convenience during calculation. The data between 2 PCA do not have significant differences, therefore covariance and correlation matrix is roughly the same. |
| Plot the scores for the first two components. What do you notice? Investigate the outliers, and the raw data for each of these unusual observations. What do you conclude about those observations? |  |
| Exclude the unusual observations and refit the PCA model. |  |
| Remarks |  |

## Some references

* PCA: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
* Standardization: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
* PCA in Python: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
* Learn more about PCA: <https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/pca-exercises>