CSC3022H: Machine Learning

Assignment 5:

Principal Component Analysis (PCA)

Department of Computer Science
University of Cape Town

Due: Friday, 15th May, 2020, 10.00 AM

Problem Description

Figure 1 illustrates a scatter plot of 64 pairs of data-points for 2 variables – that is, average rainfall (mm) in July and January for 64 selected places. See attached text file of raw data: *2018-AvgRainfall(mm)*.

Implement (in C++) a PCA algorithm [Lever et al., 2017], [Smith, 2002], to find the covariance matrix and *two* (2) principal components of this data-set. Results should answer the following questions:

1. What are the Eigenvalues for the principal components 1 and 2?
2. What are the Eigenvectors for the principal components 1 and 2 (showing July and January component values for each)?
3. Compute the values for the covariance matrix.
4. What is the total variance?
5. What proportion (as a percentage) of total variance do principal components 1 and 2 ”explain”?

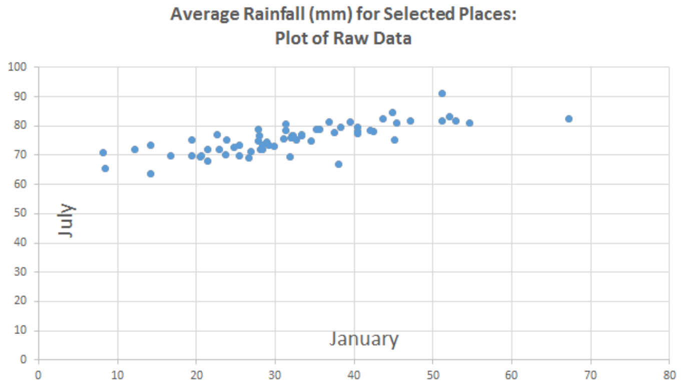

Figure 1: Average rainfall (mm) for selected places in January and July, 2018.

In a ZIP file, place the source code, makefile, and output text file (answers to questions 1 – 5).

Upload the ZIP file to Vula by 10.00 AM, Friday 15th of May.

References

[Lever et al., 2017] Lever, J., Krzywinski, M., and Altman, N. (2017). Points of significance: Principal component analysis. *Nature Methods*, 14(1):641–642.

[Smith, 2002] Smith, L. (2002). *A tutorial on Principal Components Analysis.* On Vula.