Research Methodology for Data Science CS4125

Assignment C

This assignment is concerned with the evaluation and implementation in MATLAB of methods that were discussed in the lectures. To solve the Task 1, 2, and 3 you only need MATLAB's matrix data structures, linear algebra operations (such as matrix multiplication, transposition and the computation of eigenvalues and eigenvector) and its visualization capabilities.

Task 1: *Principal Component Analysis (PCA)*

Implement the two algorithms for computing PCA modes, which we discussed in the lecture. The first algorithm uses the covariance matrix and the second algorithm uses the Gram matrix. Test your algorithms on at least two different high-dimensional data sets. Discuss and visualize the resulting PCA modes and discuss differences of the performance of the two computational schemes (e.g. computation time, storage requirements).

Task 2: *Fast Approximation of PCA*

Implement the two algorithms for fast PCA approximation we discussed, the snapshot PCA and the Nyström method. Compare the performance and accuracy of the two approximation algorithms on test examples and discuss the drawbacks and benefits of the methods you experience in your experiments. Make a recommendation on what method to use in praxis (you can distinguish between different scenarios).

Task 3: *Multidimensional Scaling (MDS)*

Implement the (classical) MDS algorithm discussed in the lecture. For at least two data sets, compute the dissimilarity matrix (a data set can also just be a similarity matrix or a dissimilarity matrix). Compute the corresponding MDS embeddings to a low-dimensional space. Visualize and discuss the results and report eigenvalues and stresses of the embeddings.

Task 4: *Stochastic Neighbor Embedding (SNE)*

Use the Matlab's tSNE function to compute embeddings for at least two data sets (you can re-use data sets from the other tasks). Compute also MDS embeddings for these data sets and discuss differences between the embeddings computed with MDS and tSNE.

Task 5: *Report*

Write a report that describes and illustrates

- the algorithms implemented
- the tests for correctness of the implemented algorithms you performed
- the data sets you used for your experiments and the steps you took to prepare the data such that they can serve as input to your implementation
- the results of your experiments and your conclusions drawn from your experiments

In addition, you can report on the division of labor amongst the group members.

Required deliverables on Brightspace:

- For Tasks 1, 2, 3, and 4 provide the MATLAB notebooks (.m files) and example data sets. Pack all the files in one ZIP archive.
- The report should be one PDF file

Deadlines: Draft submission June 9, 2021; final submission June 18, 2021.