University of Science,
Vietnam National University

Faculty of Information Technology
Computer Science Department

CSC10108
Data Visualization

Lab Instructor:

ILab 03: T-distributed Stochastic Neighbor Embedding Visualization

Revision History

Revision	Date	Author(s)	Description
1.0	May 6, 2023	LN Nam	Initial, release first version in English.

Contents

1 Learning Objectives

In this assignment, we will explore an impactful data visualization technique called "t-SNE" (t-Distributed Stochastic Neighbor Embedding), a variation of Stochastic Neighbor Embedding [1] that offers significant advantages. t-SNE excels in optimizing visualizations by effectively reducing crowding tendencies and producing clearer representations of high-dimensional data [2]. Compared to other methods, it offers smoother optimization processes, enhancing its utility in practical applications of data visualization.

2 Notes and Constraints

Note the following:
List of constraints when doing this lab:

- Work without a report will not be graded.
- Members who do not contribute to the project will not receive points.
- Reference sources (if any) need to be fully recorded in the report in the References section. Note that it is necessary to distinguish between referencing and plagiarism.
- Individuals or groups that commit cheating and dishonesty will receive 0 points in the course.
- Name the assignment category MSSV_ILab03, with MSSV being the student number, compress the entire submission into 1 file before submitting. If the size is > 20MB, upload it to an external storage service such as Google Drive or OneDrive, and then submit the link. Last but not least, please keep the link public for at least 2 years.

3 Problem Statement

For this assignment, your task is to delve into the intricacies of t-Distributed Stochastic Neighbor Embedding, commonly abbreviated as t-SNE. To kickstart your exploration, it's recommended to peruse the paper "*Visualizing Data using t-SNE*" authored by **Laurens van der Maaten**. Once you've gathered ample insights, your next step entails crafting a comprehensive report encapsulating your understanding of t-SNE. Your report should span a *minimum of five pages*, delving into the theoretical foundations, optimization techniques, practical applications, and comparative analysis with other dimensionality reduction methods. Through meticulous research and thoughtful analysis, y ou'll elucidate the nuances of t-SNE, demonstrating a thorough grasp of its principles and implications in data visualization.

For demonstrations, your task involves implementing t-SNE using libraries like scikit-learn on the dataset utilized in the previous PCA lab. Through this implementation, you will elucidate the *disparities between t-SNE and PCA*, showcasing their respective strengths and weaknesses. By comparing the visualizations generated by both techniques, you'll gain valuable insights into how they handle dimensionality reduction differently and discern their impact on data representation and interpretation.

4 Limitations

You can leverage the support of thirdparty libraries such as scikit-learn to use PCA, and t-SNE. Extra points for impressive demonstrations.

5 Evaluation Criteria

Your assignment will be evaluated based on the following criteria:

Criteria	**Mark**
Studying t-SNE	45%
Implementation t-SNE	30%
Making comparison between t-SNE and PCA.	15%
Overall comprehension of the submitted source code.	10%
Bonus points	10%
Total	**110%**

6 What to Submit

You have to submit these things to complete your assignment:

- ❑ **Docs Folder**: This folder should contain your report files in .doc, .docx, or .pdf formats, with a strong recommendation for .pdf to ensure compatibility and preservation of formatting. Your report should cover several key areas:
  - **Your Information**: Include your group name and student IDs.
  - **Requirement Fulfillment**: Discuss how fully each project requirement has been met.
  - **Study algorithm**: Provide thorough explanations of the algorithms used, include running examples, and offer commentary on the code.
  - **Presentation Style**: Aim for clarity in your report, using illustrations where helpful to convey your points effectively.
- ❑ **Source_Codes Folder**: This directory should house all the source code for your project. It will primarily contain Jupyter notebooks and Python scripts developed by yourself. If you have code in languages other than Python, please include clear instructions for its use.

By organizing your submission in this manner, you'll help ensure that your work is clearly presented, easily navigable, and thoroughly documented, reflecting the depth and breadth of your project efforts.

References

[1] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[2] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.