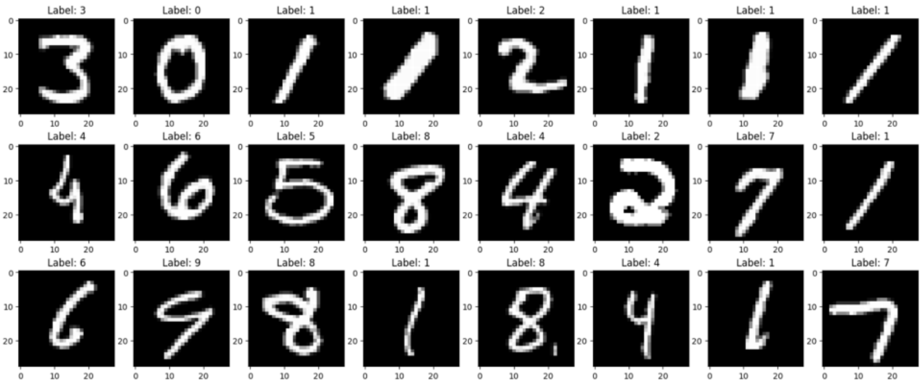
Assignment 10/3: t-SNE (Homework)

Colab file: Week10_t-SNE_Homework - Colab

Dataset: MNIST_784

Sample Data:



t-SNE Visualization of MNIST with KMeans Evaluation

explores the use of t-SNE for visualizing the MNIST dataset and evaluates different perplexity values using KMeans clustering.

Process Overview

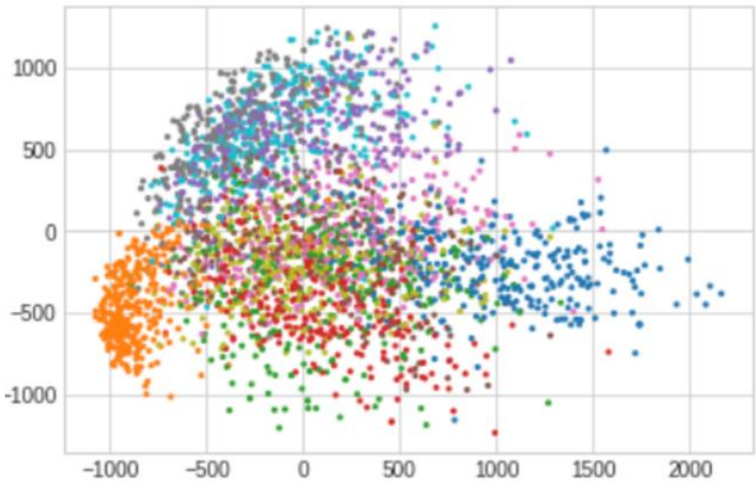
1. **Data Loading and Preprocessing:**
    - The MNIST dataset is loaded using *fetch_openml.*
    - Two alternatives are explored:
        - **Alternative 1:** Standardize the data using StandardScaler before applying t-SNE.
        - **Alternative 2:** Do not standardize the data and apply t-SNE directly.
2. **t-SNE Parameter Tuning (Perplexity):**
    - A loop iterates through different perplexity values (defined in perplexity_values).
    - For each perplexity:
        - **t-SNE Transformation:** t-SNE is applied to the data with the current perplexity, learning_rate = 1000, max_iter = 2500, and random_state = 42.
        - **KMeans Clustering:** KMeans clustering

(with n_clusters = 10 and random_state = 42) is performed on the t-SNE reduced data.

- **Inertia Evaluation:** The KMeans inertia is calculated as a metric to evaluate the quality of the clustering for the given perplexity. Lower inertia generally indicates better clustering.
- **Visualization:** A scatter plot is generated to visualize the t-SNE transformed data, colored by the true labels (digits 0-9).

- The perplexity that results in the lowest KMeans inertia is chosen as the best perplexity.

3. **Final Visualization with Best Perplexity:**
    - Using the best perplexity found, t-SNE is applied again to the data.
    - The final visualization of the MNIST dataset in 2D space is created with the best perplexity.
4. **Analysis of Perplexity Impact:**
    - A plot shows the relationship between perplexity and the KMeans inertia.
5. **Comparison of Alternatives:**
    - The inertia of the best perplexity for Alternative 1 and the inertia for Alternative 2 are compared.



This picture is shown as Naively using PCA

Hyperparameters

- **t-SNE:**
  - n_components = 2
  - learning_rate = 1000
  - max_iter = 2500
  - random_state = 42
- **Perplexity:** Iterated through perplexity_values =
  [20, 21, 22, 23, 24, 25, 28, 30, 31, 32, 33, 34, 35, 40, 41, 42, 43, 44, 45, 50, 60].
  The best perplexity is selected based on minimizing the KMeans inertia.
- **KMeans:** n_clusters = 10, random_state = 42

Note:

- ***learning_rate:*** If the learning rate is too high, the data may look like a 'ball' with any point approximately equidistant from its nearest neighbours. If the learning rate is too low, most points may look compressed in a dense cloud with few outliers.
- ***max_iter:*** Maximum number of iterations for the optimization.

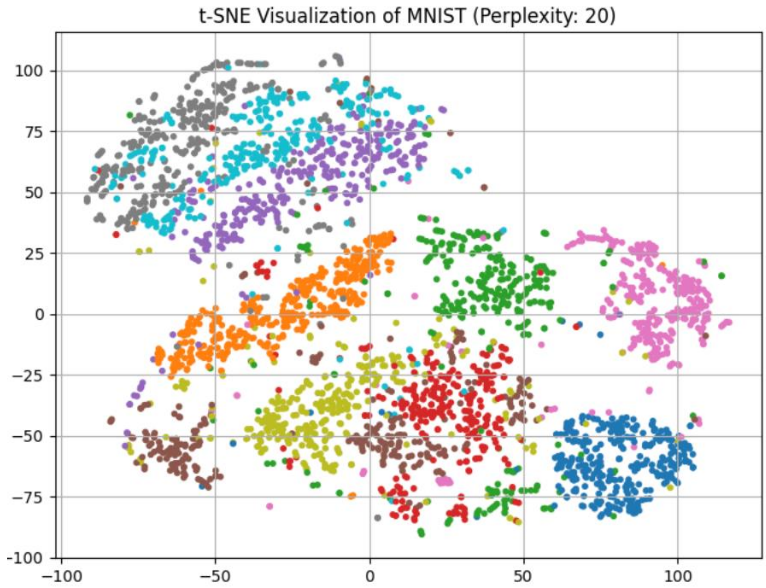
Tuning Procedure

1. **Iterate through Perplexity Values:** The code iterates through a range of perplexity values to identify the one that produces the best visualization and clustering.
2. **Evaluate with KMeans:** For each perplexity value, t-SNE is applied to the data, and the resulting 2D representation is clustered using KMeans.
3. **Minimize KMeans Inertia:** The KMeans inertia is used as a metric to evaluate the clustering quality. The perplexity that minimizes inertia is considered the best choice.
4. **Visualization:** Scatter plots are generated for each perplexity to visually assess the clustering quality.
5. **Final Visualization:** The best perplexity is used to create the final t-SNE visualization of the MNIST dataset.

Results

The code produces the following outputs:

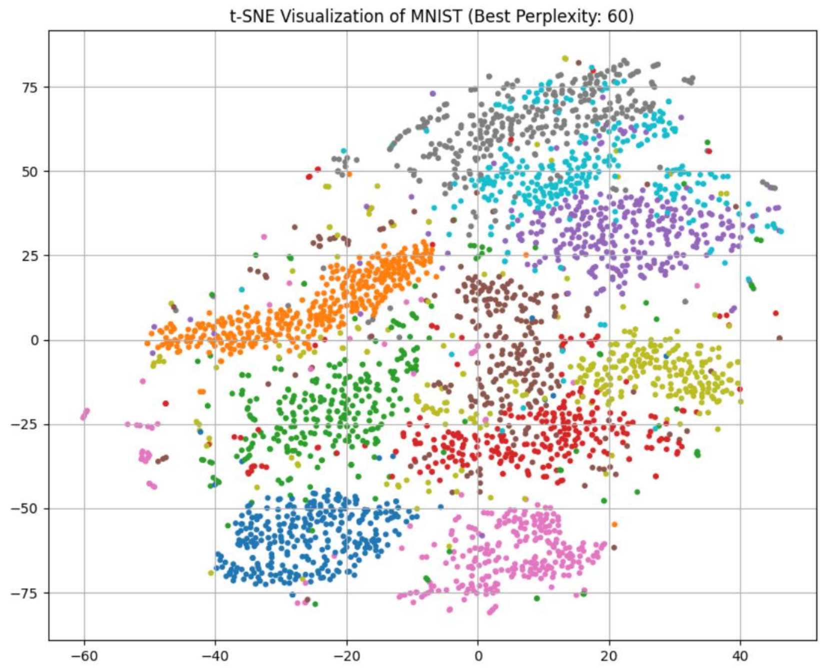
- **Scatter Plots:** A series of scatter plots showing the 2D representation of the MNIST data after t-SNE transformation for each perplexity value.



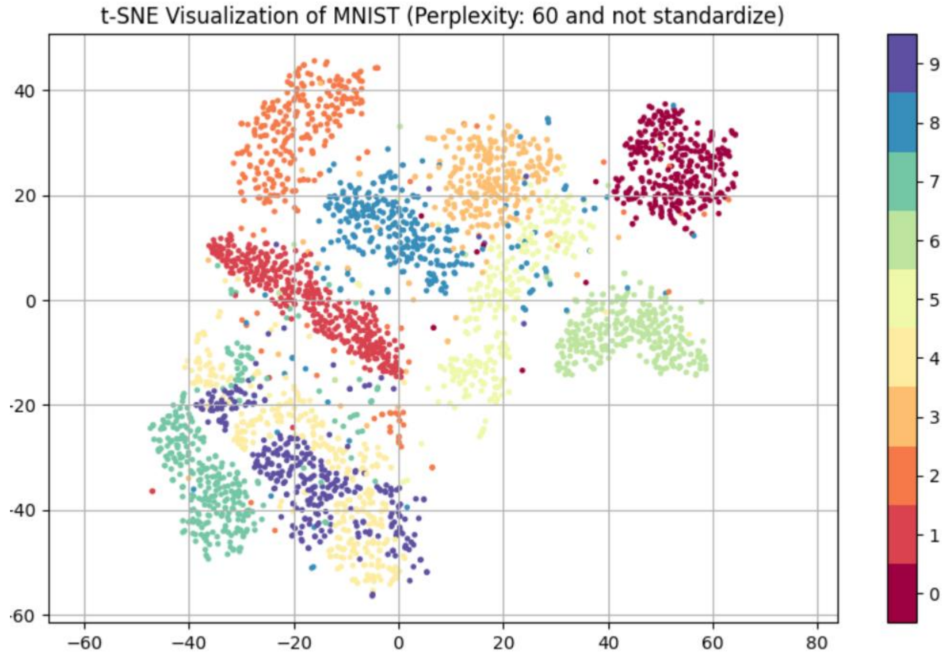
For example, This is the reduced data which used perplexity is equaled 20.

- **Best Perplexity:** The perplexity value that minimizes the KMeans inertia is reported.
- **Final Visualization:** A scatter plot using the best perplexity showing the 2D representation of the MNIST data.

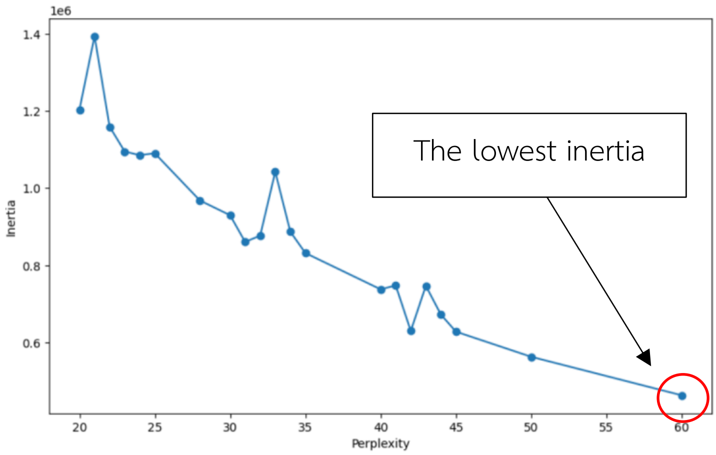
Alternative 1:



Alternative 2:



- **Perplexity vs. Inertia Plot:** A plot demonstrating how the KMeans inertia changes with different perplexity values.



- **Comparison of Alternatives:** The inertia for both alternatives is reported to determine which approach provides better results.

Alternatives	Inertia (KMean, k = 10)
1	462944.0
2	296208.67

Finally, the alternative 2, which is not standardize data is the better solution to t-SNE. So choose the **alternative 2** (Not standardize, and perplexity = 60).

How to achieve beautiful results?

The good results are achieved by carefully tuning the t-SNE hyperparameters, specifically the perplexity. The perplexity parameter controls the local neighborhood size used by t-SNE. By **iterating through different perplexity values and evaluating the resulting clustering quality with KMeans inertia**, we are able to find a perplexity that optimally balances local and global structure in the data.

Additionally, exploring alternative approaches like **standardizing or not standardizing** the data before applying t-SNE can influence the visualization and clustering performance, and it is important to compare these approaches.

By utilizing the combination of t-SNE, KMeans and evaluating the results visually, we are able to obtain an effective and insightful visualization of the complex structure within the MNIST dataset.