PROBLEM SET

PROBLEM SET DESCRIPTION

This problem set aims to provide an in-depth understanding of dimensionality reduction techniques and their applications on real-world datasets. The focus is on implementing and comparing the performance of t-SNE with other widely used methods such as PCA, Isomap, and LLE. The t-SNE will be implemented on the Fashion MNIST dataset, which consists of images of clothing items, and is a widely used benchmark dataset in computer vision and machine learning.

1) Data Preparation:

Load the Fashion MNIST dataset using TensorFlow and preprocess using NumPy. The pixel values should be normalized and the images should be flattened to achieve faster results. Additionally, standardization of the dataset is recommended to ensure better performance of the algorithms.

2) t-SNE Implementation:

Implement the t-SNE algorithm in Python using only NumPy and matplotlib/seaborn. Start with a perplexity of 30 and a learning rate of 200 for the optimization. Use the following algorithm:

Algorithm 1: Simple version of t-SNE

Data: $X = \{x_1, x_2, ..., x_n\}$,
cost function parameter: perplexity $Perp$,
optimization parameters: iterations T, learning rate η, momentum $\alpha(T)$.
Result: $\Upsilon^{(T)} = y_1, y_2, ..., y_n$.
Use binary search to find var σ_i for each datapoint x_i which produces P_i with the given fixed perplexity $Perp$.
Compute pairwise affinities $p_{j|i}$ with σ_i
set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
sample initial solution $\Upsilon^{(0)} = y_1, y_2, ..., y_n$ from $(0, 10^{-4})I$
for $t = 1$ *to* T **do**
    compute low-dimensional affinities q_{ij}
    compute gradient $\frac{C}{\Upsilon}$
    set $\Upsilon^{(t)} = \Upsilon^{(t-1)} + \eta \frac{\partial C}{\partial \Upsilon} + \alpha(t)(\Upsilon^{(t-1)} - \Upsilon^{(t-2)})$
end

3) Hyperparameter Tuning:

Experiment with different perplexity values (e.g., 5, 10, 50) and learning rates (e.g., 10, 100, 500) to see how they affect the quality of the t-SNE visualization. Visualize the results using Matplotlib and compare the different parameter settings.

4) Visualization:

Use the t-SNE algorithm with the best hyperparameters to visualize the Fashion MNIST dataset in two dimensions. The resulting embeddings should be plotted and color-coded based on their true labels for visual interpretation. The quality of the t-SNE visualization can be evaluated based on the separation of the clusters and the preservation of the local and global structure of the data.

5) Evaluation:

Implement PCA, Isomap, and LLE using scikit-learn, and compare the results with the baseline t-SNE algorithm. Visual comparison can be made using Matplotlib, and the performance of the algorithms can be evaluated based on their ability to preserve the structure and reduce the dimensionality of the data.

PROBLEM SET PEDAGOGICAL VALUE

This problem set has significant pedagogical value as it provides students with a comprehensive understanding of dimensionality reduction techniques and their importance in data analysis and visualization. By implementing t-SNE and other algorithms, students can explore their strengths and limitations and gain a deeper understanding of their applications in various fields such as machine learning, computer vision, and data science.

1) Understanding the importance of dimensionality reduction:

Dimensionality reduction is an important concept in machine learning, as it enables efficient computation and analysis of high-dimensional datasets. By reducing the number of features, the computation time can be significantly reduced, and the resulting visualization can be more easily interpreted. t-SNE is an excellent example of a nonlinear technique that can be used to visualize complex high-dimensional datasets.

2) Exploring hyperparameter tuning:

Hyperparameter tuning is an important aspect of machine learning, and students can learn about the impact of hyperparameters on the performance of the algorithms. By

experimenting with different hyperparameters of t-SNE, such as perplexity and learning rate, students can learn about the effect of hyperparameter tuning on the resulting visualization and understand the importance of parameter selection.

3) Visualizing high-dimensional data:

Visualizations are a powerful tool for exploring high-dimensional datasets, and t-SNE can help students understand the relationships, similarities, and differences between different clusters. By visualizing the embeddings generated by t-SNE, students can gain insights into the structure of the dataset and identify patterns that are not easily visible in the original high-dimensional space.

4) Understanding the limitations of visualization:

While t-SNE can provide useful visualizations, it is important to understand the limitations of visualization and the fact that it may not always reveal all the underlying patterns or relationships in the data. Students can learn about the limitations of visualization and explore other analysis techniques that can complement t-SNE, such as clustering, classification, and regression.