t-SNE

Tushar B. Kute,
http://tusharkute.com

t-SNE

- t-SNE (t-Distributed Stochastic Neighbor Embedding) is a popular machine learning algorithm for dimensionality reduction and visualization.
- It is particularly effective for high-dimensional data.
- Developed by Laurens van der Maaten and Geoffrey Hinton in 2008, t-SNE helps to visualize high-dimensional data by mapping it into a two or three-dimensional space, making it easier to see patterns, clusters, and relationships.

T-SNE : Concepts

- High-dimensional Space: Data in its original form can have many dimensions (features).
- Low-dimensional Embedding: The goal is to represent this high-dimensional data in a lower-dimensional space (typically 2D or 3D) while preserving the structure and relationships of the data points as much as possible.
- Similarity Preservation: t-SNE aims to keep similar data points close together and dissimilar points far apart in the lower-dimensional space.

t-SNE vs PCA

- Both t-SNE and PCA are dimensional reduction techniques that have different mechanisms and work best with different types of data.
- PCA (Principal Component Analysis) is a linear technique that works best with data that has a linear structure.
  - It seeks to identify the underlying principal components in the data by projecting onto lower dimensions, minimizing variance, and preserving large pairwise distances.

t-SNE vs PCA

- t-SNE is a nonlinear technique that focuses on preserving the pairwise similarities between data points in a lower-dimensional space. t-SNE is concerned with preserving small pairwise distances whereas, PCA focuses on maintaining large pairwise distances to maximize variance.
- In summary,
  - PCA preserves the variance in the data, whereas t-SNE preserves the relationships between data points in a lower-dimensional space, making it quite a good algorithm for visualizing complex high-dimensional data.

t-SNE Working

- t-SNE models a point being selected as a neighbor of another point in both higher and lower dimensions. It starts by calculating a pairwise similarity between all data points in the high-dimensional space using a Gaussian kernel. The points that are far apart have a lower probability of being picked than the points that are close together.
- Then, the algorithm tries to map higher dimensional data points onto lower dimensional space while preserving the pairwise similarities.
- It is achieved by minimizing the divergence between the probability distribution of the original high-dimensional and lower-dimensional. The algorithm uses gradient descent to minimize the divergence. The lower-dimensional embedding is optimized to a stable state.

t-SNE Advantages

- Visualization:
  - t-SNE is excellent for visualizing clusters and the structure of high-dimensional data in 2D or 3D plots.
- Non-linear Mapping:
  - It captures non-linear relationships in data, making it more powerful than linear methods like PCA (Principal Component Analysis).

t-SNE Disadvantages

- Computationally Intensive:
  - t-SNE can be slow and resource-intensive, especially for large datasets.
- Parameter Sensitivity:
  - It has several parameters (e.g., perplexity, learning rate) that require tuning and can significantly impact the results.
- Interpretability:
  - The resulting axes in the low-dimensional space do not have a straightforward interpretation.

Thank you

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com