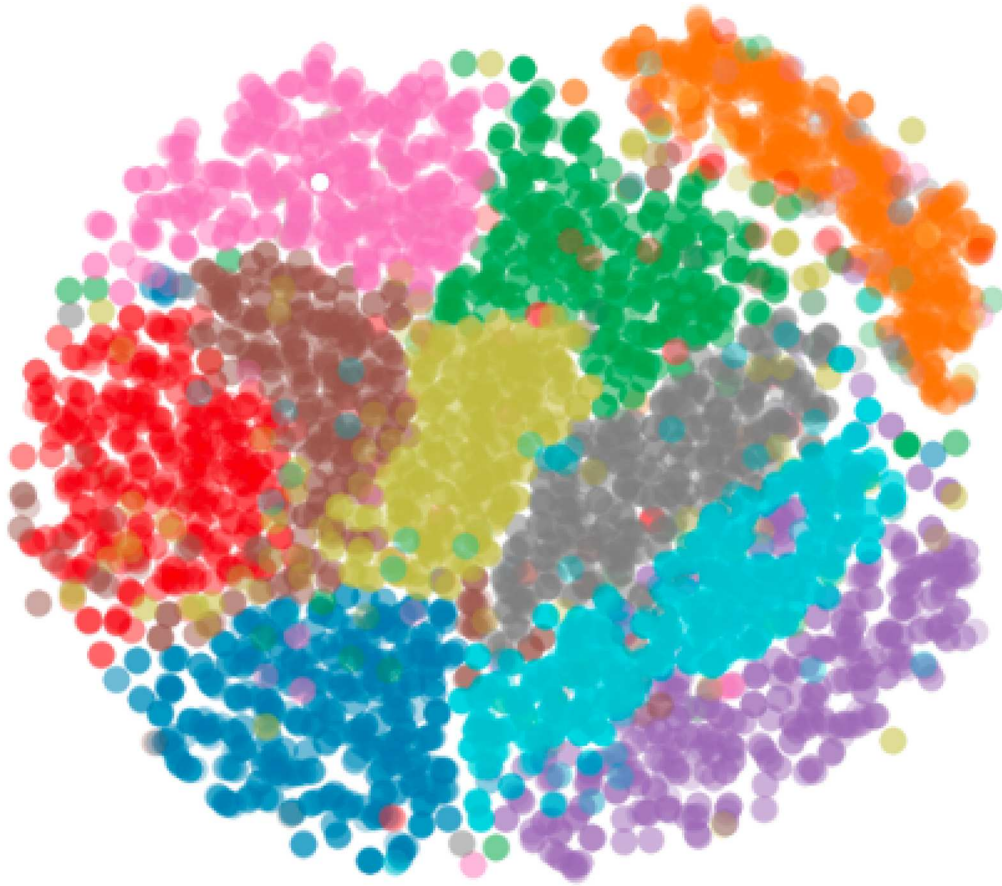
t-Distributed Stochastic Neighbour Embedding (t-SNE)



DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Key Concepts

Unsupervised Learning

In unsupervised learning, an algorithm is trained on a dataset without any labeled output or target variable.

Dimensionality Reduction

It is an unsupervised learning technique where the number of data inputs are reduced to a manageable size while also preserving the integrity of the dataset as much as possible.

The two common methods for dimensionality reduction are:

1. Principal Component Analysis (PCA)

2. t-Distributed Stochastic Neighbour Embedding (t-SNE)

What is t-SNE?

t-SNE is a non-linear dimensionality reduction technique.

It is a probabilistic approach to place samples from high-dimensional space into low-dimensional space so as to preserve the identity of neighbors.

It finds an embedding so that original high-dimensional sample distribution is approximated well by the resulting low-dimensional sample distribution.

t-SNE Steps

1. Prepare the data: Prepare the data by selecting the appropriate features, scaling the data if necessary, and organizing it in a format suitable for the algorithm.

2. Calculate pairwise similarities: Calculate pairwise similarities between all of the data points in the high-dimensional space. This can be done using a Gaussian kernel or another similarity metric such as cosine similarity or Euclidean distance.

t-SNE Steps

3. Compute probability distributions: Using the pairwise similarities, we then compute the probability distribution of each point being selected as a neighbor by other points. The probability distributions are computed separately for the high-dimensional space and the low-dimensional space (usually 2 or 3 dimensions).

t-SNE Steps

4. Optimize the low-dimensional embeddings: Minimize the difference between the two probability distributions using a gradient descent algorithm. This is done by adjusting the low-dimensional embeddings to better match the pairwise similarities of the high-dimensional data.

5. Visualize the results: Once the optimization is complete, the low-dimensional embeddings can be visualized on a scatter plot or other type of graph to reveal patterns and clusters in the data.

PCA vs t-SNE

PCA is a useful technique for simple and fast dimensionality reduction, while t-SNE is a more powerful technique for preserving the local structure of the data and revealing patterns and clusters. The choice between PCA and t-SNE depends on the specific requirements of the problem at hand and the characteristics of the data being analyzed.

Disadvantages

t-SNE is computationally expensive and can be sensitive to the choice of parameters, making it more difficult to implement and optimize than PCA

Follow **#DataRanch** on LinkedIn for more...

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Follow **#DataRanch** on LinkedIn for more...

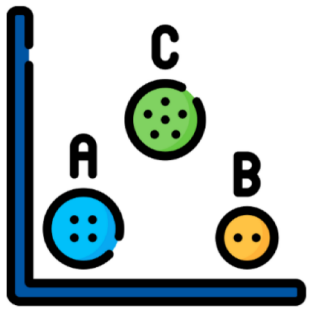
What is Supervised Learning?

DATARANCH.org

What is Unsupervised Learning?

DATARANCH.org

Clustering



DATARANCH.org

Principal Component Analysis

DATARANCH.org

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

info@dataranch.org

linkedin.com/company/dataranch