What is t-SNE?

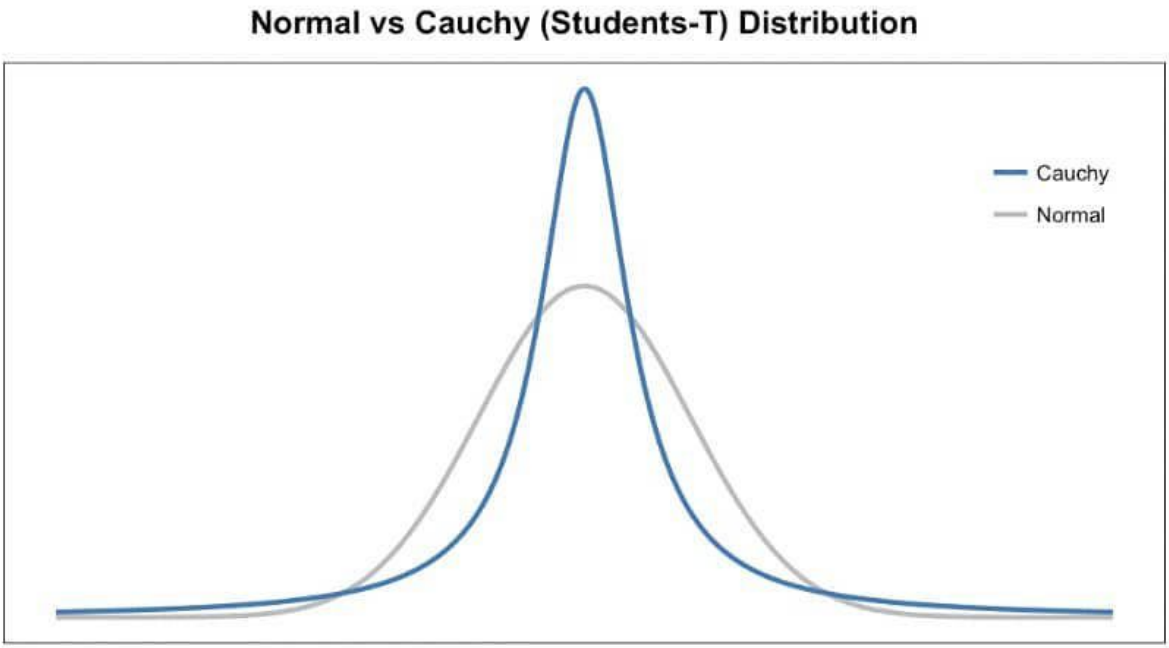
- t-SNE is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data. In simpler terms, t-SNE gives you a feel or intuition of how the data is arranged in a high-dimensional space. It was developed by Laurens van der Maatens and Geoffrey Hinton in 2008.
- It is extensively applied in image processing, NLP, genomic data and speech processing.

How t-SNE works

- **We are not going into math details but we explain the algorithm in detail.**
- **Step 1, measure similarities between points in the high dimensional space. Think of a bunch of data points scattered on a 2D space. For each data point (xi) we'll centre a Gaussian distribution over that point.**
- **Then we measure the density of all points (xj) under that Gaussian distribution. Then renormalize for all points. This gives us a set of probabilities (Pij) for all points. Those probabilities are proportional to the similarities.**
- **The similarity of points is calculated as the conditional probability that a point Xi would choose point Xj as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian (normal distribution) centred at Xi.**

t-Distributed Stochastic Neighbor Embedding (t-SNE)

How t-SNE works



Normal vs Student t-distribution source @medium

- **Step 2 is similar to step 1, but instead of using a Gaussian distribution you use a Student t-distribution with one degree of freedom, which is also known as the Cauchy distribution (Figure 3). This gives us a second set of probabilities (Qij) in the low dimensional space. As you can see the Student t-distribution has heavier tails than the normal distribution. The heavy tails allow for better modeling of far apart distances.**

How t-SNE works

- It then tries to minimize the difference between these conditional probabilities (or similarities) in higher-dimensional and lower-dimensional space for a perfect representation of data points in lower-dimensional space.
- We want the two map structures to be similar. We measure the difference between the probability distributions of the two-dimensional spaces using Kullback-Liebler divergence (KL divergence).
- Note Kullback-Leibler divergence or KL divergence is a measure of how one probability distribution diverges from a second, expected probability distribution.

Time and Space Complexity

- As you might have observed, that the algorithm computes pairwise conditional probabilities and tries to minimize the sum of the difference of the probabilities in higher and lower dimensions. This involves a lot of calculations and computations. So the algorithm is quite heavy on the system resources.
- t-SNE has a quadratic time and space complexity in the number of data points. This makes it particularly slow and resource draining while applying it to data sets comprising of more than 10,000 observations.

What does t-SNE actually do?

- t-SNE a non-linear dimensionality reduction algorithm finds patterns in the data by identifying observed clusters based on similarity of data points with multiple features. But it is not a clustering algorithm it is a dimensionality reduction algorithm. This is because it maps the multi-dimensional data to a lower dimensional space, the input features are no longer identifiable. Thus you cannot make any inference based only on the output of t-SNE. So essentially it is mainly a data exploration and visualization technique.
- But t-SNE can be used in the process of classification and clustering by using its output as the input feature for other classification algorithms.

Few things about t-SNE

- For the algorithm to execute properly, the perplexity should be smaller than the number of points. Also, the suggested perplexity is in the range of (5 to 50)
- sometimes, different runs with same hyper parameters may produce different results.
- Cluster sizes in any t-SNE plot must not be evaluated for standard deviation, dispersion or any other similar measures. This is because t-SNE expands denser clusters and contracts sparser clusters to even out cluster sizes. This is one of the reasons for the crisp and clear plots it produces.
- Distances between clusters may change because global geometry is closely related to optimal perplexity. And in a dataset with many clusters with a different number of elements one perplexity cannot optimize distances for all clusters.

Few things about t-SNE

- Patterns may be found in random noise as well, so multiple runs of the algorithm with different sets of hyperparameter must be checked before deciding if a pattern exists in the data.
- Different cluster shapes may be observed at different perplexity levels.
- Topology cannot be analyzed based on a single t-SNE plot, multiple plots must be observed before making any assessment.