Title of the Presentation

t-SNE: A Way to Visualize Multidimensional Dataset

Date: 13 December, 2019
Speaker: Subhrajyoty Roy
Mentor: Dr. Ayanendranath Basu, Interdisciplinary Statistical Research Unit (ISRU), ISI, Kolkata

---

Extended Abstract

t-SNE (t-distributed Stochastic Neighbour Embedding) is a statistical technique for performing unsupervised dimension reduction of Multidimensional data developed by Laurens van der Maaten and Geoffrey Hinton. It is of utmost importance to visualize the data (or perform some exploratory analysis) before proceeding with any kind of advanced statistical analysis, to get a feel for the data, as well as effectively determine courses of action to analyze the data. t-SNE is now used for visualization in an extensive range of applications, including Computer Security Research, Music Analysis, Text Analysis, Cancer Research, Bioinformatics and Biomedical Signal Processing. The main effectiveness of t-SNE is to identify patterns and natural clusters of the data automatically, in an unsupervised way. The technique is a variation of Stochastic Neighbor Embedding that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. This is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints.

From a working point of view, whereas PCA and other classical techniques focus on retaining global pairwise distances between all pairs of points, t-SNE focuses on modeling pairwise distances at a local level. This allows it to effectively represent Swiss Roll type dataset, for which preserving the Geodesic distance is more meaningful than preserving the Euclidean distance. Also, the technique can be implemented for large real-world datasets with millions of observations, using Barnes-Hut Approximation. There are also some variants like Parametric t-SNE, Multiple maps t-SNE which is designed specifically to solve certain types of problems, which is of extreme importance in various applied fields.

In recent years, there were some advancements regarding theoretical aspects of the performance of t-SNE to visualize and separate natural clusters of the high dimensional data into embedding space, which was already empirically observed through its astounding performance in the visualization of real-world datasets.