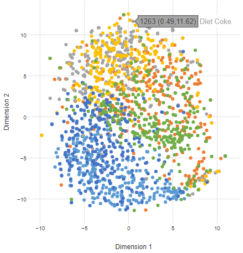
t – Distributed Stochastic Neighbor Embedding

Introduction:

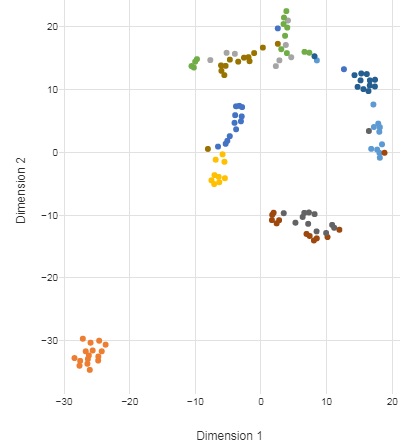
t-SNE is one of the popular algorithms in machine learning which is used to reduce dimensionality problem. It basically takes a high dimensional dataset and converts it into a low dimensional graph which retains a lot of information present in the original dataset. It uses mathematics in the background and accomplishes the task.

It generally reduces or eliminates the curse of dimensionality problem which refers to low performance of the model due to the large number of features. It was developed by Goffrey Hinton and Laurens van der Mateen.

Input:



Output:



Working:

Step 1:

Detrmine “similarity” of all the points with respect to all the points.

To find the similarity:

For each cluster of points, Measure the distance between two points in a graph. Plot the distances on the normal curve. The length of the vertical line which joins the point on the curve and the point on axis is the unscaled similarity. Thus close points have high similarity and distant points have low similarity.

Step 2:

Scale these similarities so that they sum up to one. This is done to equalize the wide curves and the narrow curves as the graph of a cluster varies with another. Usually the curves are wide if the density of point is less. This is done by dividing each score by sum of all scores. Now scaled similarities are ready.

Step 3:

Consider 2 points A and B. Similarity Score of A w.r.t B might not be the same as the Similarity Score of B w.r.t A. So t-SNE averages both of them. Similarity Score of A w.r.t A is zero.

Step 4:

Now project he points of original data on number line vertically. Measure the distance. Plot it on t-distributed curve. End the process.

Note: t-SNE has “Perplexity” parameter equal to the expected density.