ASSIGNMENT : 7

Name: FALAK KHAN
Batch: MACHINE LEARNING AND AI BATCH A3

1) Please explain your understanding of TSNE.

Ans: **t-distributed stochastic neighbor embedding** (**t-SNE**) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.

It is based on Stochastic Neighbor Embedding originally developed by Sam Roweis and Geoffrey Hinton where Laurens van der Maaten proposed the t-distributed variant. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

2) What is dimensionality-reduction?

Ans: Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

3) What is LDA?

Ans: Linear Discriminant Analysis or LDA is a dimensionality reduction technique. It is used as a pre-processing step in Machine Learning and applications of pattern classification. The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.

4) Explain the curse of dimensionality.

Ans: Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data is referred to as ‘Curse of Dimensionality’.

5) What is the difference between PCA and TSNE?

Ans:

	PCA	t-SNE
1.	It is a linear Dimensionality reduction technique.	It is a non-linear Dimensionality reduction technique.
2.	It tries to preserve the global structure of the data.	It tries to preserve the local structure (cluster) of data.
3.	It doesn't work well as compared to t-SNE.	It is one of the best dimensionality reduction techniques.
4.	It does not involve Hyperparameters.	It involves Hyperparameters such as perplexity, learning rate and number of steps.
5.	It gets highly affected by outliers.	It can handle outliers.
6.	PCA is a deterministic algorithm.	It is a non-deterministic or randomized algorithm.
7.	It works by rotating the vectors for preserving variance.	It works by minimizing the distance between the point in a gaussian.
8.	We can find decide on how much variance to preserve using eigen values.	We cannot preserve variance instead we can preserve distance using hyperparameters.