Prof. Dr. Carsten Meyer

Fachbereich Informatik & Elektrotechnik

Fachhochschule Kiel

Machine Learning – Summer Term 2020

Exercise sheet 1

**Submission due:**

**Exercise 1 ( Unsupervised Learning ):**

a ) Briefly explain the following terms related to Unsupervised learning: (e.g. 1-2 sentences for each term)

Clustering.

Organizing set of Cluster of objects, based on set of similar characteristics in the dataset.

Association

An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Semi-supervised Machine Learning

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems. Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

Hard and Soft Clustering

In hard clustering, each data point either belongs to a cluster completely or not.e.g. if the customers were to be segmented in 5 groups then each customer is put into one group out of the 10 groups.

In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

b). What are different types of clustering methodologies ? Brefly explain each with examples.

Connectivity-based models

* Builds model based on distance connectivity.
* e.g Hierarchical Clustering

Centroid models

* Representing each cluster by a single mean vector.
* e.g. K-means Clustering

Distribution models

* Clusters are modeled using statistic distributions.
* e.g. Expectation–Maximization (EM) algorithm using multivariate normal distribution.

Density models

* Defining clusters as connected dense region in the dataset
* e.g. DBSCAN( Density-Based Spatial Clustering of Applications with Noise )

Neural modes

* These models can be characterized as one or combination of more models above.
* e.g. Self-organizing maps in Unsupervised neural network

Grid based models

* Used for multi-dimensional dataset.
* Creating grids and performing Comparison on grids( or cells ).
* e.g. STING ( a STatistical INformation Grid approach )

**Exercise 2 ( Clustering ):**

**a) What are the Challenges associated with clustering on high-dimensional data ? Brefly explain Curse of dimensionality.**

Since clustering is finding different patterns within the data with a goal that resulting cluster captures "Natural" structure of the data. e.g. gene and proteins cluster with similar functionality.

Cluster should be similar( or related ) to one another and different from ( or unrelated to ) the objects in other cluster. the greater the difference between group( cluster ) the better the clustering, but in most case the definition of what constitutes a cluster is not well defined and clusters are not well separated from one another.

The visualization of dimensionality more than 3 is out of our intuition and with the increase of dimension many complexities may arise which could only be represented with mathematical models. The relevancy of distance measures like euclidean distance decreases since computation of pairwise distance of all points becomes computationally challenging.

e.g. Clustering Text Documents where, if word-frequency vector is used, the number of dimension is = size of vocabulary.

Some of the methodologies discussed on 1. b) may not be application on high-dimensional data due to Curse of Dimensionality so they depend on the result and the choice of Dimensionality reduction techniques like PCA, ICA, UMAP e.t.c.

Curse of Dimensionality is the problem when the dimensionality increases, the volume of the space increases so fast the data becomes sparse, which

* Can be harder to visualize due to the growth of number of possible values with each dimension.
* Can affect the methods requiring statistical significance, Since the amount of data needed to support the result grows exponentially with dimension.

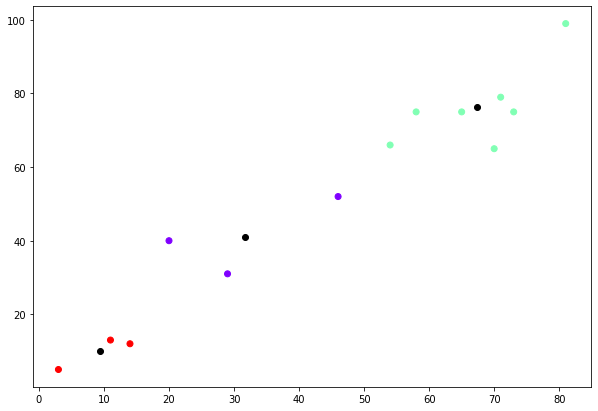
Some of the approaches used for high dimensional data clustering are :

* Subspace clustering
* Projected Clustering
* Hybrid approaches
* Correlation clustering etc.

b) Given the simple dataset X, a**pply** Kmeans clustering algorithm and compute the final centroid values for each clusters with number of cluster = 3.

X = [3,5], [11,13], [14,12], [20,40], [29,31],[46,52],[54,66], [65,75], [71,79], [58,75], [70,65], [73,75], [81,99]

Hint : use *Scikit-learn* and convert dataset to *numpy array.*



Ref : Kmeans.ipynb

The centroids of the given clusters represented by black dot are ( 9.33, 10 ),( 31.67, 41 ), ( 67.43, 76.29 ).

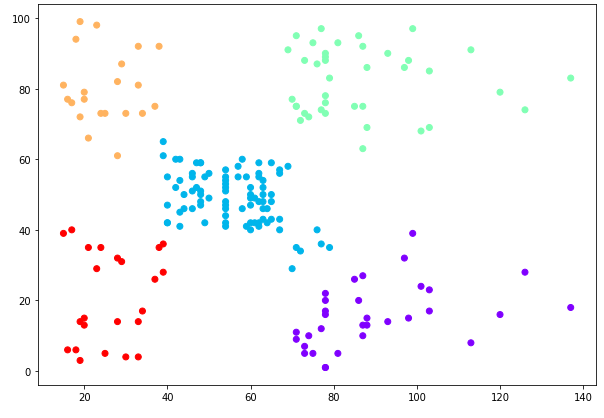
c) Provided with this exercise sheet is a customer\_shopping dataset. perform Customer segmentation using hierarchical clustering, find clusters of interest to the Shopping mall with the customer spending score depending on annual income and interpret the result

Dataset information : 200 Samples

Attributes:

* CustomerID
* Genre( Male | Female )
* Age
* Annual Income( in Thousand $ )
* Spending Score( 1- 100 )

The spending score signifies how often a person spends money in a mall with scale (1-100). where 100 being the highest spender.

Ref : HierarchicalClustering.ipynb

Given 5 clusters

* Bottom right are customer with high salary, low spending
* Bottom left are customer with low salary, low spending
* Top right are the customer with high salary, high spending
* Top left are the customer with low salary , high spending

The cluster in the middle is of interest to the shopping mall because those are the most stable customers. Furthermore, the cluster analysis can be performed according to age or genre of customers for different insights.

**Exercise 3 (Dimensionality Reduction ) :**

**a) State Dimensionality reduction and typical techniques used to perfom reduction of dimensionality.**

Given : Collection of N high-dimensional objects x\_1, ... x\_n.

* Exploration of the arrangement of the high-dimensional objects in data space.
  + Data Visualization.
  + Structure of the data( e.g. Clusters, Local Structures, ... )
* Reduction of High dimensional objects by building a map in which distances between points reflect similarities in data.
  + Minimize some objective function that measures the discrepancy between similarities in the data and similarities in the map.( i.e. reflect the data in low dimension preserving similarities/characteristics. )
  + Termed as Embedding or Reduction or Multidimensional Scaling.

Typical Techniques

Matrix Factorization

e.g. PCA, Linear Auto-encoder, Word2Vec, Latent Dirichlet Allocation e.t.c.

Neighbor Graphs

e.g. ISOMAP, t-SNE, LLE, UMAP, Laplacian Eigenmaps, LTSA e.t.c.

b) Attached with this Exercise sheet is a mice protein expression dataset. perform dimensionality reduction with Principal component analysis .

*Expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning.*

*Classes:*

* *c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)*
* *c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)*
* *c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)*
* *c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)*
* *t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)*
* *t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)*
* *t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)*
* *t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)*

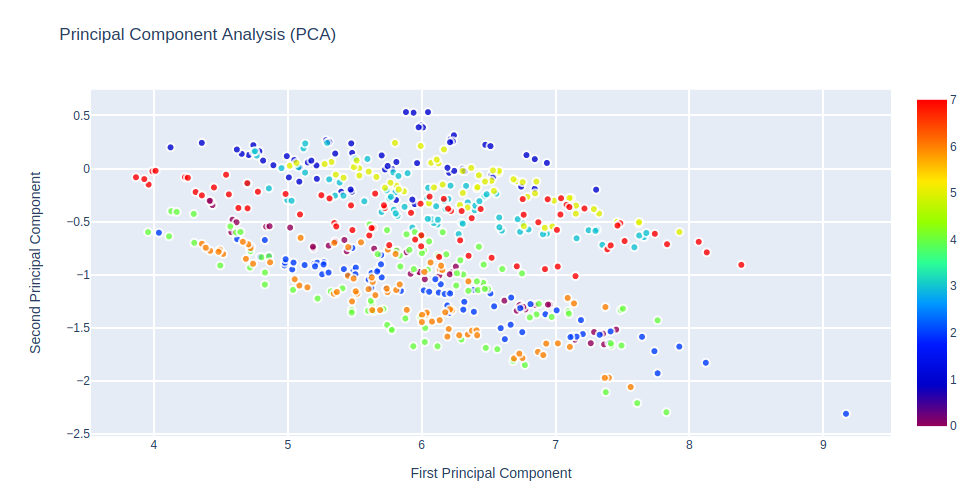
*Attribute Information:*

* *(1) Mouse ID*
* *(2-78) Values of expression levels of 77 proteins; the names of proteins are followed by â€œ\_nâ€ indicating that they were measured in the nuclear fraction. For example: DYRK1A\_n*
* *(79) Genotype: control (c) or trisomy (t)*
* *(80) Treatment type: memantine (m) or saline (s)*
* *(81) Behavior: context-shock (CS) or shock-context (SC)*
* *(82) Class: c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m (*[*https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression*](https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression)*)*

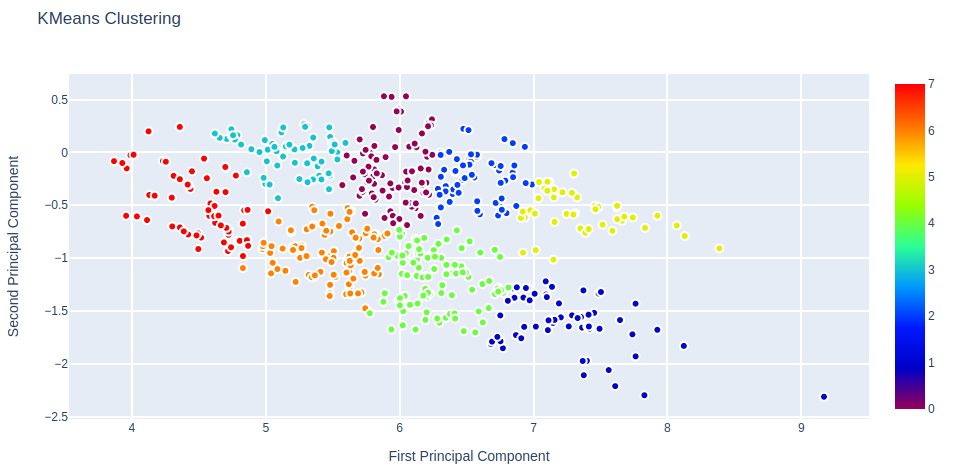
**Solution :**

Followin*g* figure shows the linear dimensionality reduction using Singular Value Decomposition of the data from 77D to lower 2 dimensional space.

Ref : miceProtein.ipynb



c) Apply Kmeans clustering on dataset on 3. b) them and discuss the resulting clusters.



Ref : miceProtein.ipynb

The resulting clusters shows the separation of samples in a groups with similar protein expressions. The number of centroids generated = no. of classes = 8

Exercise 4 (t-SNE and PCA for data visualization ) :

a) Discuss the influence of different parameter to the implementation on t-SNE to visualize the data in 2D space.

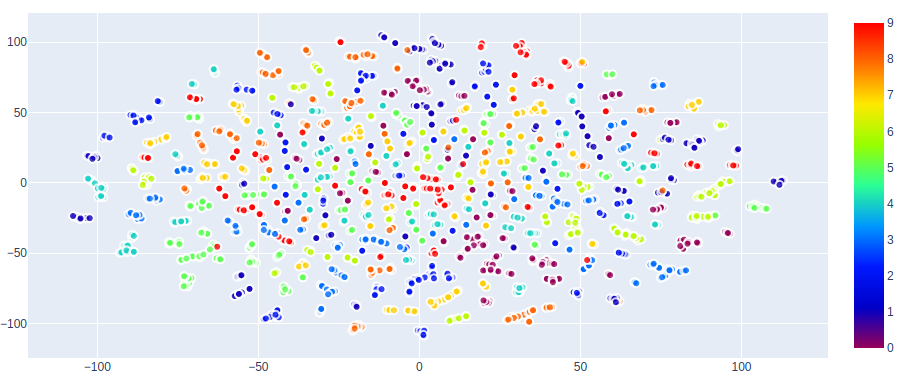
t-SNE is an algorithm for exploring high-dimensional data. A non-linear dimensionality reduction technique well suited for embedding high-dimensional data for visualization in a low dimensional space( 2 or 3) suitable for human observation. It works in following steps :

* Construction of Probability distribution over pairs of high dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked
* Defining a similar probability distribution over the points in the low-dimensional map, minimizing the Kullback-Leibler(KL) divergence between the two distributions with respect to the locations of the points in the map.

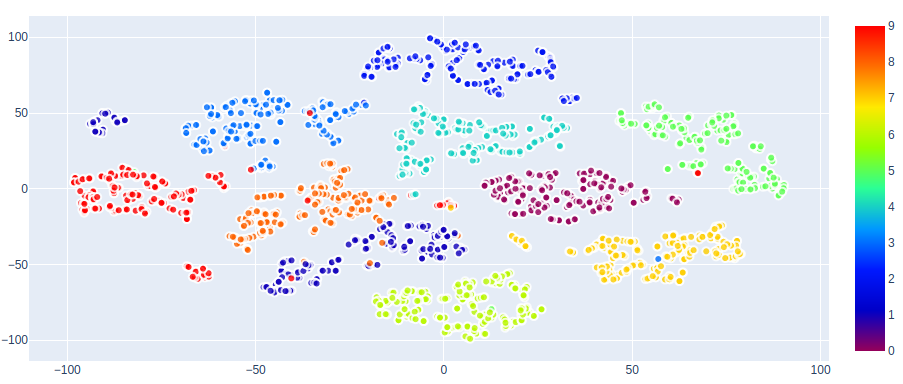
Different parameter used with t-SNE were observed keeping other parameter to default and results are discussed below:

Perplexity( default = 30)

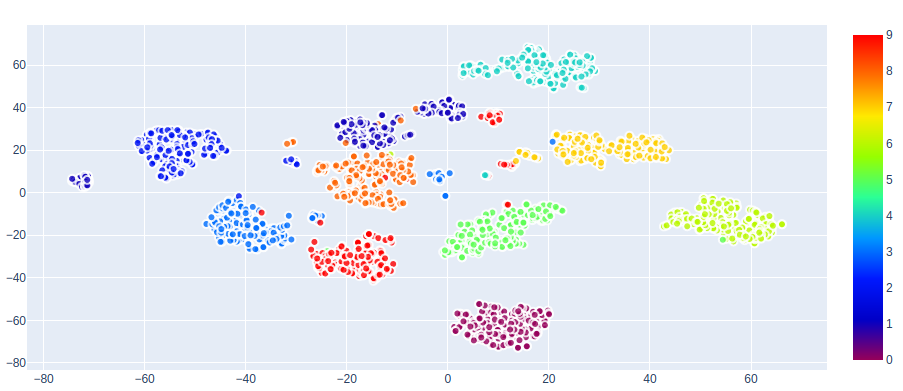
*perplexity = 1*



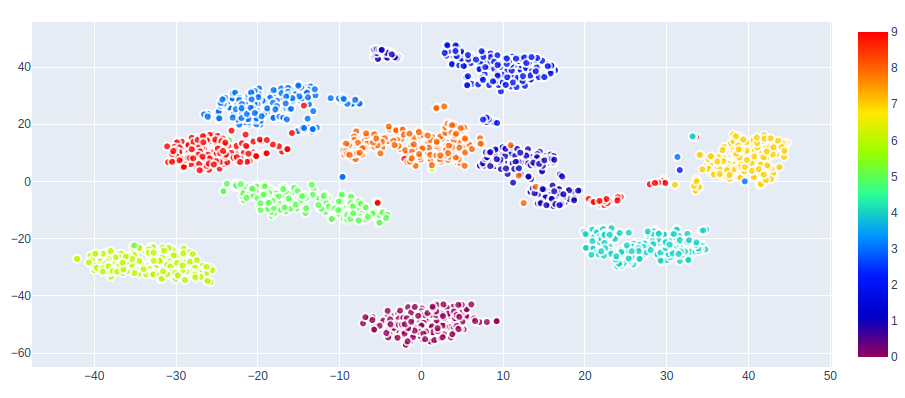
*perplexity = 5*

**

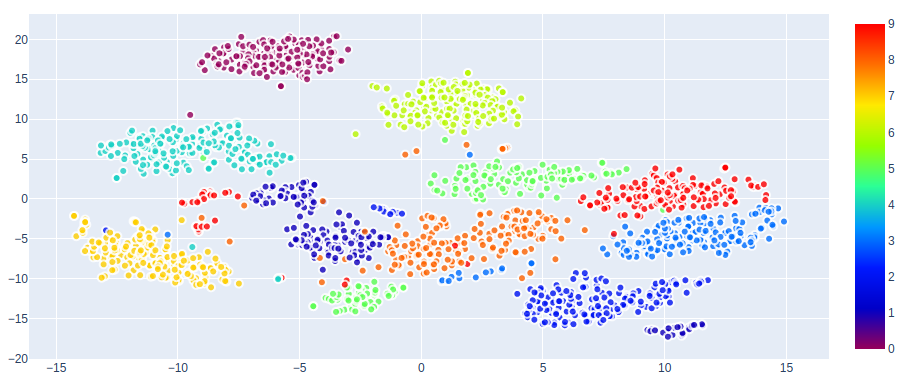
*perplexity = 25*

**

*perplexity = 50*

**

*perplexity = 200*

**

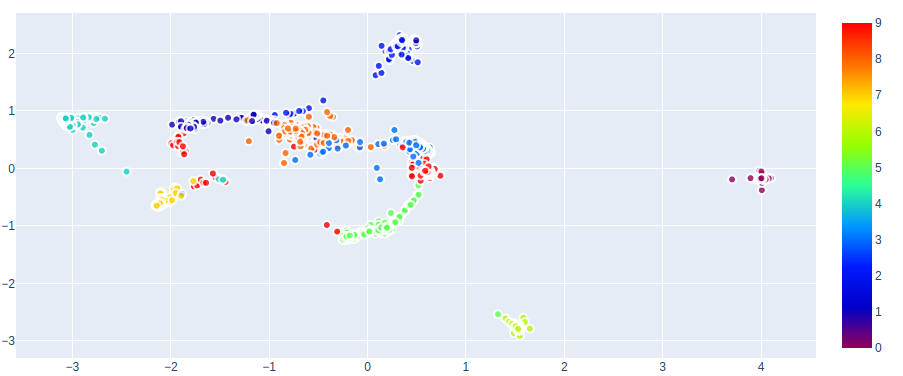
Perplexity is in simple terms, balance of attention between local and global aspect of the dataset( i.e. it sets the number of effective nearest neighbors ). Starting from zero with random projection, with the increase in perplexity the separability of cluster increases with the suitable range of (10 - 100). With further increase it start to converge again

Learning rate (default = 200)

Learning rate value starting from 1, where clusters are not well separated with a well separated range (10 - 8000 ). Further increase of learning rate will converge them to a single cluster. t-SNE is better at visualizing high dimensional data than PCA. but it is prone to getting stuck at local minima. where this parameter may be varied to get a probable escape.

Number of iteration ( default = 1000 )

Minimum possible value : 250.

****

Starting with 250 as shown in the given figure, starts being effective from 500 showing different cluster of interest. increasing after a certain point( more than 1000 ) will not result in better result and only increase time complexity.

Topology

Exploration of different topological information requires more than one plot with different experimental values of perplexity.

Distance

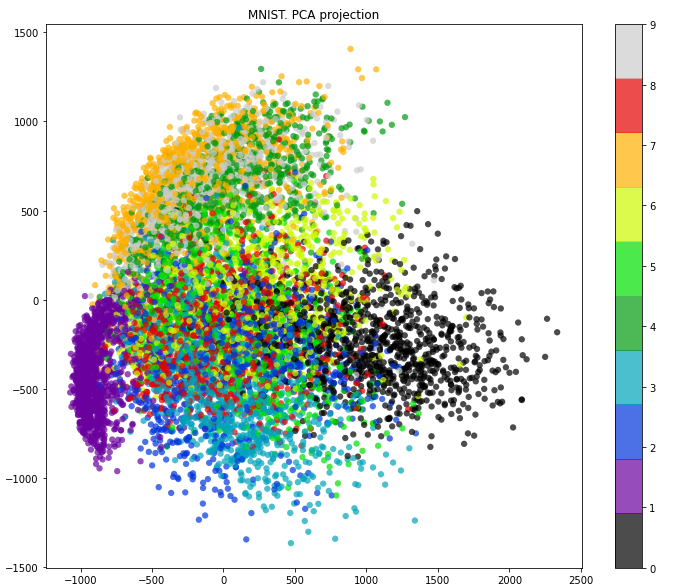
t-SNE does not retain distances but probabilities, so measuring some error between Euclidean distances in high dimensional and low dimensional is not much use. However, with the same data and perplexity the KL divergence can be compared.

The flexibility of t-SNE to explore high-dimensional structure is a blessing but with this blessing comes the complexity of interpretablity.

**b) Compare and contrast t-SNE with other linear ( PCA ) and Non-linear( ISOMAP ) Algorithms for visualization.**

In PCA the hyperplane is selected such that when all the points in high-dimension are projected, the variance is maximized( i.e axis of Maximum variance )

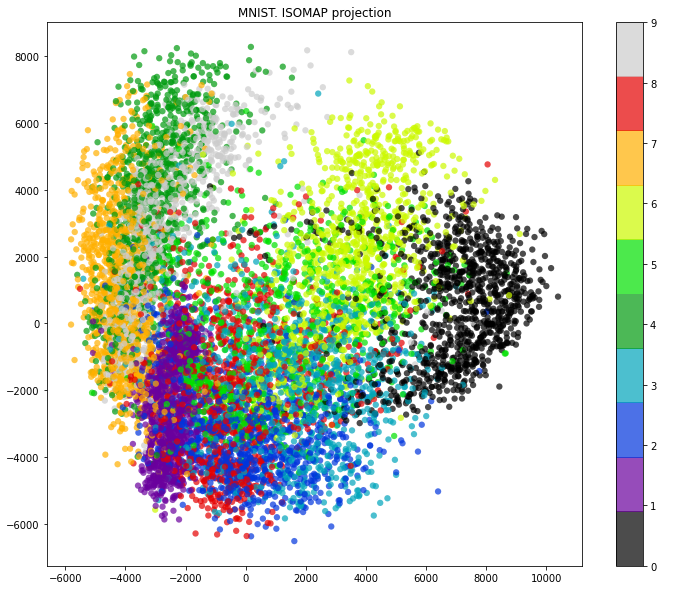
Ref : tSNE.ipynb



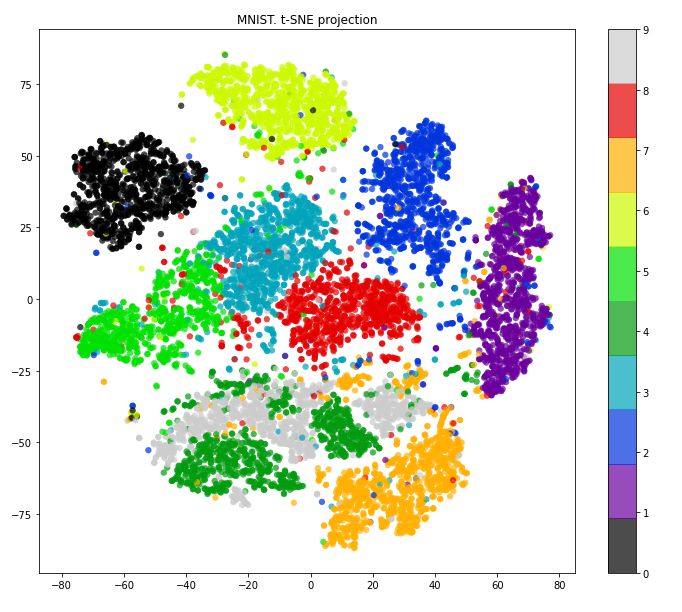
As shown in the given figure digit 1 and 0 are well separated since they have less pixels in common than other digits. Although the principal component analysis shows the variability in dataset from the best angle possible , these linear structure are not enough to capture global geometry( i.e. preserving the distance between points in our visualization as close as possible to the distance in original space)

Another type of linear dimensionality reduction technique LDA is similar to PCA but is supervised and the axis of maximal class separation is selected( i.e. maximal marginal separation between classes )

Isomap is a Non-linear approach which attempts to recover original embedding of hidden data.



It is observed to be able to better separate the clusters than PCA. Although it tries to capture the estimated intrinsic geometry from high dimensional space , is not suitable for clear visualization in 2 D space.



t-SNE tries to preserve the topology of the data and hence better separate the clusters. for every point it constructs a notion of which other points are it's neighbors, trying to make all points to have same number of neighbors, then tries to embed them so that these points have the same number of neighbors. It is better at visualizing high dimensional data than PCA. but it is prone to getting stuck at local minima.

These minima could be overcome by using more data and carefully selecting hyper-parameters.

|  |  |
| --- | --- |
| **PCA** | **t-SNE** |
| It tries to preserve linear structure | It tries to preserve topology ( neighborhood structure ) |
| it is deterministic. | Different initialization may lead different results. |
| It is Computationally faster | It is Computationally Intense |
| less flexible for visualization | more flexible for visualization |
|  | It can capture structure of complex manifolds |

There are abundance of techniques and variations to explore. It is harder to map high dimensional data into low dimensions and preserve all the structure so an approach must make trade-offs, sacrificing one property to preserve another.