t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE)

Icke-linjär dimensionreduceringsteknik för data exploration och visualisera högdimensionell data

Ger en känsla för hur datan är fördelad

Dimensionsreducering är att presentera högdimensionell data i 2-3 dimensioner

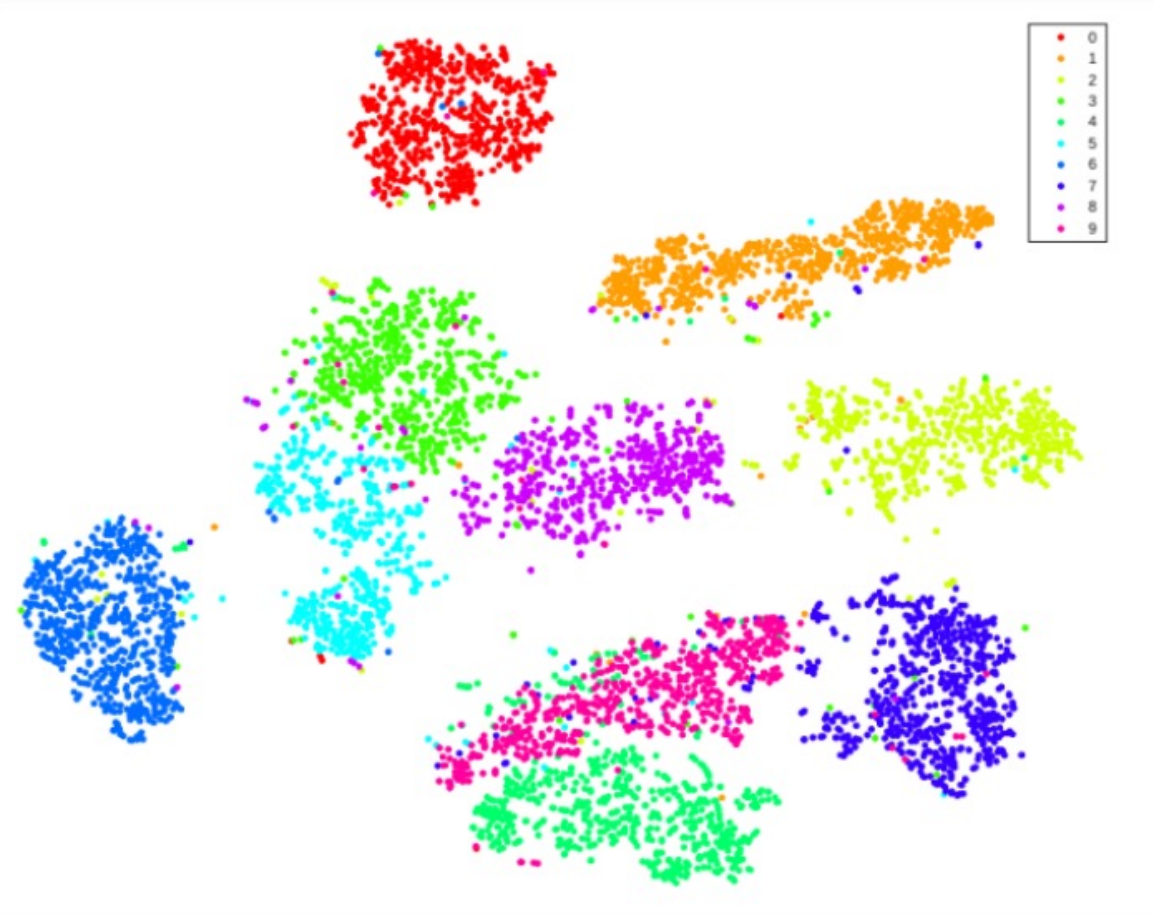


Figure 1 : Illustration of t-SNE on MNIST dataset

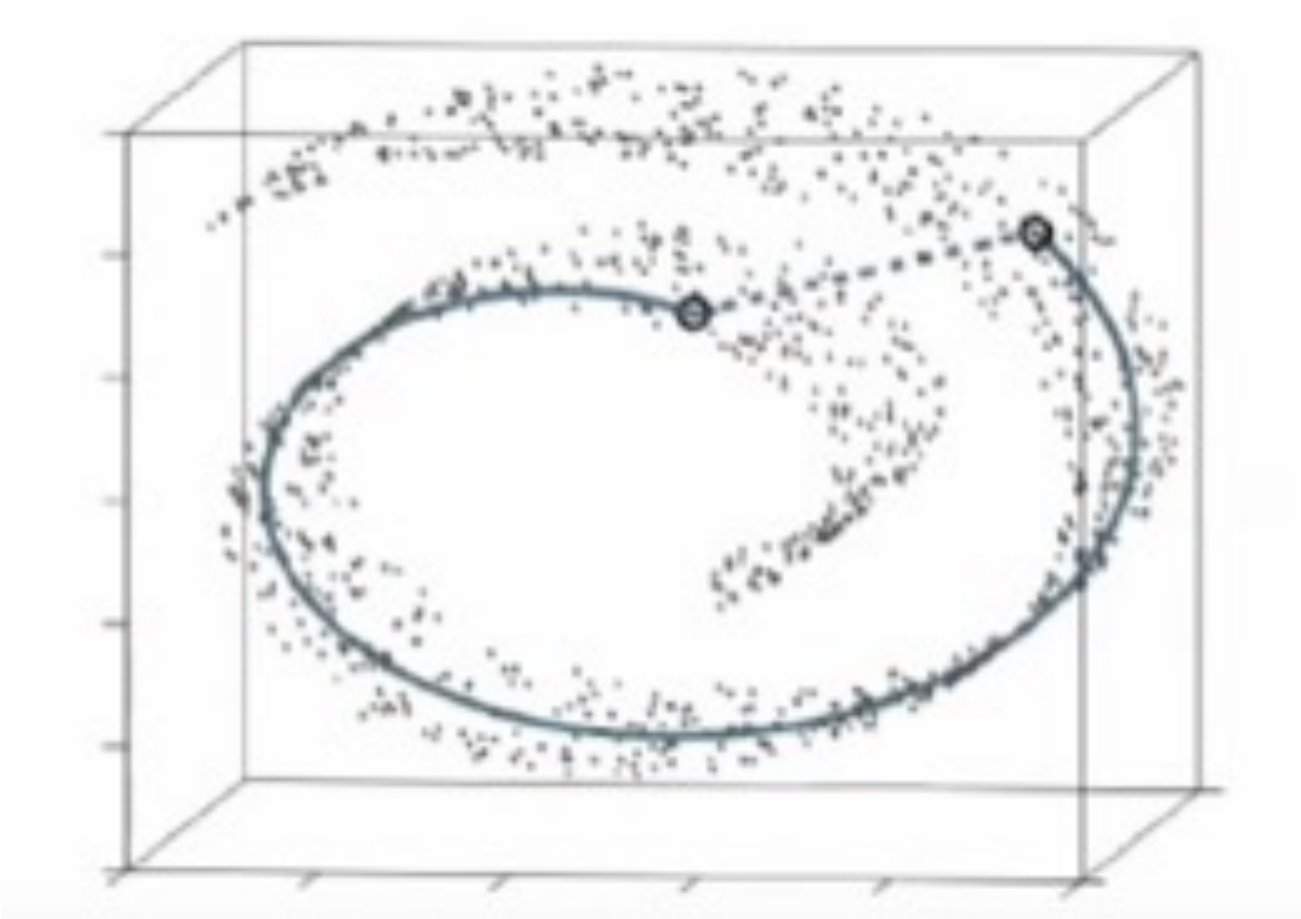
Jämförelse PCA

PCA är linjär dimensionsreducering

PCA maximerar varians och behåller stora parvisa distanser.

Dvs saker som är olikt hamnar långt isär

Detta kan vara problematisk i de icke-linjära fallen



t-SNE

- t-SNE hittar mönster i datan baserad på likheter i datapunkternas features
- Likheten beräknas med conditional probability (betingad sannolikhet) att punkt A kommer att välja punkt B som sin granna
- t-SNE försöker minimera skillnaden mellan dessa conditional probabilities (likheter) båda i högdimensionella fallet och i lågdimensionella fallet
- På så sätt kan den presentera datapunkterna så bra som möjligt i låga dimensioner
- Algoritmen involverar många beräkningar så den tar mycket plats och tar lång tid att köra
- Algoritmen har kvadratisk tidskomplexitet efter antal datapunkter

UMAP

Uniform Manifold Approximation and Projection (UMAP)

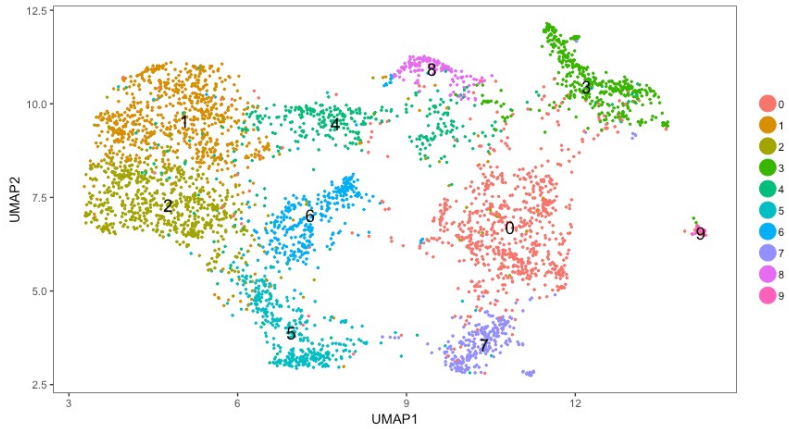
Kan användas som visualiseringsteknik och dimensionsreducering

Hanterar icke-linjäritet

UMAP modellerar mångfaldet till datan med ett fuzzy topologisk rum. Algoritmen söker motsvarande fuzzy rum i låg dimension av datan och använder denna

Mångfald = topologisk rum som liknar et n-dimensionell euklidisk rum

Fuzzy = ett sant värde kan vara ett värde mellan 0 och 1 (i motsättning till binära där det antingen är 0 eller 1)



Jämförelse t-SNE

- t-SNE är väldigt långsam för stora datamängder
- t-SNE är bra på likheter i kluster, men inte kluster som är lika andra kluster
- t-SNE är bra för visualisering i 2-3 dimensioner, men inte så bra för feature selection