Dimensionality Reduction Techniques

SJSU ID - 015939963
Name - Harika Nalam

Different dimensionality reduction techniques are applied to Image and tabular data.

Image data used is digits data. The digits dataset is a dataset of handwritten digits and each feature is the intensity of one pixel of an 8 x 8 image.

Tabular data is used to mushroom data. The dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota families. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended.
The dimensionality techniques used are

- Principal Component Analysis
- Singular Valued Decomposition
- Local Linear Embedded
- Isometric Mapping
- t Distributed Stochastic Neighbor Embedding
- Uniform Manifold Approximation and Projection

Time taken to reduce the dimensionality of Image and tabular dataset:

Image dataset:

Technique Name	Time taken
PCA	0.029
SVD	0.01
LLE	1.75
t-SNE	35.02
ISOMAP	3.75
UMAP	9.79

Tabular dataset:

Technique Name	Time taken
PCA	0.05
SVD	3.49
LLE	4.05
t-SNE	193.86
ISOMAP	44.89
UMAP	47.91

Principal Component Analysis(PCA):
PCA is one of the powerful methods but sometimes it fails as it assumes the data can be linearly modeled.
PCA transforms data by projecting it onto a set of orthogonal axes.
When compared based on the time taken to build the model, PCA takes less time compared to other dimensionality techniques.
But when compared with visualization of data with new principal components, PCA does not show the clear distinction between the data points with new features compared to other techniques.

Singular value Decomposition(SVD):
SVD uses a matrix representation of the dataset, eliminating the fewer important data to produce low-dimensional data.
SVD is a factorization of that matrix into three matrices(orthonormal, diagonal, and orthonormal matrices).
SVD is applied for image compression and recommended systems.

Local Linearly Embedded(LLE):
Local Linear Embedding preserves the local properties of data i.e, points nearby in higher dimension should be closer in the lower dimension.
The LLE algorithm is used to map data into low dimensional data by searching the k nearest neighbors for each data point, then reconstruction of properties using the weighted sum of k nearest neighbors. Finally, reconstruct the data points using reconstructed weights from its neighbors in the reduced dimension.
LLE performs relatively poorly on the MINST data as the dataset consists of multiple manifolds.

Isometric Mapping(ISOMAP):

ISOMAP is a non-linear dimensionality reduction method that preserves geodesic distances(distance between points on curved surfaces) in the lower dimensional space.

ISOMAP approximates both the global and local structure of the dataset in the low-dimensional data.

ISOMAP performs poorly when the dataset is not well sampled.

t Distributed Stochastic Neighbor Embedding(t-SNE):

tSNE captures the local structure of high dimension and also preserves the global structure of data like clusters. It provides well-segregated clusters.

tSNE is a complex model and it takes more time to build the model compared to other techniques.

It is one of the best models, tSNE forms a well-segregated cluster of data points.

Uniform Manifold Approximation and Project(UMAP):

UMAP is a non-linear dimensionality reduction technique. It is a neighbor-based dimensionality reduction technique that handles both numeric and categorical data.

Umap assumes that data is uniformly distributed on a locally connected Riemannian manifold.