

Dimensionality Reduction

It is the process of reducing the number of attributes in the dataset while keeping as much of the variation in the original dataset as possible.

- **Importance of Dimensionality reduction:**

- Less training time, less computational and increase the performance of algorithms.
- Avoid problem of overfitting.
- Useful for data visualization.
- Takes care of multicollinearity: occurs when features are highly correlated with one or more of the other features. It affects the performance of regression and classification models.
- Used for image compression: minimize the size in bytes of images while keeping as much of quality of the image.

- **Dimensionality reduction techniques:**

- **For numerical Data:**

1. Variance Inflation Factor (VIF).
2. Principal Component Analysis (PCA).
3. Singular value Decomposition (SVD).
4. t-Distributed Stochastic Neighbor Embedding (t-SNE).
5. Isomap.
6. Locally Linear Embedding (LLE).
7. Multidimensional Scaling (MDS).
8. Canonical correlation Analysis. (CCA).

- **For categorical Data:**

1. Linear Discriminant Analysis (LDA).
2. Independent Component Analysis (ICA) .

- **For both:**

1. (PCA) .
2. (LLE).
3. (MDS).
4. Isomap

- **Linear techniques:**

- **Variance Inflation Factor (VIF):** is commonly used to detect and address multicollinearity among predictor variables in a linear regression model. Multicollinearity can lead to unstable and unreliable estimates of the regression coefficients, which can affect the model's performance. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

Where:

R^2 is the R-squared value of the regression model that uses the predictor variable in question as the independent variable and all other predictor variables as the dependent variables.

$$VIF = \frac{1}{1 - R_i^2}$$

If:

- VIF equal to 1 = variables are not correlated.
- VIF between 1 and 5 = variables are moderately correlated.
- VIF greater than 5 = variables are highly correlated.

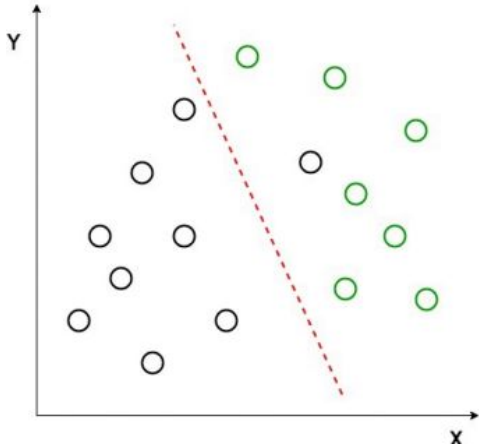
To use VIF in machine learning, one would first fit a linear regression model with all the predictor variables, then calculate the VIF for each predictor variable. A VIF value greater than 5 or 10 is generally considered to indicate a problematic degree of multicollinearity. In this case, one should consider removing one or more correlated predictor variables from the model and re-fit the model.

- **Linear Discriminant Analysis (LDA):** used for classification task.

used for modelling difference in groups spreading two or more classes. it used to project the features in higher dimension space into a lower dimension space.

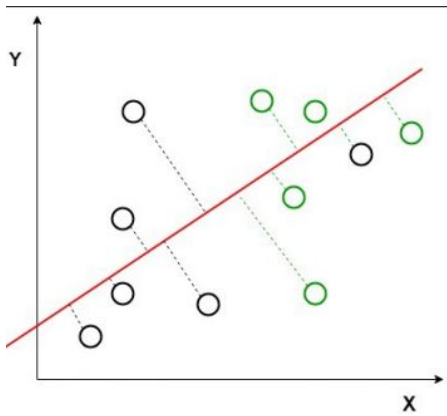
- Example:

Suppose we have to sees of data points belonging to two different classes when the data are plotted on the 2D plane, there is no line that can separate the classes. LDA is used with reduces the 2D graph into a 1D in order to maximize the separability between the classes.



Here LDA uses both axes (x,y) to create new axis and projects data onto a new axis in away to maximize the separation of two classes and reducing the 2D graph into 1D graph.

- Two criteria for LDA by creating new axis:
 - maximize distance between means of a classes.
 - minimize the variation within each class.



in the graph, it can be seen that a new axis (read is generated and plotted in 2D such that it maximize distance between means of a classes and minimize the variation, newly generated axis increases the separation between the data points of a classes.



- Limitations:
 - data should be normally distributed
 - data should contain known class labels
- Difference between LDA and PCA:
 - LDA: find a linear combination of input features that optimizes class separability.
 - PCA: attempts to find a set of uncorrelated components of maximum variance in dataset.
- extensions to LDA:

Quadratic Discriminant Analysis (QDA): Each class uses its own estimate of variance or covariance when there are multiple input variables.

Flexible Discriminant Analysis(FDA): for non-linear combinations of inputs.

Regularized Discriminant Analysis (RDA):Introduces reg. into estimate of the variance moderating the influence of different variables on LDA.

- **Singular value decomposition (SVD):** deals with decomposing a matrix into a product of 3 matrix.

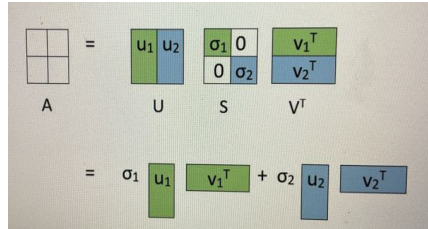


$$A = U S V^T$$

U: is $m \times m$ matrix of left singular vector.

S: is an $m \times n$ diagonal matrix of singular values.

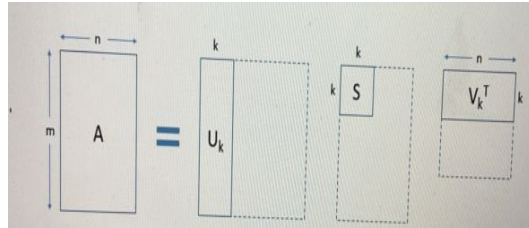
V: is an $n \times n$ matrix of right singular vector.



$$A = U S V^T$$

$$= \sigma_1 \begin{bmatrix} u_1 \end{bmatrix} \begin{bmatrix} v_1^T \end{bmatrix} + \sigma_2 \begin{bmatrix} u_2 \end{bmatrix} \begin{bmatrix} v_2^T \end{bmatrix}$$

The decomposition allow us to express our original matrix as a linear combination of low-rank matrices.



in practical application, will observe that only the first few, say k , singular values are large. The rest approach zero. All terms except the first few can be ignored without losing much of the information.

- summarize:
using SVD, we are able to present our large matrix A by 3 matrices: U, S, V .
Helpful in large computations.
can obtain a k -rank approximation of A , through selecting the first k singular value, and truncate the 3 matrices.

- Applications:
First we need to know four things:
 1. SVD is the decomposition of a matrix A into 3 matrices U , S and V .
 2. S : Singular value is the importance values of different features.
 3. rank is measure of the unique information stored in matrix. Higher rank more the information importance.
 4. eigenvectors: directions of maximum spread or variance of data.

Data compression: used to reduce the dimensionality of a dataset by identifying the most important features.

Data denoising: to remove noise from dataset by identifying and removing the least important components or features.

Data visualization: used to project high-dimensional data onto a lower-dimensional space, to make it easier to visualize and understand the relationships between different features in data.

- **Non-linear methods:**

- **Kernel PCA:** uses a kernel function to project dataset into higher dimensional feature space. It is similar to the idea of support vector machine.

- various kernel types:

Linear: is defined as the inner product of two vectors. used to find hyperplane. $k(x,y) = x \cdot y$.

Polynomial : is defined as the inner product of two vectors raised to a Power.

Used to find non-linear decision boundaries. $k(x,y) = (x \cdot y + c)^p$ c:constant. p:degree of the polynomial.

RBF(Radial basis function): (Gaussian Kernel): defined as the exponentiated squared euclidean distance between two vectors. $K(x, y) = \exp(-\gamma \|x - y\|^2)$

gamma: controls the width of the kernels and determines the influence of each training example of decision boundary. A larger value of gamma means that each training example has greater influence of decision boundary, but a large value may lead to overfitting.

is popular choice for kernels, due to ability to model complex relationships and its computational efficiency.

- **T-distributed Stochastic Neighbor Embedding (t-SNE):** is non-linear dimensionality reduction technique. it is to visualize higher-dimensional features in two or three dimensional space.

How does work?

finds a patterns in the data based on the similarity (correlation) of data points with features, the similarity of points is calculated as the conditional probability that a point A would choose point B as its neighbour. then tries to minimize the difference between these conditional Probabilities (or similarities) in higher and lower dimensional space for a perfect representation of data points in lower him space.

- **Multidimensional Scaling (MDS):** used to project high onto low-dimensional space while preserving the pairwise distances between the data points as much as possible. It is based on the concept of distance and aims to find a projection of the data that minimize the differences between the distances in the original space and distances in lower-dimensional space. It is implemented using numerical optimization such as gradient descent or simulated annealing, to minimize the difference between distances.

$$\text{stress} = \sqrt{\frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n (d_{ij} - \hat{d}_{ij})^2}$$

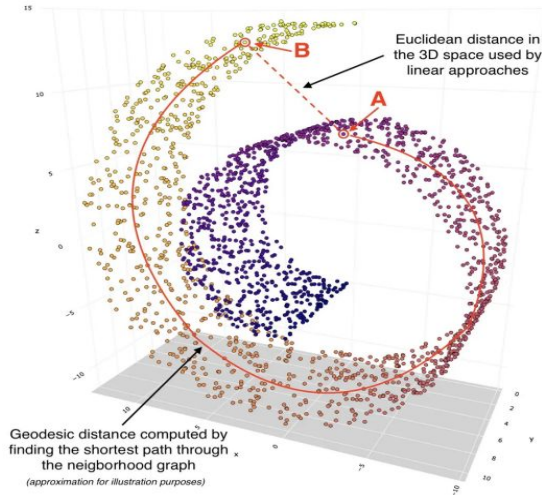
where d_{ij} is the distance between data points i and j in the original space, \hat{d}_{ij} is the distance between data points i and j in the lower-dimensional space, and n is the number of data points. The stress function is a measure of the deviation of the distances in the lower-dimensional space from the distances in the original space and is used to evaluate the quality of the projection.

How Multidimensional Scaling (MDS) is compared to other dimensionality reduction technique techniques?

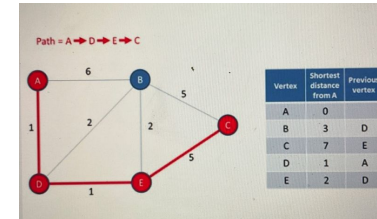
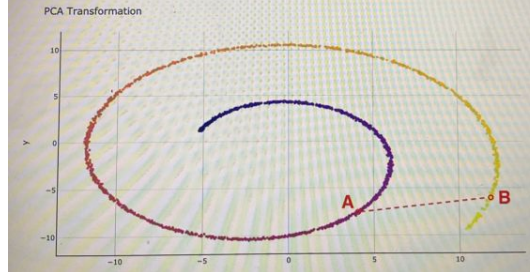
MDS is commonly compared to other dimensionality reduction techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), to understand how it differs from these techniques and when it may be more appropriate to use.

1. MDS is based on the concept of distance and aims to find a projection of the data that minimizes the differences between the distances in the original space and the distances in the lower-dimensional space. In contrast, PCA and t-SNE are based on the concept of variance and entropy, respectively, and aim to find a projection of the data that maximizes the variance or entropy in the lower-dimensional space. This means that MDS is more focused on preserving the relationships between the data points, while PCA and t-SNE are more focused on summarizing the data and finding the most relevant dimensions.
2. MDS can be applied to a wide range of data types, including numerical, categorical, and mixed data. In contrast, PCA and t-SNE are more suited to numerical data, and may not be as effective with categorical or mixed data. This makes MDS a more versatile and flexible technique and allows it to handle complex, multi-modal data sets.
3. MDS uses numerical optimization algorithms to find the projection that minimizes the stress function, and that best preserves the pairwise distances between the data points. In contrast, PCA and t-SNE use linear algebra and stochastic algorithms, respectively, to find the projection that maximizes the variance or entropy in the lower-dimensional space. This means that MDS is a more flexible and adaptable technique, and can find projections that are different from those produced by PCA or t-SNE.

- **Isometric Mapping (Isomap):** is a technique that combines several different algorithms. It to use a nonlinear way to reduce dimensions while preserving local structures.
 1. Use knn approach to find the is nearest neighbors of every data point. K: number of neighbors
 2. Once the neighbors are found, construct the neighborhood graph where points are connected to each other if they are each other's neighbors, Data points that are not neighbors remain unconnected.
 3. Compute the shortest path between each pair of data points it is Dijkstra's algorithm, this step is also described as finding geodesic distance between points.
 4. Use multidimensional scaling (MDS) to compute lower dimensional embedding. Given distances between each pair of points are known MDS, places each object into the n-dim. space such that the between point distances are preserved as possible.



We can see that these two points are relatively close to each other within the 3D space. If we used a linear dimensionality reduction approach such as PCA, then the Euclidean distance between these two points would remain somewhat similar in lower dimensions. See PCA transformation chart below:



Dijkstra's Algorithm.