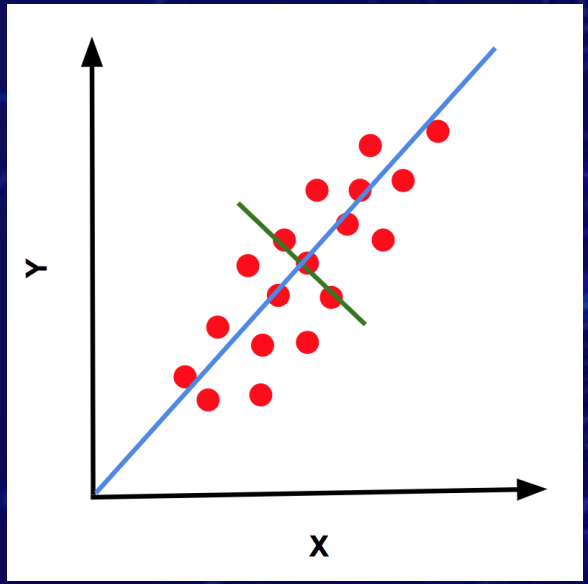
edunet foundation | Proudly supported by SAP | code unnati

Unit 3.4
Principal Component Analysis in Machine Learning



Reference

Disclaimer

The content is curated from online/offline resources and used for educational purpose only

Applications

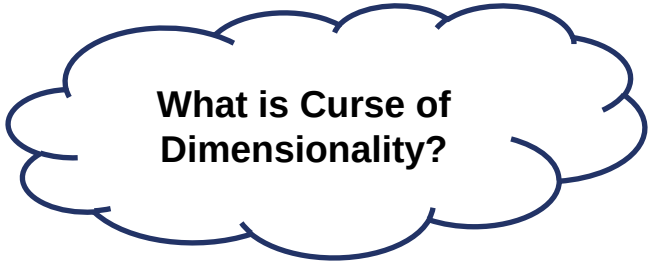
- Image compression
- Customer Profiling
- Food science field
- Banking Field
- Customer Perception
- Finance Field
- Healthcare Industries

Learning Objectives

- Introduction
- What is Curse of Dimensionality?
- Why Curse of Dimensionality?
- Dimensionality Reduction Techniques
- Principal Component Analysis
- Principal Component Analysis Example
- Steps for PCA Algorithm
- Applications of PCA in Machine Learning
- Hands On

Introduction

- Data forms the foundation of any machine learning algorithm, without it, Data Science can not happen.
- Machine Learning in general works wonders when the dataset provided for training the machine is large and concise.
- Usually having a good amount of data lets us build a better predictive model since we have more data to train the machine with.
- However, using a large data set has its own pitfalls.
- The biggest pitfall is the curse of dimensionality.



What is Curse of Dimensionality?

- Suppose there is a dataset with 50 features. Now let's assume you intend to build various separate machine learning models from this dataset.
- The difference between these models is the number of features.

Model-1	Model-2	Model-3	Model-4
No of Features=10	No of Features=20	No of Features=35	No of Features=45
Accuracy=50%	Accuracy=80 %	Accuracy=72%	Accuracy=62%

- The model-2 has more information than model-1 because its number of features is comparatively higher. So, the accuracy of model-2 is more than that of model-1.

What is Curse of Dimensionality?….

- With the increase in the number of features, the model's accuracy increases.
- However, after a specific threshold value, the model's accuracy will not increase, although the number of features increases.
- Because a model is fed with a lot of information, making it incompetent to train with correct information.
- The phenomenon when a machine learning model's accuracy decreases, although increasing the number of features after a certain threshold, is called the curse of dimensionality.
- The curse of dimensionality can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data.

Hughes Phenomenon

- The Hughes Phenomenon shows that as the number of features increases, the classifier's performance increases as well until we reach the optimal number of features.
- Adding more features based on the same size as the training set will then degrade the classifier's performance.

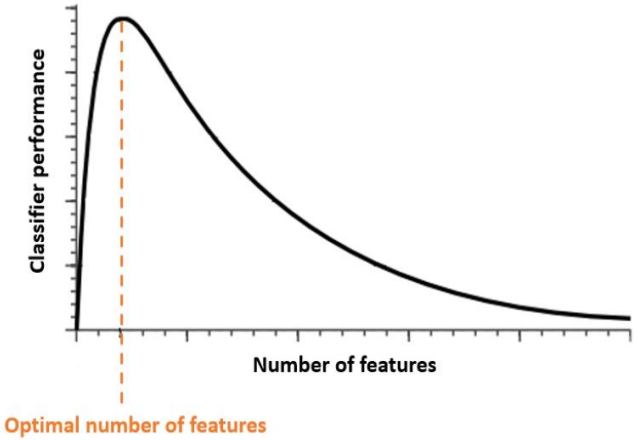


Image: Hughes Phenomenon

Why Curse of Dimensionality?

- Data sparsity is an issue that arises when you go to higher dimensions. Because the amount of space represented grows so quickly that data can't keep up, it becomes sparse, as seen below.

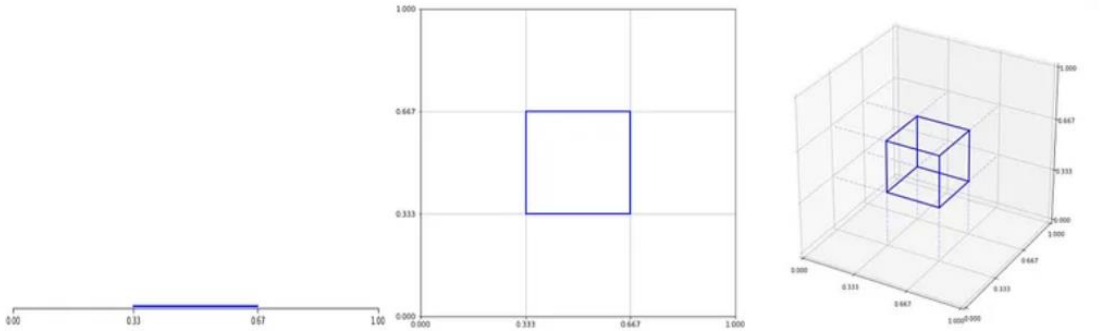
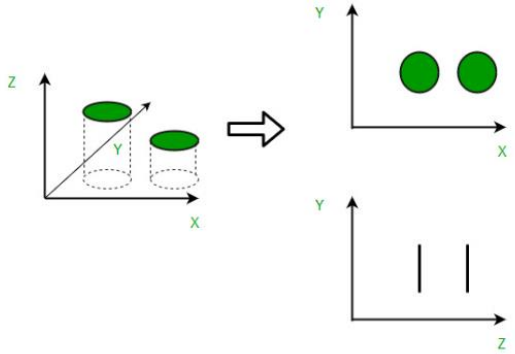


Image: Difference between the space acquired against total space (1/3 , 1/9, 1/27)

How do we lift the curse of dimensionality?

- The solution to Curse of dimensionality is dimensionality reduction.
- Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible.
- In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data.

In the given figure , where a 3-D feature space is split into two 2-D feature spaces, and 2-D space into simple line



Dimensionality Reduction Techniques

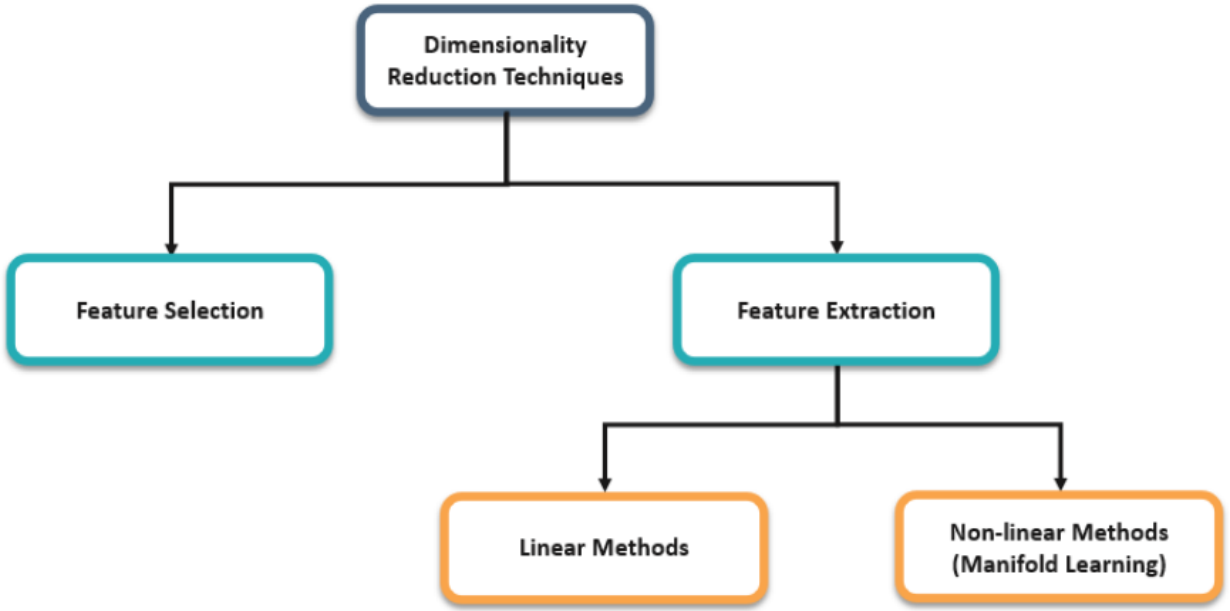
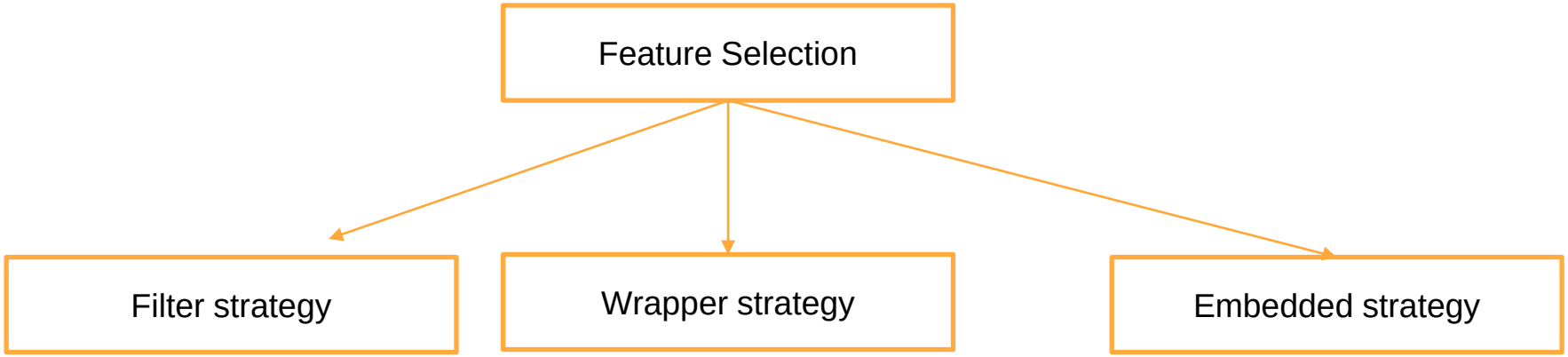


Image: Dimensionality Reduction Techniques

Feature Selection

- The feature selection method aims to find a subset of the input variables (that are most relevant) from the original dataset.
- For example: Dropping Feature having 70% null values.



Feature Selection Example

- Consider a table which contains information on old cars. The model decides which cars must be crushed for spare parts.

Model	Year	Miles	Owner

Model	Year	Miles

- Dropping Owner feature from dataset

Feature Extraction

- Feature extraction is a dimensionality reduction technique.
- Unlike feature selection, which selects and retains the most significant attributes, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes.
- Feature extraction projects a data set with higher dimensionality onto a smaller number of dimensions.
- It is useful for data visualization, since a complex data set can be effectively visualized when it is reduced to two or three dimensions.

Feature Extraction Algorithms

- Here are some famous feature extraction algorithms:
  - Principal Component Analysis
  - Linear Discriminant Analysis
  - Generalized Discriminant Analysis (GDA)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

Principal Component Analysis (PCA)

- Principal Component Analysis or PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into smaller k (k<<p) number of uncorrelated variables called principal components while retaining as much of the variation of the original data as possible.

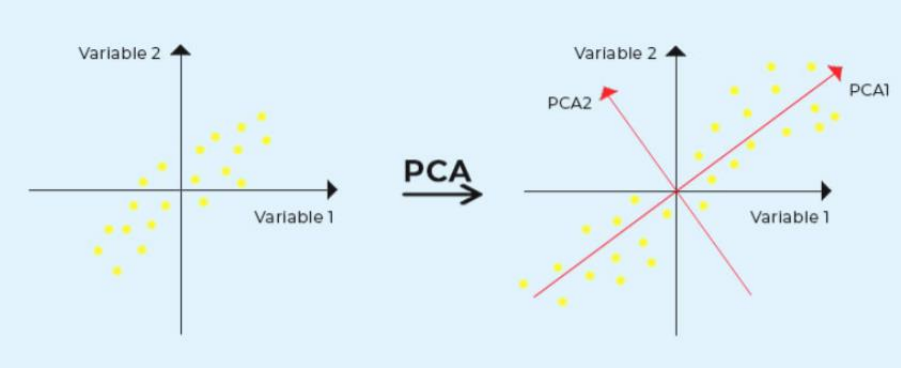


Image: PCA

Principal Component Analysis (PCA).....

- PCA finds a lower-dimensional representation of data by constructing new features (called principal components) which are linear combinations of the original features.
- The Principal Components are a straight line that captures most of the variance of the data. Principal Components have a direction and magnitude.

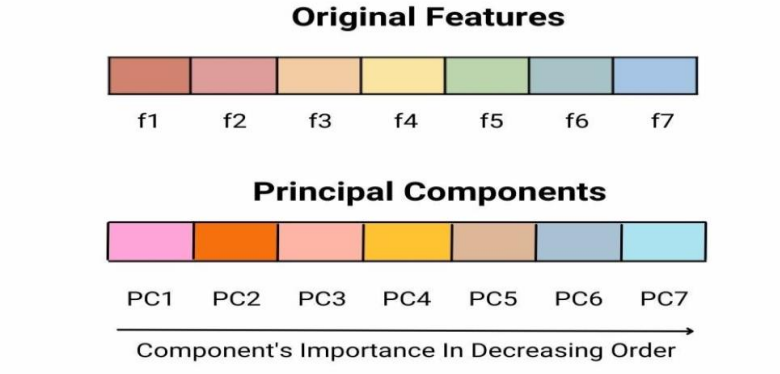
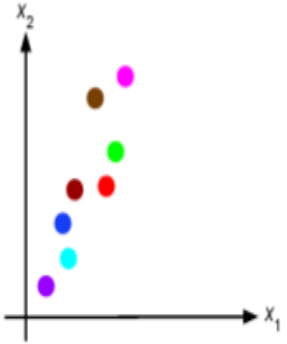


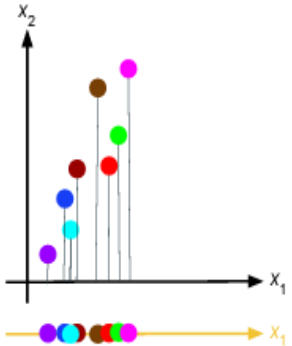
Image: Principal Component

Principal Component Analysis Example

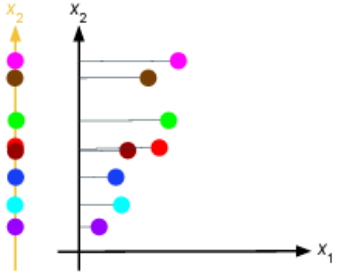
- Let us imagine we have a dataset containing 2 different dimensions. Let the dimensions be X1 and X 2 as given below.
- We are trying to reduce 2 dimension to 1 dimension which has as maximum variance.



Scatter Plot of Data



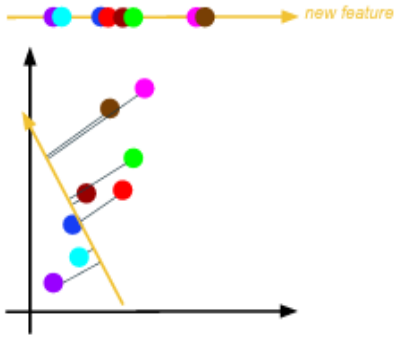
Projection of data directly to X -Axis



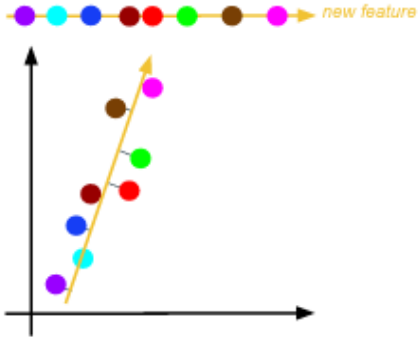
Projection of data directly to Y -Axis

Principal Component Analysis Example

- PCA try to find unit vector in new axis which has maximum variance



Rejected
(Not Maximum Variance)



PC1
(Maximum Variance)

Steps for PCA Algorithm

1. Standardize the data: PCA requires standardized data, so the first step is to standardize the data to ensure that all variables have a mean of 0 and a standard deviation of 1.

$$Z = \frac{X - mean(X)}{\sigma(X)}$$

2. Calculate the covariance matrix: The next step is to calculate the covariance matrix of the standardized data. This matrix shows how each variable is related to every other variable in the dataset. Given below covariance matrix for three X , Y , Z features.

$$\text{Covariance of data} = \begin{bmatrix} cov(X,X) & cov(Y,X) & cov(Z,X) \\ cov(X,Y) & cov(Y,Y) & cov(Z,Y) \\ cov(X,Z) & cov(Y,Z) & cov(Z,Z) \end{bmatrix}$$

Image: Covariance Matrix of three variable

Steps for PCA Algorithm…..

3. Calculate the eigenvectors and eigenvalues: The eigenvectors and eigenvalues of the covariance matrix are then calculated. The eigenvectors represent the directions in which the data varies the most, while the eigenvalues represent the amount of variation along each eigenvector.

4. Choose the principal components: The principal components are the eigenvectors with the highest eigenvalues. These components represent the directions in which the data varies the most and are used to transform the original data into a lower-dimensional space.

5. 5. Transform the data: The final step is to transform the original data into the lower-dimensional space defined by the principal components.

Lab 1 Step by step implementation of PCA using NumPy and pandas

Lab 2 Implementation of PCA using inbuilt function

Summary

- Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data into a lower-dimensional representation. It achieves this by creating new features called principal components. These components are linear combinations of the original features and capture the most significant patterns or variations within the data.

- The process begins by constructing the first principal component, which captures the maximum variance in the data. Subsequent components capture decreasing amounts of variance while being orthogonal to the earlier ones. Principal components can be visualized as axes in the high-dimensional space, representing directions that capture data variance. Each component has a direction and magnitude, indicating its alignment and significance in explaining the data's variance.

- PCA's key role is to simplify complex data by extracting and condensing vital information into a lower-dimensional space. This aids in visualization, interpretation, and analysis. Understanding principal components' direction, magnitude, and significance empowers us to effectively reduce data dimensionality while retaining core insights

Quiz

Question 1: What is the primary purpose of Principal Component Analysis (PCA)?

A) To increase the dimensionality of the data.

B) To identify outliers in the dataset.

C) To transform data into a lower-dimensional representation while retaining essential information.

D) To select all features from the original dataset.

Quiz

Question 1: What is the primary purpose of Principal Component Analysis (PCA)?

A) To increase the dimensionality of the data.

B) To identify outliers in the dataset.

C) To transform data into a lower-dimensional representation while retaining essential information.

D) To select all features from the original dataset.

Answer: C) To transform data into a lower-dimensional representation while retaining essential information.

Quiz

Question 2: What are the principal components in PCA?

A) New data points that are generated from the original data.

B) New features that capture the most significant patterns or variations within the data.

C) The highest frequency components in the data.

D) Features that are uncorrelated with each other.

Quiz

Question 2: What are the principal components in PCA?

A) New data points that are generated from the original data.

B) New features that capture the most significant patterns or variations within the data.

C) The highest frequency components in the data.

D) Features that are uncorrelated with each other.

Answer: B) New features that capture the most significant patterns or variations within the data.

Quiz

Question 3: What does the first principal component capture in PCA?

A) The least variance in the data.

B) The maximum variance in the data.

C) The average of the data.

D) The median of the data.

Quiz

Question 3: What does the first principal component capture in PCA?

A) The least variance in the data.

B) The maximum variance in the data.

C) The average of the data.

D) The median of the data.

Answer: B) The maximum variance in the data.

Quiz

Question 4: What is the relationship between principal components and data variance in PCA?

A) Principal components are random vectors unrelated to data variance.

B) Principal components are perpendicular to each other and do not capture variance.

C) Principal components capture the directions of maximum variance in the data.

D) Principal components are inversely proportional to data variance.

Quiz

Question 4: What is the relationship between principal components and data variance in PCA?

A) Principal components are random vectors unrelated to data variance.

B) Principal components are perpendicular to each other and do not capture variance.

C) Principal components capture the directions of maximum variance in the data.

D) Principal components are inversely proportional to data variance.

Answer: C) Principal components capture the directions of maximum variance in the data.

Quiz

Question 5: What advantage does Principal Component Analysis (PCA) offer in data analysis?

A) It increases the complexity of the dataset.

B) It generates new features that are highly correlated with the original ones.

C) It helps in identifying outliers in the dataset.

D) It simplifies data visualization and interpretation by reducing dimensionality.

Quiz

Question 5: What advantage does Principal Component Analysis (PCA) offer in data analysis?

A) It increases the complexity of the dataset.

B) It generates new features that are highly correlated with the original ones.

C) It helps in identifying outliers in the dataset.

D) It simplifies data visualization and interpretation by reducing dimensionality.

Answer: D) It simplifies data visualization and interpretation by reducing dimensionality.

References

- https://www.geeksforgeeks.org/principal-component-analysis-pca/
- https://www.analyticsvidhya.com/blog/2022/07/principal-component-analysis-beginner-friendly/
- https://www.javatpoint.com/principal-component-analysis

Thank you...!