Module 4

Dimensionality Reduction: Overview

Dimensionality Reduction Overview

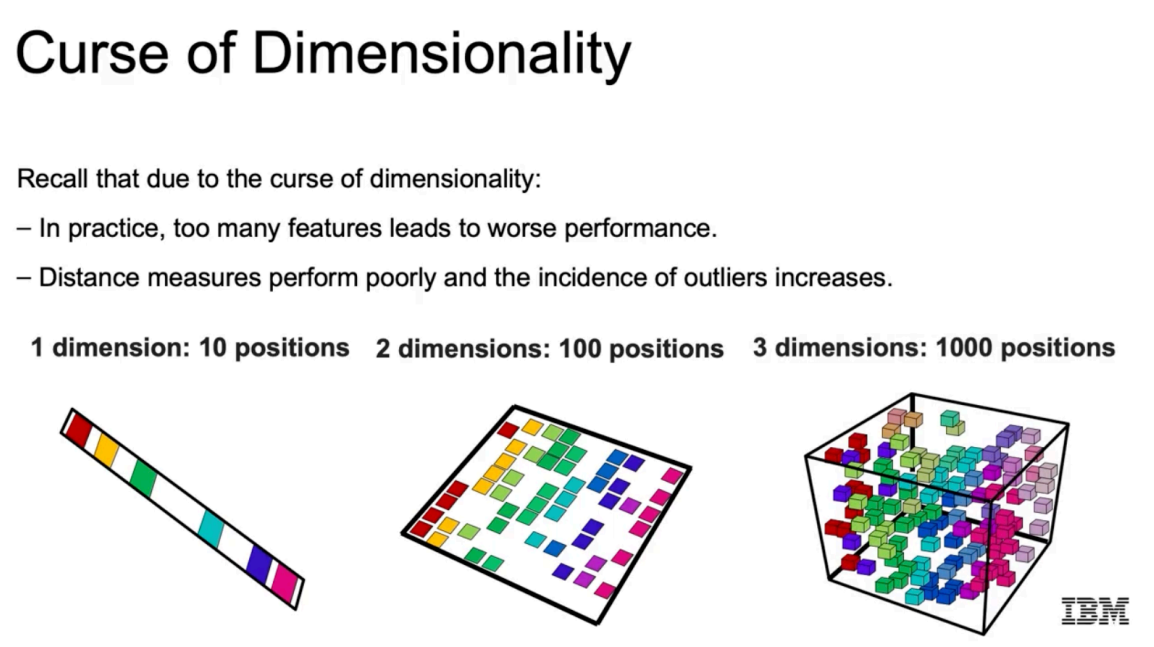
- The goal is to address the curse of dimensionality by creating lower-dimensional representations of data.
- Techniques like Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) are introduced to reduce dimensions while maintaining data integrity.

Principal Component Analysis (PCA)

- PCA transforms original features into new features in a lower-dimensional space, preserving variance.
- It involves projecting data points onto a line derived from correlated features, effectively reducing dimensions.

Curse of Dimensionality

- As dimensions increase, the number of observations needed to cover the data space grows exponentially, complicating model performance.
- Reducing dimensions can improve model efficiency and reduce the incidence of outliers in high-dimensional datasets



Dimensionality Reduction: Principal Component Analysis

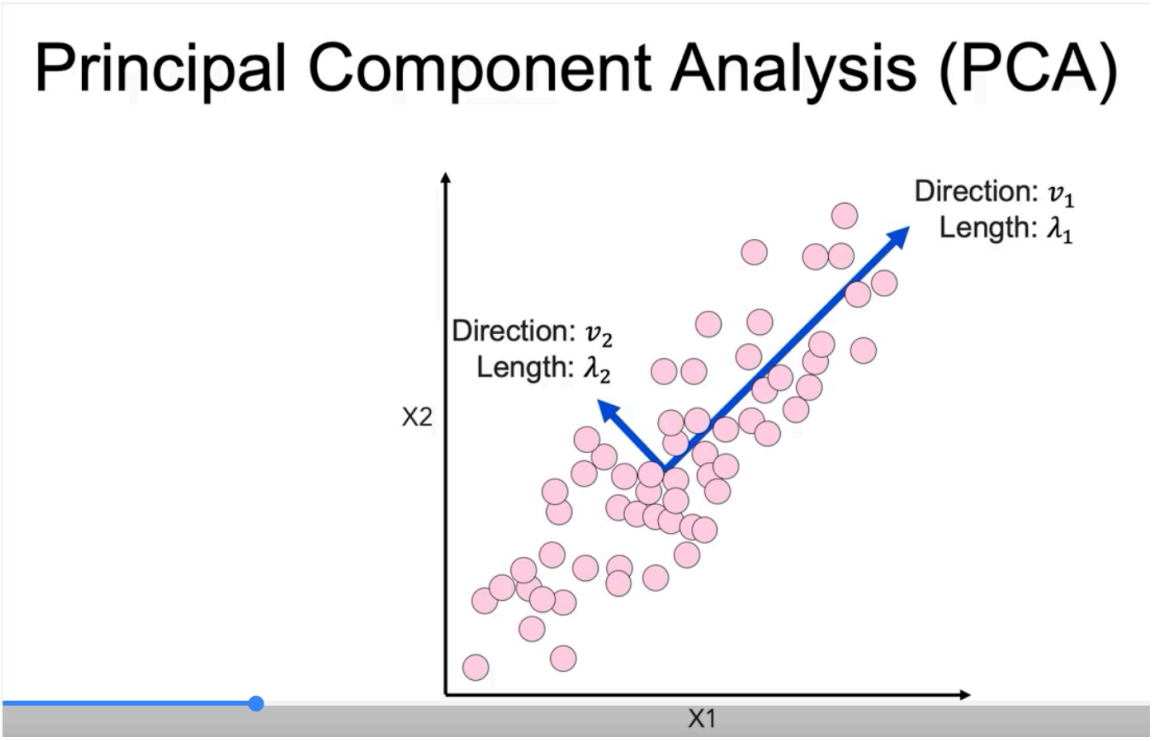
- PCA reduces the dimensionality of datasets by identifying the primary directions (vectors) of variance in the data.
- The primary right singular vector captures the most variance, while the secondary right singular vector provides additional axes for data representation.

Singular Value Decomposition (SVD)

- SVD decomposes a dataset matrix into three matrices (U, S, V), where S contains the lengths of the vectors indicating their importance.
- The diagonal values in S are sorted, with larger values indicating more significant principal components.

Applying PCA with Scikit-learn

- PCA can be implemented using Scikit-learn by specifying the number of components to reduce the dataset.
- Proper scaling of data is crucial before applying PCA to ensure accurate projections and maintain variance.



Single Value Decomposition (SVD)

– SVD is a matrix factorization method normally used for PCA.

– Does not require a square data set.

– SVD is used by Scikit-learn for PCA.