Mathematical Foundations-I (Probability & Statistics) [AI&DS-ST106]

PROJECT REPORT

BY SNEHA SHINDE
RISHITA MERCHANT
SANSKRITI DHAR

Singular Value Decomposition (SVD)

Singular Value Decomposition is a matrix factorization technique that decomposes a matrix into three constituent matrices: U, Σ, and V^T. Given a data matrix X, SVD allows us to represent it as $X = U\Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix containing the singular values. SVD can be utilized for dimensionality reduction by truncating the singular value matrix Σ, keeping only the most significant singular values and their corresponding columns in U and V.

Applications Of SVD:

1. **Dimensionality Reduction** -SVD helps reduce dimensionality by projecting high-dimensional data onto a lower-dimensional space, capturing the most significant patterns. It retains the most important information while filtering out noise, making data easier to visualize and model.
2. **Latent Semantic Analysis** - SVD in Latent Semantic Analysis (LSA) is a powerful technique in NLP for understanding the hidden meanings and relationships within text. It's used to reduce the dimensionality of a document-term matrix, identify latent concepts, and facilitate tasks like document clustering, classification, and information retrieval
3. **Feature extraction** - SVD extracts features by decomposing a data matrix into orthogonal components, retaining only the top k singular values and vectors that capture the most variance.
   This reduces dimensionality, denoises data, and yields new uncorrelated features that improve model performance
4. **AI and ML** - SVD is used in AI and ML to simplify complex data by breaking it into simpler components. It reduces data noise and redundancy, helping algorithms focus on the most important patterns. Common uses include data compression, feature selection, and improved model performance. It also helps reveal hidden structures, like latent topics in text analysis or collaborative filtering in recommendation systems. Overall, SVD makes data easier to interpret, manage, and model, leading to better insights and predictions.

Principal Component Analysis (PCA)

PCA (Principal Component Analysis) is a dimensionality reduction technique used in data analysis and machine learning. It helps you to reduce the number of features in a dataset while keeping the most important information. It changes your original features into new features these new features don't overlap with each other and the first few keep most of the important differences found in the original data.

PCA is commonly used for data preprocessing for use with machine learning algorithms. It helps to remove redundancy, improve computational efficiency and make data easier to visualize and analyse especially when dealing with high-dimensional data.

Advantages of PCA

1. Dimensionality reduction: By determining the most crucial features or components, PCA reduces the dimensionality of the data, which is one of its primary benefits. This can be helpful when the initial data contains a lot of variables and is therefore challenging to visualize or analyze.
2. Feature Extraction: PCA can also be used to derive new features or elements from the original data that might be more insightful or understandable than the original features. This is particularly helpful when the initial features are correlated or noisy.
3. Data visualization: By projecting the data onto the first few principal components, PCA can be used to visualize high-dimensional data in two or three dimensions. This can aid in locating data patterns or clusters that may not have been visible in the initial high-dimensional space.
4. Noise Reduction: By locating the underlying signal or pattern in the data, PCA can also be used to lessen the impacts of noise or measurement errors in the data.
5. Multicollinearity: When two or more variables are strongly correlated, there is multicollinearity in the data, which PCA can handle. PCA can lessen the impacts of multicollinearity on the analysis by identifying the most crucial features or components.

Applications of PCA:

1. PCA is used to visualize multidimensional data.
2. It is used to reduce the number of dimensions in healthcare data.
3. PCA can help resize an image.
4. It can be used in finance to analyze stock data and forecast returns.
5. PCA helps to find patterns in the high-dimensional datasets.

Limitations of PCA

1. Interpretability: Although principal component analysis (PCA) is effective at reducing the dimensionality of data and spotting patterns, the resulting principal components are not always simple to understand or describe in terms of the original features.
2. Information loss: PCA involves choosing a subset of the most crucial features or components in order to reduce the dimensionality of the data. While this can be helpful for streamlining the data and lowering noise, if crucial features are not included in the components chosen, information loss may also result.
3. Outliers: Because PCA is susceptible to anomalies in the data, the resulting principal components may be significantly impacted. The covariance matrix can be distorted by outliers, which can make it harder to identify the most crucial characteristics.
4. Scaling: PCA makes the assumption that the data is scaled and centralized, which can be a drawback in some circumstances. The resulting principal components might not correctly depict the underlying patterns in the data if the data is not scaled properly.
5. Computing complexity: For big datasets, it may be costly to compute the eigenvectors and eigenvalues of the covariance matrix. This may restrict PCA's ability to scale and render it useless for some uses.