

Unsupervised Machine Learning

Assignment 1

SAI KUMAR MURARSHETTI

LEWIS ID: L30079224

Principal Component Analysis (PCA) It is a dimensionality reduction approach that is frequently used in unsupervised machine learning. It is used to identify patterns in data and present them in a way that highlights the differences between them.

Data Preparation: PCA begins with a dataset containing m observations of n variables. These variables could represent features or attributes of the data.

Mean Centered: PCA frequently get by centering the data, which means subtracting each variable's mean from all observations. This phase centers the data around its starting point.

Correlation Matrix Determination: After centering the data, PCA generates the correlation matrix. This matrix shows the relationships between each pair of variables. The correlation of two variables indicates that they vary together. A positive covariance indicates the variables increase or decrease simultaneously, whereas a negative covariance shows that one variable tends to increase while the other decreases.

Eigenvalue Decomposition: The correlation matrix is then separated into its eigenvalues. This method generates eigenvectors and eigenvalues. Eigenvectors are the directions along which the data varies the greatest, and eigenvalues represent the degree of the variation in those directions.

Principal Component Analysis (PCA): ranks the eigenvectors in descending order according to their specific eigenvalues. The eigenvector with the highest eigenvalue shows the direction of largest variance in the data and is known as the initial principal component. The following principal components contain decreasing amounts of variance and generally perpendicular to the previous components.

Dimension Reduction: PCA allows users to reduce the dimensionality of the information by selecting a subset of its most important principal components. Usually, the first k principal components represent a significant percentage of the total variance (e.g., 95%). Estimating data into principal components reduces dimensions from n to k.

Data Restoration: If required, this reduced-dimensional data can be converted into its original feature representation. This reconstruction includes dividing the predicted information by the transpose of the specified principal components and restoring the mean which was removed during mean centered.

Unsupervised Machine Learning

Assignment 1

SAI KUMAR MURARSHETTI

LEWIS ID: L30079224

PCA is commonly used for several instances, including data visualization, noise reduction, feature extraction and increasing subsequent training algorithms by reducing the dimensionality of the data being analyzed. But it is important to remember that PCA assumes linear correlations between variables and may not perform well when non-linear interactions present. In these cases, nonlinear dimensionality reduction approaches such as t-distributed sequential network encoding (t-SNE) or autoencoders may be more suitable.

Linear Discriminant Analysis (LDA) is a technique used in both supervised and unsupervised machine learning though it is most typically associated with supervised learning. a situation such as the context of unsupervised learning, LDA can be used to reduce dimensionality and extract features.

In unsupervised learning, LDA identifies the linear combinations of features which most effectively separate the classes in the data without the use of class labels. It aims to convey the data into a lower-dimensional domain while retaining the greatest possible separation between classes.

Data Preparation: Identical to various reduction of dimensionality approaches, LDA begins with a dataset of observations with many features.

Within-class Scatter Matrix: LDA produces overall within-class scatter matrix, which represents the arrangement of data inside each class. It is calculated by adding together the scatter matrices for every class.

Between-class Scatter Matrix: LDA also generates a between-class scatter matrix, which measures the variance of various classes. It is generated through comparing averages of several classes.

Eigenvalue Decomposition: The next step is performing an eigenvalue decomposition on the aggregate within- and between-class scatter matrices. This generates eigenvectors and eigenvalues.

Dimensionality Reduction: LDA selects the eigenvectors with largest eigenvalues. These eigenvectors represent the directions in which the data is most easily separated into classes. The data is eventually displayed on the eigenvectors, resulting in dimensionality reduction.

Unsupervised Machine Learning

Assignment 1

SAI KUMAR MURARSHETTI
LEWIS ID: L30079224

While Linear Discriminant Analysis (LDA) in unsupervised learning does not generally include class labels during dimensionality reduction, the reduced-dimensional data it generates may serve as an input for task classification in a supervised learning environment.

LDA is particularly useful when class labels are known upfront of time or when maximum class separability is required. It is useful in visualizing high-dimensional data, identifying key features and improving classification algorithm performance. whereas like PCA, LDA assumes linear correlations between variables so can overlook complex non-linear patterns in the data. In certain situations, nonlinear dimensionality reduction techniques such as multivariate and deep learning approaches may be more suitable.