Class Note, Daniel Zalewski

1. **Dimensionality reduction**:
2. **Definition** – Dimensionality reduction involves transforming data from a high-dimensional space to a lower-dimensional space while retaining essential data characteristics.
3. **Objective** – The main goal of dimensionality reduction is to simplify data analysis by reducing the number of variables while preserving relevant information.
4. **Motivation** - High dimensionality can lead to data sparsity, increased computational complexity, and overfitting issues.
5. **Methods**:
* **Principal component analysis (PCA)** – Performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In practice, the covariance matrix of the data is constructed and the eigenvectors on this matrix are computed.
* **Non-negative matrix factorization (NMF)** – The Non-Negative Matrix Factorization method in dimensionality reduction is a technique primarily used in data analysis and machine learning. It allows for the decomposition of non-negative matrices into the product of two non-negative matrices, typically of smaller dimensions.
* **Kernel PCA** – Kernel PCA is extension of the traditional PCA method that allows for non-linear dimensionality reduction. While PCA find linear combinations of features that capture the maximum variance in the data, Kernel PCA extends this capability to handle non-linear relationships by using a kernel trick.
* **Graph-based kernel PCA** – Graph-based kernel PCA is an extension of kernel PCA that incorporates graph theory concepts into dimensionality reduction process. This approach combines the power of kernel PCA with the ability to capture the intrinsic structure of the data as presented by graph.
* **Linear discriminant analysis (LDA)** – Linear discriminant analysis is a technique used to find linear combinations of features that best separate different classes or groups in the data. It is a supervised technique, meaning it utilizes information about class membership during the learning of data transformation.
* **Generalized discriminant analysis (GDA)** – GDA is an extension LDA that allows for more flexibility in handling non-linear relationships between features and class labels. While LDA assumes that the features allow a multivariate Gaussian distribution within each class, GDA relaxes this assumption by allowing for different types of distribution within each class.
* **Autoencoder** – An autoencoder is a type of neural network used to learn an efficient data representation by training the encoding and decoding of data.
* **t-distributed Stochastic Neighbor Embedding (t-SNE)** – t-SNE is a technique used for dimensionality reduction and visualization of high-dimensional data in lower-dimensional space, typically two or three dimensions. Unlike linear methods such as PCA, t-SNE focuses on preserving local relationships between data points rather than global structure.
* **Uniform Manifold Approximation and Projection (UMAP)** – UMAP is a dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space. UMAP aims to preserve both local and global structure of data, making it effective for a wide range of datasets.
1. **Applications** – Dimensionality reduction is applied in fields such as signal processing, pattern recognition, bioinformatics, and others where large amounts of data or variables are present.
2. **Benefits** – By reducing dimensionality, it becomes easier to visualize data, speed up, analysis, reduce model errors and improve result interpretability.
3. **Challenges** – Dimensionality reduction requires selecting appropriate techniques and considering potential model errors, and improve result interpretability.
4. **Interpretation** – After applying dimensionality reduction, it’s crucial to interpret and considering potential information loss when reducing the number of dimensions.
5. **Principal component analysis (PCA)**:
6. **Dimensionality reduction with PCA** – PCA allows for the reduction of data dimensionality by transforming the original variables *X1*, *X2*, …, *Xp* into a new set of variables called principal components *PC1*, *PC2*, …, *PCk* where *k* is the number of components chosen to be retained. Each observation *i* is represented by *k* principal components.
7. **Mathematical Foundations of PCA**:
* **Eigenvalue Analysis** - Eigenvalues of the covariance matrix ∑ represent the variance explained by each principal component.
* **Eigenvector Analysis** - The eigenvectors corresponding to the eigenvalues λ*i* represent the directions in the new variable space (principal components) along which the data have the most variability.
* **Data Transformation** – Data are transformed from the original space to the principal component space using matrix *V*, which contains eigenvectors as columns. The transformation can be expressed as: *Y = X ∙ V*, where *Y* is the data in the new space.
1. **Interpretation of Principal Components**:
* The first principal component *PC1* contains the highest variance of the data. It can be interpreted as the direction in space that best explains the variability in the data.
* Subsequent principal components explain decreasing portions of the variance in the data, but may still represent significant structures or patterns.
1. **Applications of PCA**:
* **Dimensionality Reduction** – By selecting only a few principal components, one d=can effectively reduce the number of dimensions in the data while retaining important information.
* **Data Visualization** – Principal components can be used to visualize data in lower-dimensional space, facilitating understanding of the data structure.
* **Noise Reduction** – Insignificant components with low variance can be discarded, leading to noise reduction in the data.
* **Clustering** – PCA can aid in identifying groups or clusters in data by analyzing their structure in the principal component space.
1. **Latent variable model:**
* A latent variable model is a statistical model that involves unobservable variables, known as latent variables, which are inferred from observable variables. Those models are used to represent complex relationships among observed variables by assuming the existence of hidden factors that influence the observed data.
1. **Concept of Latent Variables** – Latent variables are variables that are not directly observed but are inferred from observed variables. They represent underlying concepts, traits, or factors that cannot be measured directly but are assumed to affect the observed data.
2. **Components of Latent Variable Models**:
* **Observed Variables**: These are the variables that are directly measured or observed in the dataset. They are also referred to as manifest variables.
* **Latent Variables**: These are unobserved variables that are assumed to influence the observed variables. They represent the underlying structure or patterns in the data.
1. **Assumptions**:
* Latent variable models typically assume that the observed variables are influenced by the latent variables and possibly by error terms.
* They also assume certain relationships or dependencies among the observed variables and between the observed and latent variables.
1. **Types of Latent Variable Models**:
* **Factor Analysis** – In factor analysis, the observed variables are assumed to be linear combinations of smaller number of unobserved factors. It is commonly used for dimensionality reduction and identifying underlying dimensions in the data.
* **Structural Equation Modeling (SEM)** – SEM is a multivariate statistical technique that incorporates both observed and latent variables. It allows for the estimation that represent the hypothesized relationships.
* **Latent Class Analysis (LCA)** – LCA is used to identify unobserved subgroups or classes within a population based on patterns of responses to observed categorical variables.
* **Latent Dirichlet Allocation (LDA)** – LDA is a generative probabilistic model used for topic modeling in text data. It assumes that documents are generated from a mixture of latent topics, and each word is associated with one of these topics.
* **Hidden Markov Models (HMMs)** – HMMs are used to model time-series data with latent states. They assume that observed data are generated by a sequence of hidden states, and the transitions between states are governed by probability distributions.
1. **Applications**:
* Latent variable models are widely used in various fields, including psychology, sociology, economics, biology, and natural language processing.
* They are used for tasks such as modeling human behavior, analyzing survey data. Understanding market segmentation, clustering documents, and predicting future states in time-series data.

Bibliography:

* James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
* https://en.wikipedia.org/wiki/Dimensionality\_reduction
* https://en.wikipedia.org/wiki/Principal\_component\_analysis
* https://en.wikipedia.org/wiki/Latent\_variable\_model