Dimensionality reduction

- We can visualize 2D or 3D data using scatter plots, But what about 4D, 5D or nD, we can use pair plots to do that but up to what extent we can analyze pair plots, At max we can do that till 6D. Then what after that, here comes Dimensionality reduction.
- Dimensionality reduction is a technique to reduce the nD data to 2D or 3D where we can visualize the data.With more variables, comes more trouble! And to avoid this trouble, dimension reduction techniques comes to the rescue.

There are two components of dimensionality reduction

Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:

1. Filter
2. Wrapper
3. Embedded

Feature extraction: This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Dimensionality reduction

Benefits of Dimension Reduction

- Space required to store the data is reduced as the number of dimensions comes down
- Less dimensions lead to less computation/training time
- Some algorithms do not perform well when we have a large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful
- It takes care of multicollinearity by removing redundant features.
- Reducing the dimensions of data to 2D or 3D may allow us to plot and visualize it precisely.
- It is helpful in noise removal also and as result of that we can improve the performance of models.

Dimensionality reduction

Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep- in practice, some thumb rules are applied.

Dimensionality reduction

Common Dimensionality Reduction Techniques

1. Missing Values: **If we encounter missing values, what we do? Our first step should be to identify the reason then impute missing values/ drop variables using appropriate methods. But, what if we have too many missing values? Should we impute missing values or drop the variables?. If the information contained in the variable is not that much, you can drop the variable if it has more than ~40-50% missing values.**

2. Low Variance: **Let's think of a scenario where we have a constant variable in our data set. Do you think, it can improve the power of model? Ofcourse NOT, because it has zero variance. In case of high number of dimensions, we should drop variables having low variance compared to others because these variables will not explain the variation in target variables.**

3. Decision Trees: **It can be used as a ultimate solution to tackle multiple challenges like missing values, outliers and identifying significant variables.**

Dimensionality reduction

Common Dimensionality Reduction Techniques

4. Random Forest: Random Forest is one of the most widely used algorithms for feature selection. It comes packaged with in-built feature importance so you don't need to program that separately. This helps us select a smaller subset of features.

5. High Correlation: Dimensions exhibiting higher correlation can lower down the performance of model. Moreover, it is not good to have multiple variables of similar information or variation also known as "Multicollinearity".

6. Backward Feature Elimination: In this method, we start with all n dimensions. Compute the sum of square of error (SSR) after eliminating each variable (n times). Then, identifying variables whose removal has produced the smallest increase in the SSR and removing it finally, leaving us with n-1 input features.
Repeat this process until no other variables can be dropped.

Dimensionality reduction

Common Dimensionality Reduction Techniques

7. Forward Feature Selection: method. In this method, we select one variable and analyse the performance of model by adding another variable. Here, selection of variable is based on higher improvement in model performance.

8. Principal Component Analysis (PCA): In this technique, variables are transformed into a new set of variables, which are linear combination of original variables. These new set of variables are known as principle components. They are obtained in such a way that first principle component accounts for most of the possible variation of original data after which each succeeding component has the highest possible variance.The second principal component must be orthogonal to the first principal component.

9. Independent Component Analysis: Independent Component Analysis (ICA) is based on information-theory and is also one of the most widely used dimensionality reduction techniques. The major difference between PCA and ICA is that PCA looks for uncorrelated factors while ICA looks for independent factors.

Dimensionality reduction

Common Dimensionality Reduction Techniques

10. Factor Analysis: Let's say some variables are highly correlated. These variables can be grouped by their correlations i.e. all variables in a particular group can be highly correlated among themselves but have low correlation with variables of other group(s). Here each group represents a single underlying construct or factor. These factors are small in number as compared to large number of dimensions. However, these factors are difficult to observe. There are basically two methods of performing factor analysis. EFA (Exploratory Factor Analysis), CFA (Confirmatory Factor Analysis)

11. Independent Component Analysis: Independent Component Analysis (ICA) is based on information-theory and is also one of the most widely used dimensionality reduction techniques. The major difference between PCA and ICA is that PCA looks for uncorrelated factors while ICA looks for independent factors.

Dimensionality reduction

Common Dimensionality Reduction Techniques

12. t- Distributed Stochastic Neighbor Embedding (t-SNE)

So far we have learned that PCA is a good choice for dimensionality reduction and visualization for datasets with a large number of variables. But what if we could use something more advanced? What if we can easily search for patterns in a non-linear way? t-SNE is one such technique. There are mainly two types of approaches we can use to map the data points:

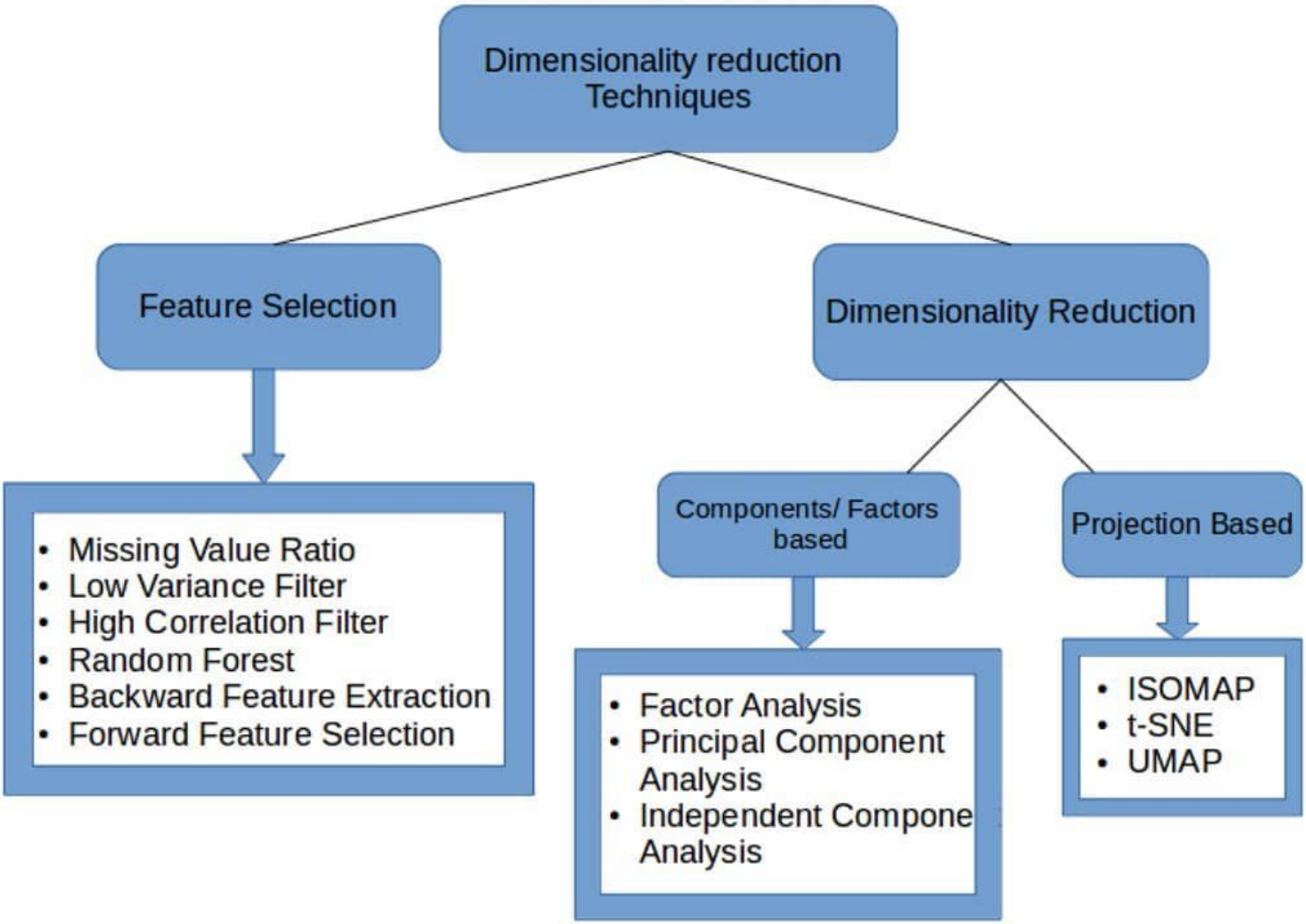
Local approaches : They maps nearby points on the anifold to nearby points in the low dimensional representation.

Global approaches : They attempt to preserve geometry at all scales, i.e. mapping nearby points on manifold to nearby points in low dimensional representation as well as far away points to far away points.

We mostly focus on PCA and t-SNE in upcoming posts

Dimensionality reduction

when to use each Dimensionality Reduction Technique



Source - analyticsvidhya.com