**Dimensional Reduction Principal Component Analysis**

### **What is Dimensionality Reduction?**

In**machine learning** we are having too many factors on which the final classification is done. These factors are basically, known as variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

### **Dimensionality Reduction Methods**

The various methods used for dimensionality reduction include:

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be both linear or non-linear, depending upon the method used. The prime linear method, called Principal Component Analysis, or PCA, is discussed below.

### **Importance of Dimensionality Reduction**

Why is Dimension Reduction is important in machine learning predictive modeling?

The problem of unwanted increase in dimension is closely related to other. That was to fixation of measuring/recording data at a far granular level then it was done in past. This is no way suggesting that this is a recent problem. It has started gaining more importance lately due to a surge in data.

Lately, there has been a tremendous increase in the way sensors are being used in the industry. These sensors continuously record data and store it for analysis at a later point. In the way data gets captured, there can be a lot of redundancy.

### **Common Methods to Perform Dimensionality Reduction**

There are many methods to perform Dimension reduction. I have listed the most common methods below:

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/02/Methods-to-perform-Dimension-Reduction-01.jpg)

*Methods to perform Dimension Reduction*

### **a. Missing Values**

While exploring data, if we encounter missing values, what we do? Our first step should be to identify the reason. Then need to 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?

### **b. Low Variance**

Let’s think of a scenario where we have a constant variable (all observations have the same value, 5) in our data set. Do you think, it can improve the power of model? Of course NOT, because it has zero variance.

### **c. Decision Trees**

It is one of my favorite techniques. We can use it as an ultimate solution to tackle multiple challenges. Such as missing values, outliers and identifying significant variables. It worked well in our Data Hackathon also. Several data scientists used decision tree and it worked well for them.

### **d. Random Forest**

Random Forest is similar to decision tree. Just be careful that random forests have a tendency to bias towards variables that have more no. of distinct values i.e. favor numeric variables over binary/categorical values.

### **e. High Correlation**

Dimensions exhibiting higher correlation can lower down the performance of a model. Moreover, it is not good to have multiple variables of similar information. You can use Pearson correlation matrix to identify the variables with high correlation. And select one of them using VIF (Variance Inflation Factor). Variables having a higher value ( VIF > 5 ) can be dropped.

### **f. Backward Feature Elimination**

In this method, we start with all n dimensions. Compute the sum of a square of error (SSR) after eliminating each variable (n times). Then, identifying variables whose removal has produced the smallest increase in the SSR. And thus removing it finally, leaving us with n-1 input features.

Repeat this process until no other variables can be dropped. Recently in Online Hackathon organized by Analytics Vidhya.

### **g. Factor Analysis**

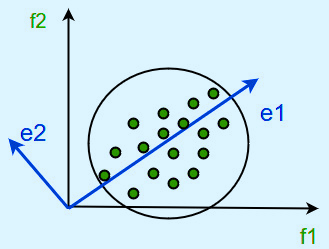
These variables can be grouped by their correlations.. Here each group represents a single underlying construct or factor. These factors are small in number as compared to a large number of dimensions. However, these factors are difficult to observe. There are basically two methods of performing factor analysis:

EFA (Exploratory Factor Analysis)

CFA (Confirmatory Factor Analysis)

### **h. Principal Component Analysis (PCA)**

Particularly, in this we need to transform variables into a new set of variables. As these are a linear combination of original variables. These new set of variables are known as principal components. Further, we need to obtain these in particular way. As first principle component accounts for the possible variation of original data. after which each succeeding component has the highest possible variance.

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/02/image-11.png)

*PCA – Dimensionality Reduction*

The second principal component must be orthogonal to the first principal component. I For two-dimensional dataset, there can be only two principal components. Below is a snapshot of the data and its first and second principal components. Applying PCA to your dataset loses its meaning.

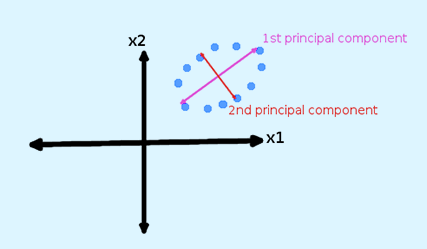
## **Reduce the Number of Dimensions**

* Dimensionality reduction has several advantages from a machine learning point of view.
* Since your model has fewer degrees of freedom, the likelihood of overfitting is lower. The model will generalize more easily to new data.
* If we are using feature selection the reduction will promote the important variables. Also, it helps in improving the interpretability of your model.
* Most of features extraction techniques are unsupervised. You can train your autoencoder or fit your PCA on unlabeled data. This can be helpful if you have a lot of unlabeled data and labeling is time-consuming and expensive.

## **Features Selection in Reduction**

* Most, important is to reduce dimensionality. Also, is to remove some dimensions and to select the more suitable variables for the problem.
* Here are some ways to select variables:
* Greedy algorithms which add and remove variables until some criterion is met.
* Shrinking and penalization methods, which will add cost for having too many variables. For instance, L1 regularization will cut some variables’ coefficient to zero. Regularization limits the space where the coefficients can live in.
* As we have to select model on particular criteria. That need to take the number of dimensions into accounts. Such as the adjusted R², AIC or BIC. Contrary to regularization, the model is not trained to optimize these criteria.
* Filtering of variables using correlation, VIF or some “distance measure” between the features.

## **Advantages of Dimensionality Reduction**

* Dimensionality Reduction helps in data compression, and hence reduced storage space.
* It reduces computation time.
* It also helps remove redundant features, if any.
* Dimensionality Reduction helps in data compressing and reducing the storage space required
* It fastens the time required for performing same computations.
* If there present fewer dimensions then it leads to less computing. Also, dimensions can allow usage of algorithms unfit for a large number of dimensions.
* It takes care of multicollinearity that improves the model performance. It removes redundant features. For example, there is no point in storing a value in two different units (meters and inches).
* Reducing the dimensions of data to 2D or 3D may allow us to plot and visualize it precisely. You can then observe patterns more clearly. Below you can see that, how a 3D data is converted into 2D. First, it has identified the 2D plane then represented the points on these two new axes z1 and z2.
* [](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/02/image7.png)
* Advantages of Dimensionality Reduction
* It is helpful in noise removal also and as a result of that, we can improve the performance of models.

**Disadvantages of Dimensionality Reduction**

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