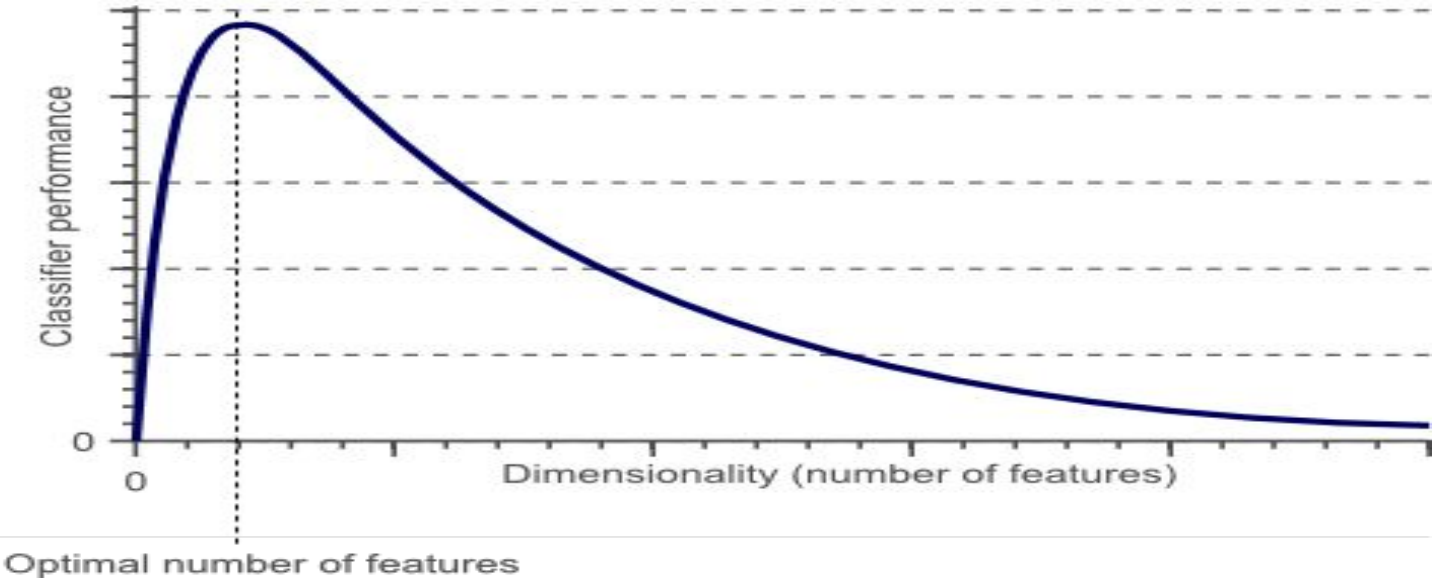
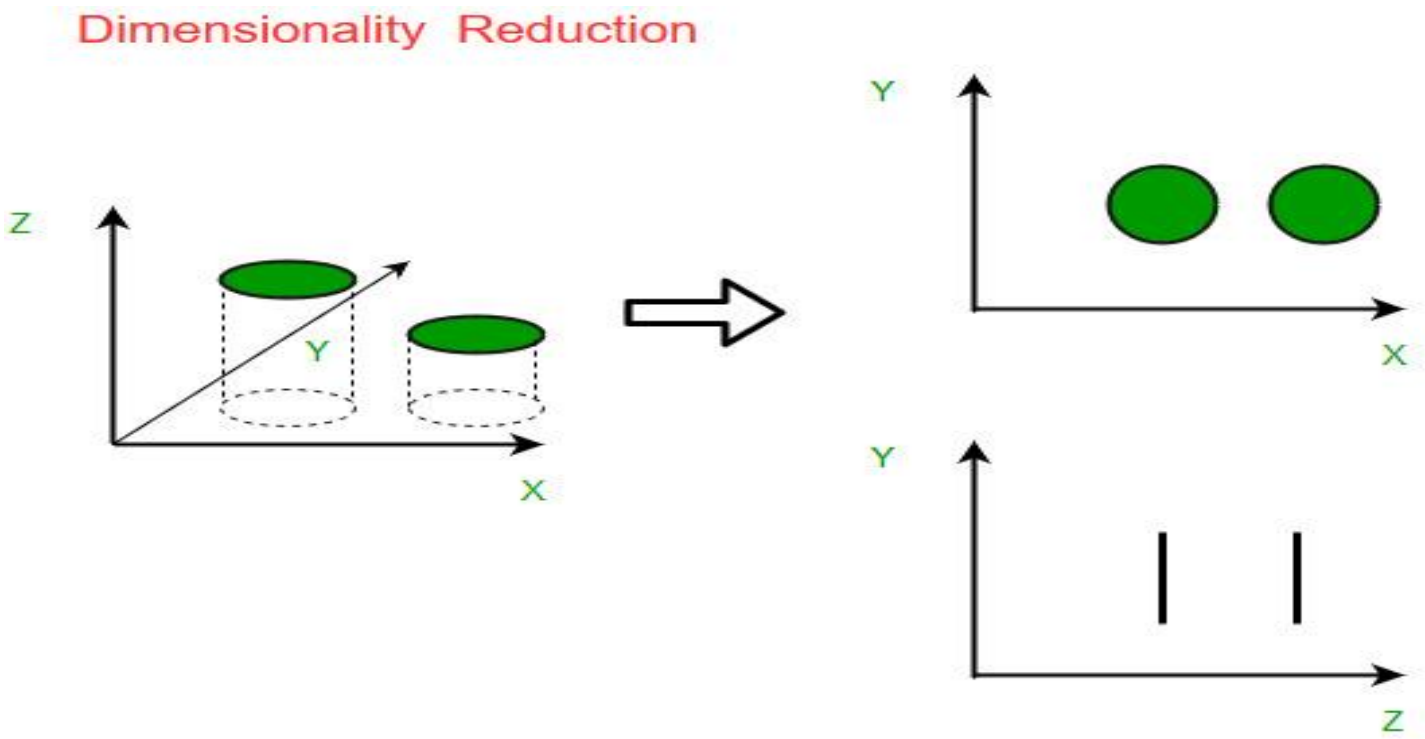
Dimensionality Reduction

- We are generating a tremendous amount of data daily. As data generation and collection keeps increasing, visualizing it and drawing inferences becomes more and more challenging.
- As the number of features increases, the model becomes more complex. The more the number of features, the more the chances of overfitting. A machine learning model that is trained on a large number of features, gets increasingly dependent on the data it was trained on and in turn overfitted, resulting in poor performance on real data, beating the purpose.



Dimensionality Reduction

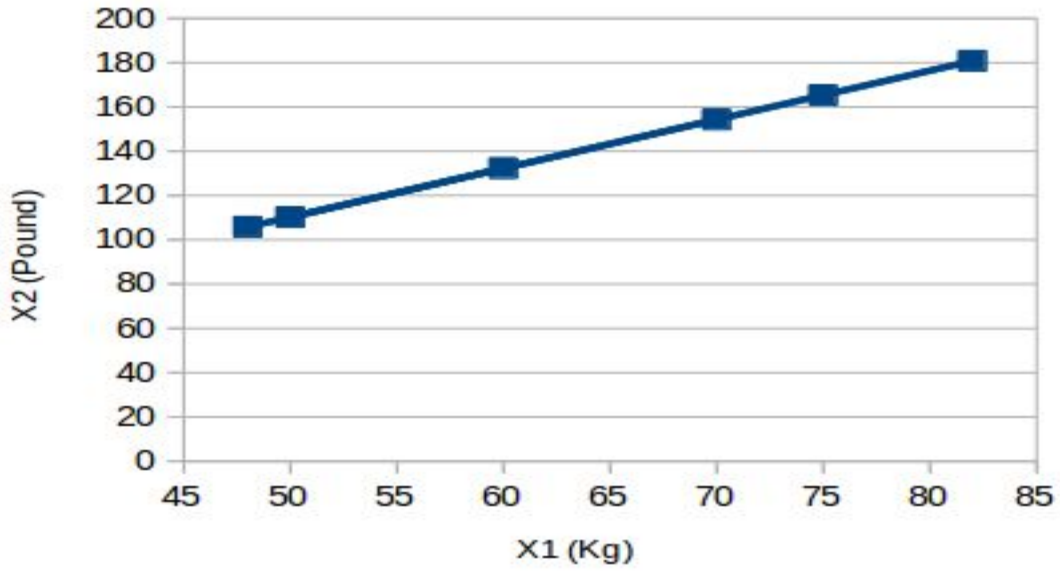
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- It can be divided into feature selection and feature extraction.



Dimensionality Reduction

INTERNSHIPSTUDIO

- Now consider a case in which we have, say 100 variables (p=100).
- Here we have weights of similar objects in Kg (X1) and Pound (X2). If we use both of these variables, they will convey similar information.



Dimensionality Reduction

INTERNSHIPSTUDIO

- So, it would make sense to use only one variable. We can convert the data from 2D (X1 and X2) to 1D (Y1) as shown below:
- Similarly, we can reduce p dimensions of the data into a subset of k dimensions (k<<n).

Components of DR

Feature selection:

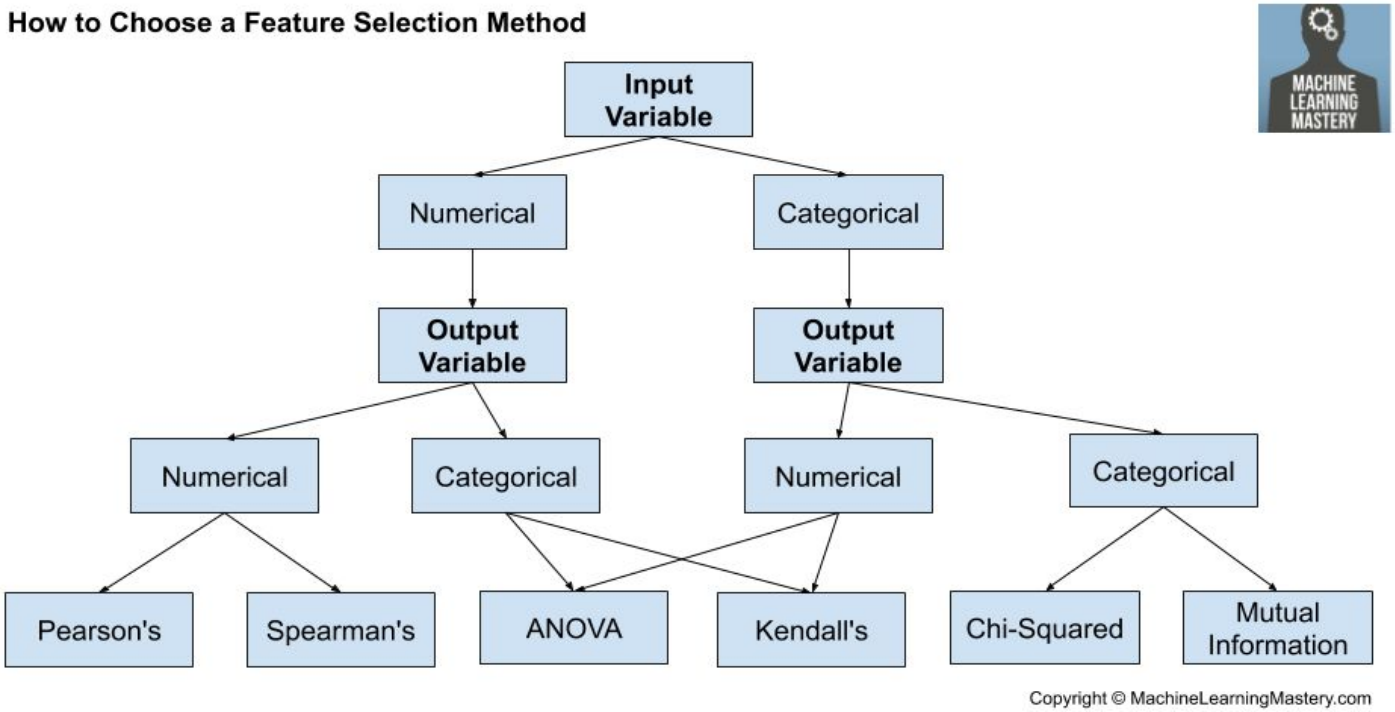
- In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  - Filter
  - Wrapper
  - Embedded

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.

Feature Selection

- **Feature selection** methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.
- Feature selection is primarily focused on removing non-informative or redundant predictors from the model.

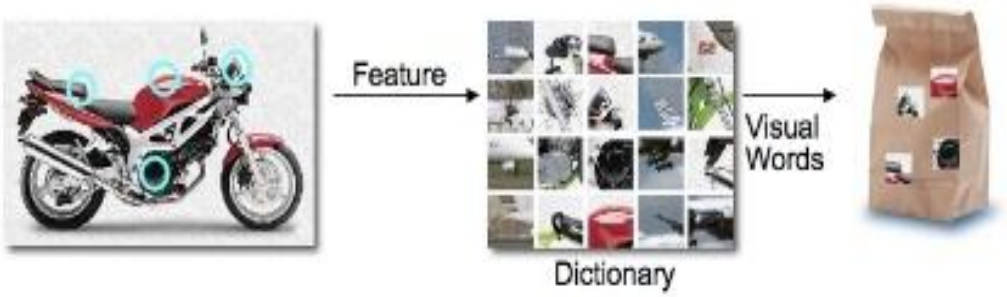


Feature Extraction

Feature extraction is a type of dimensionality reduction where a large number of pixels of the image are efficiently represented in such a way that interesting parts are captured effectively.

Feature Extraction (FE)

- Def: Feature Extraction (FE) is any algorithm that transformation raw data into features that can be used as an input for a learning algorithm.



- Examples
  - Construct bag-of-words vector from an email
  - Remove stopwords in a sentence
  - Apply PCA projection to high-dimensional data

Methods to Perform DR

Advantages of DR

The fewer features our training data has, the lesser assumptions our model makes and the simpler it will be. But that is not all and dimensionality reduction has a lot more advantages to offer, like..

- 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).

INTERNSHIPSTUDIO

- Q.1 What is Dimensionality Reduction?
- Q.2 What is the need of Dimensionality Reduction?
- Q.3 What is Feature Selection?
- Q.4 What is Feature Extraction?