Preparing Data for Modeling with scikit-learn

PREPARING NUMERIC DATA FOR MACHINE LEARNING

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Standardization and scaling

Robust scaling to mitigate effect of outliers

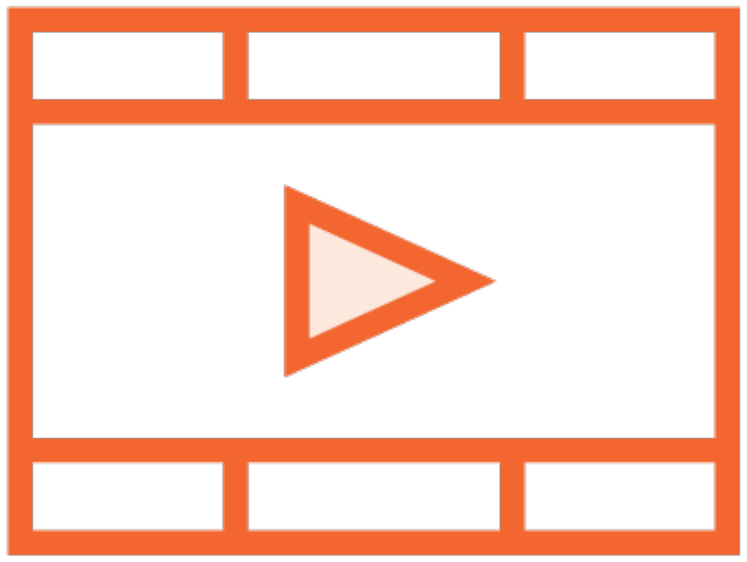
Normalization - L1, L2 and Max norm

Mapping to common distributions to fit models

Dimensionality reduction using factor analysis for pre-processing

Prerequisites and Course Outline

Prerequisites

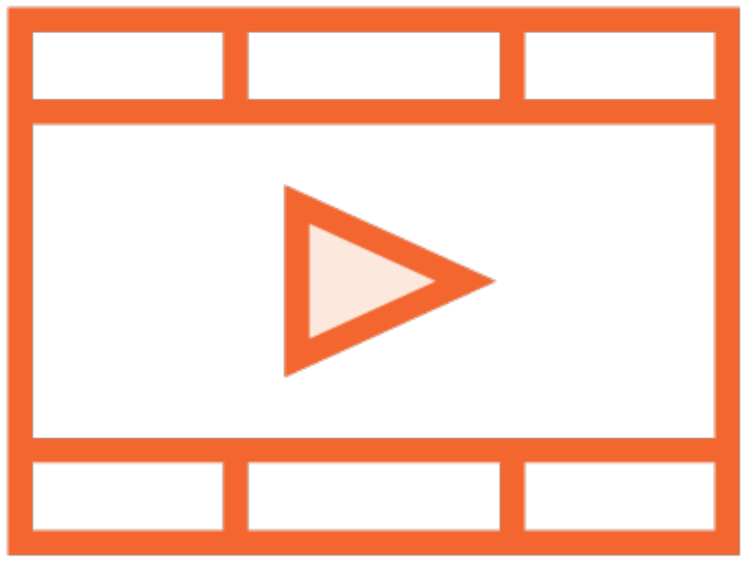


Comfortable programming in Python

Prior ML exposure recommended

High school math - mean, median, standard deviation

Prerequisite Courses



Building Your First scikit-learn Solution

Building Regression Models with scikit-learn

Building Classification Models with scikit-learn

Course Outline

Preparing numeric data

Novelty and outlier detection

Preparing text data

Preparing image data

Working with specialized datasets

Performing kernel approximations

Numeric Features in Training Data

Numeric Features

Can represent any kind of information

The range of each feature will be different

The average and dispersion of features will also be different

Comparing different features is hard

Machine learning algorithms typically do not work well with numeric data with **different scales**

Feature Scaling

Scaling

Standardization

Feature Scaling

Scaling

Standardization

Numeric values are **shifted and rescaled** so all features have the same scale i.e. within the same minimum and maximum values

Feature Scaling

Scaling

Standardization

Often data scaled to be in the range of 0 to 1, many people call this normalization

Feature Scaling

Scaling

Standardization

The feature range of data is something that you can specify

Feature Scaling

Scaling

Standardization

Does not bind values to a specific range

Feature Scaling

Scaling

Standardization

Centers data round the mean and divides each value by the variance so all features have 0 mean and unit variance

Data in One Dimension

Pop quiz: Your thoughtful, fact-based point-of-view on these numbers, please

Mean as Headline

The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Variation Is Important Too

"Do the numbers jump around?"

$$\text{Range} = X_{max} - X_{min}$$

The range ignores the mean, and is swayed by outliers - that's where variance comes in

Mean and Variance

Mean and variance succinctly summarize a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Variance and Standard Deviation

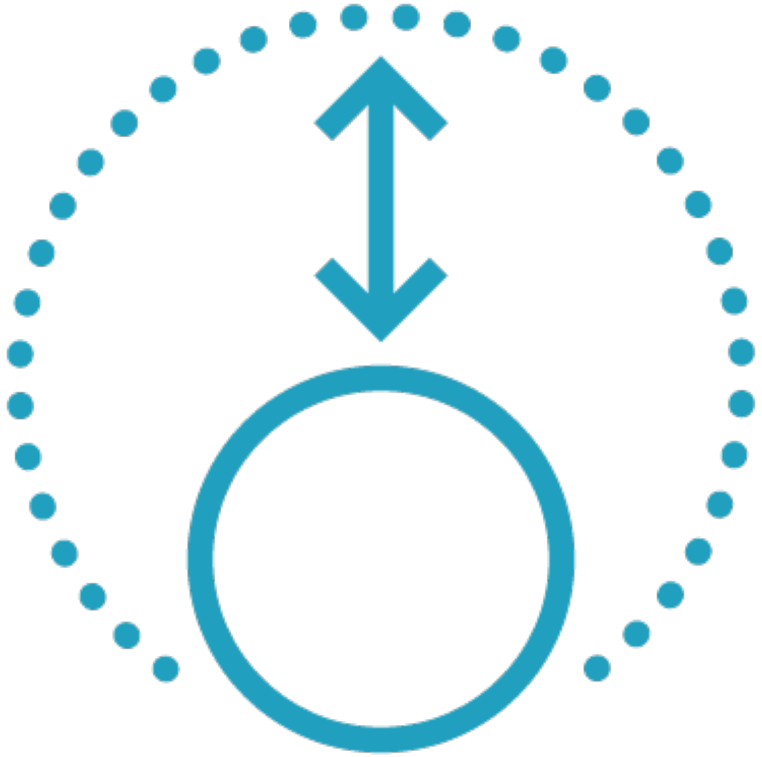
Standard deviation is the square root of variance

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

What Is Normalization?

What Is Normalization?

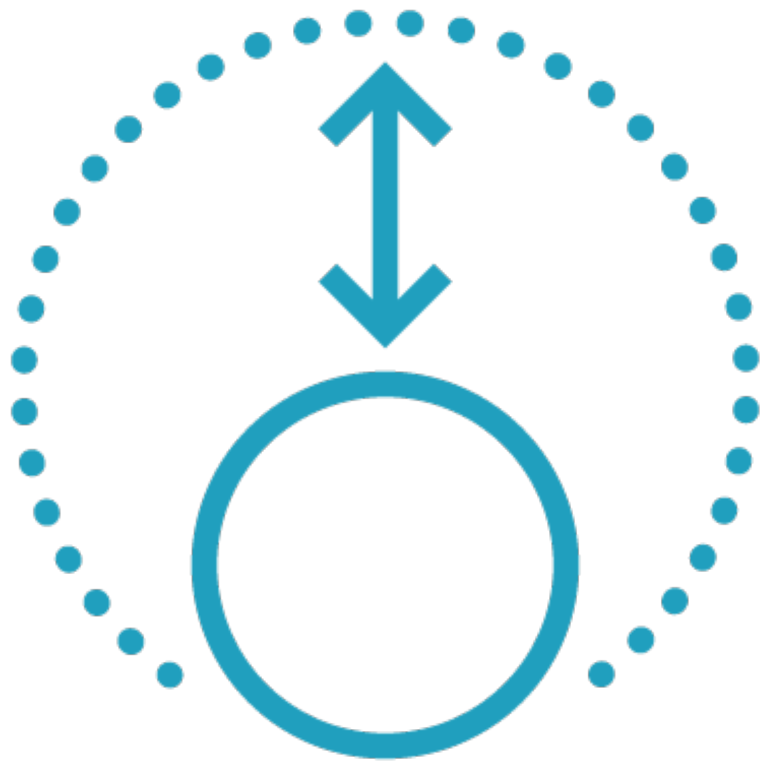


Scaling to a certain range - feature scaling

Centering at 0 and scaling to unit variance - standardization

Transforming a vector to unit norm

What Is Normalization?



Scaling to a certain range - feature scaling

Centering at 0 and scaling to unit variance - standardization

Transforming a vector to unit norm

Norm refers to the **magnitude** of the vector

Normalization

Process of scaling input vectors individually to unit norm (unit magnitude), often in order to simplify cosine similarity calculations.

Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors, widely used in ML algorithms - especially in document modeling applications.

Normalizing is a row-wise operation, while scaling is a column-wise operation

Data

$$\begin{bmatrix} x_{11} & x_{12} & & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ x_{31} & x_{32} & \cdots & x_{3k} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}$$

All of the numeric values in our dataset

Columns Represent Features

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ x_{31} & x_{32} & & x_{3k} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}$$

Standardization and scaling apply to an individual feature

Rows Represent Vectors

$$\begin{bmatrix} x_{11} & x_{12} & & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ x_{31} & x_{32} & \cdots & x_{3k} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}$$

Normalization applies to vectors i.e. to a row which represents data for a single instance

Different Norms

L1

Sum of absolute values of components of vector

L2

Traditional definition of vector magnitude

max

Largest absolute value of elements of vector

L1-norm

$$x_{new} = \frac{(x, y, z)}{|x| + |y| + |z|}$$

L2-norm

$$x_{new} = \frac{(x, y, z)}{\mathrm{sqrt}(x^2 + y^2 + z^2)}$$

Max norm

$$x_{new} = \frac{(x, y, z)}{max(abs(x, y, z))}$$

Transforming Distributions

PowerTransformer

Map features from any distribution to be as close to a Gaussian distribution as possible; useful when zero-mean, unit-variance normally distributed features are preferable.

Two Power Transforms

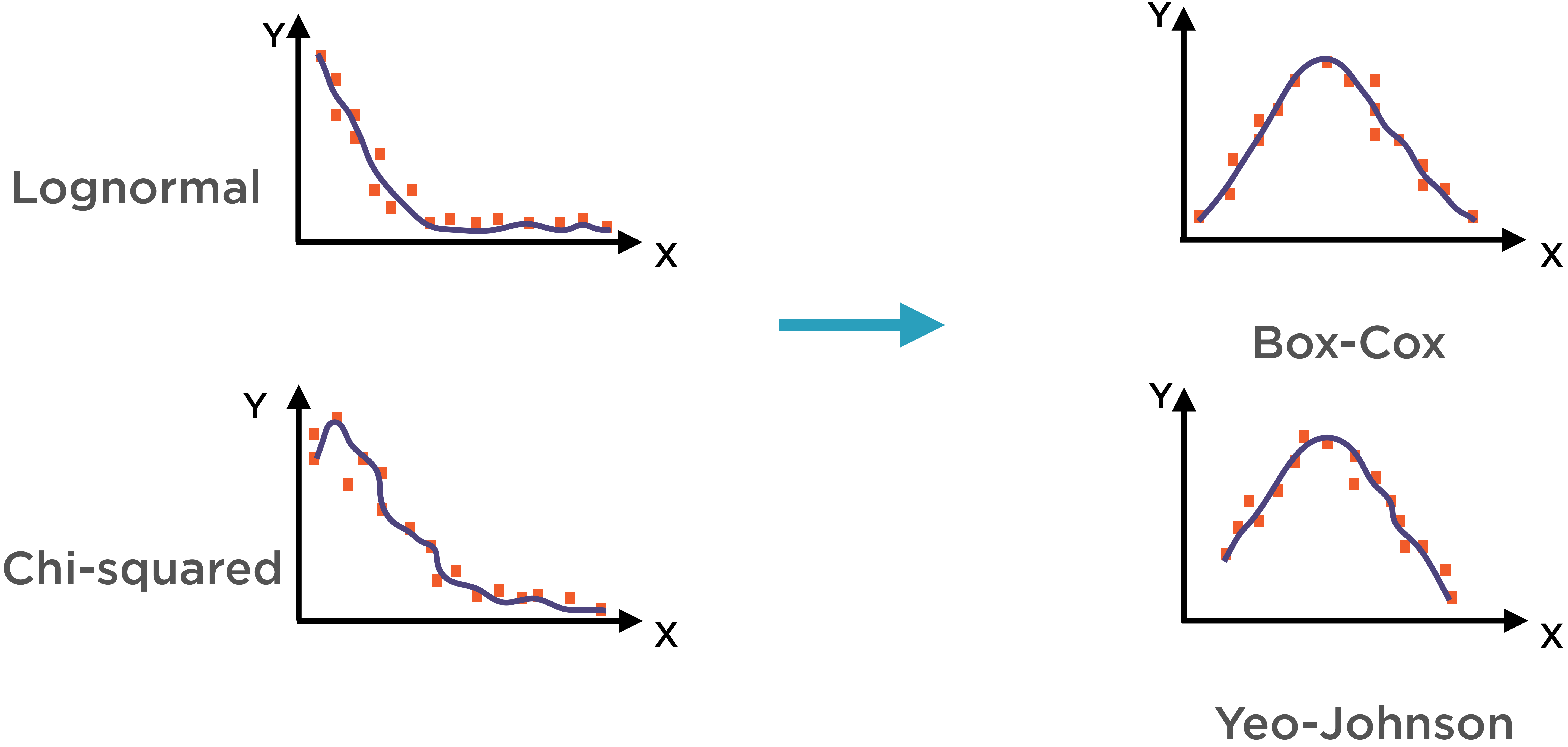
Box-Cox transform

Requires strictly positive input data

Yeo-Johnson transform

Supports positive and negative data

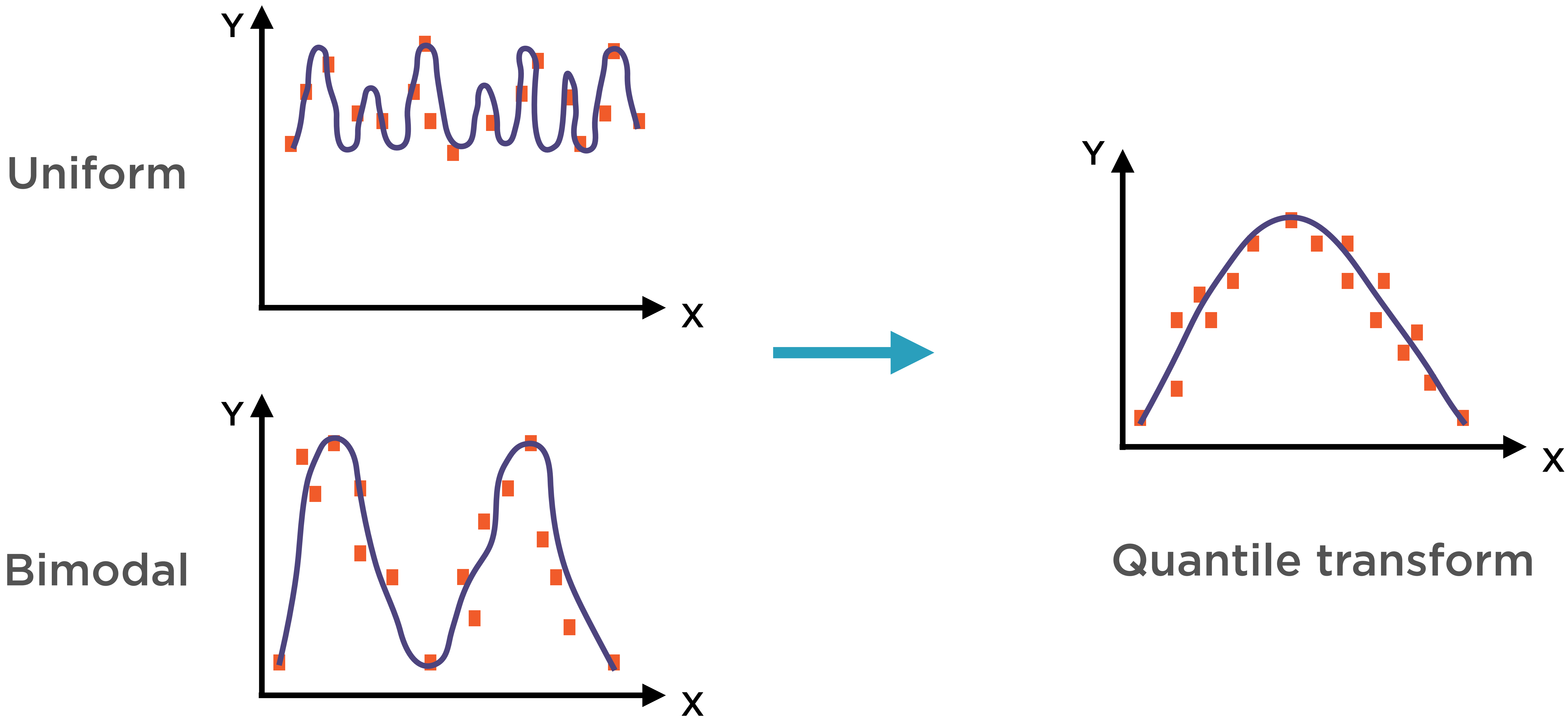
PowerTransformer



QuantileTransformer

Transforms features to follow a uniform or a normal distribution using quantile information; non-linear and might distort correlations and linear relationships.

QuantileTransformer



Demo

Calculating and visualizing summary statistics for numeric data

Demo

Using the standard scaler to standardize numeric features

Demo

Using the robust scaler to scale numeric features

Demo

Normalizing data using L1, L2 and Max norms

Demo

Transform data to a normal distribution using a quantile transformer

Demo

Perform dimensionality reduction on input features using Singular Value Decomposition (SVD)

Summary

Standardization and scaling

Robust scaling to mitigate effect of outliers

Normalization - L1, L2 and Max norm

Mapping to common distributions to fit models

Dimensionality reduction using factor analysis for pre-processing