Building Features Using Scaling and Transformations

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Quantization (binning) of counts and values

MinMax scaling, MaxAbs scaling

Generating polynomial features

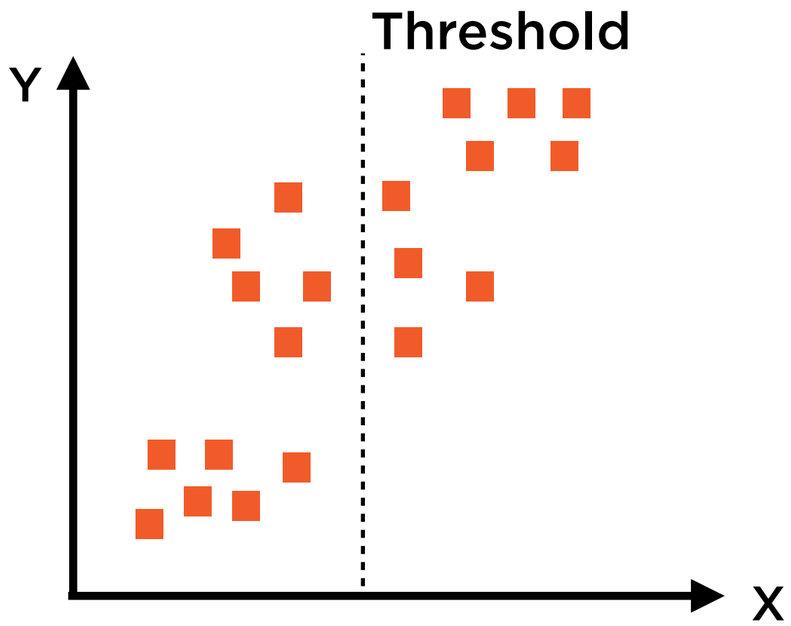
Power transformers and quantile transformers

Binarizer and KBinsDiscretizer

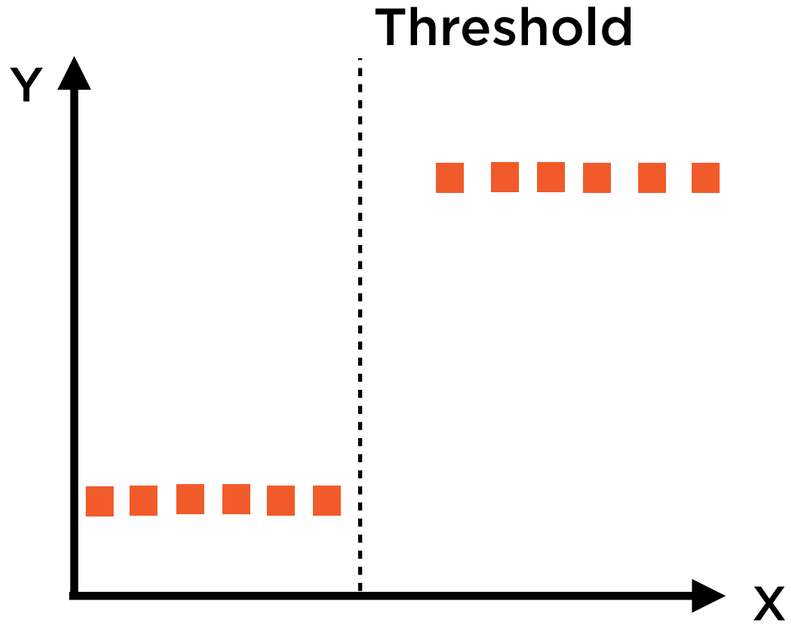
Binarizer

Converts continuous variable into a binary categorical variable based on a threshold specified by user

Binarizer



Continuous input

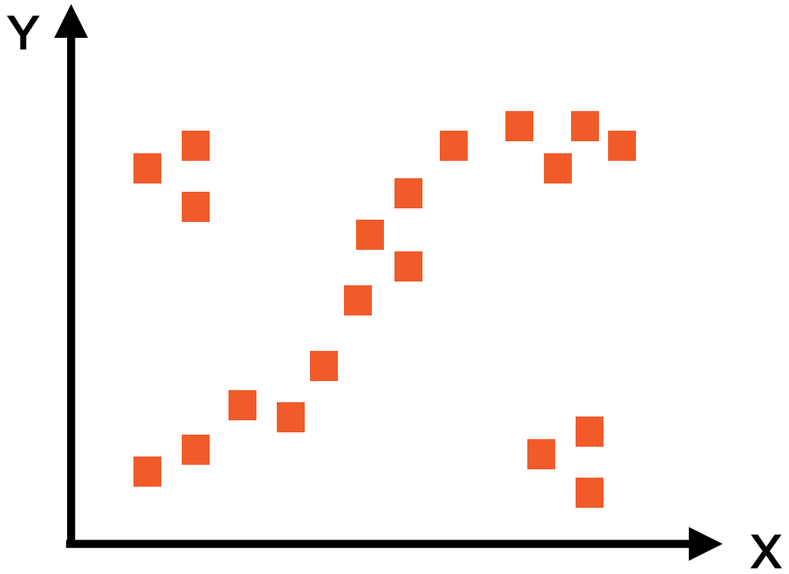


Binary categorical output

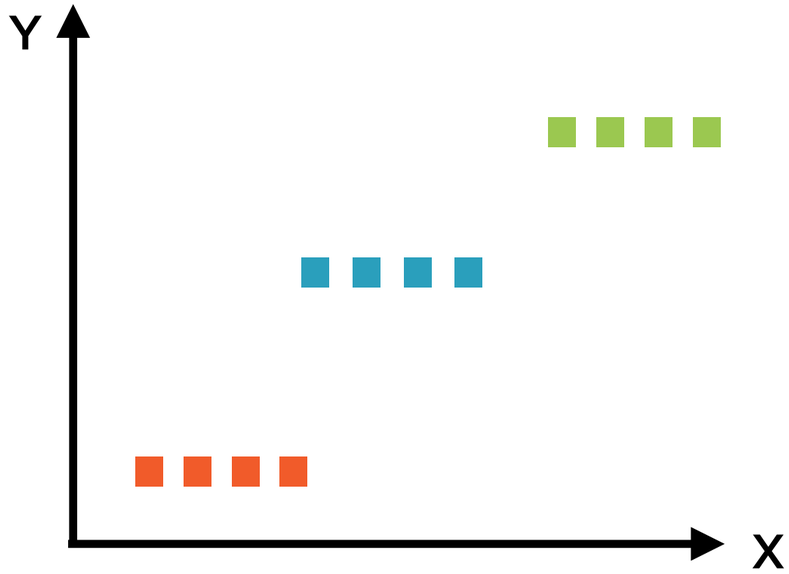
KBinsDiscretizer

Generalizes idea of binarizer; converts continuous data into categorical data arranged into a specified number of bins

KBinsDiscretizer



Before



After (3 bins)

KBinsDiscretizer Strategies

Uniform

Bin widths are constant in each feature

Quantile

All bins in each feature have approximately the same number of samples

K-means

Bins based on the centroids of a K-means clustering procedure

Demo

Converting continuous data to binary data using the binarizer

Demo

Using KBinsDiscretizer to create binned categories

Study outliers using category bins

MaxAbsScaler and MinMaxScaler

Max-Abs Scaler

$$z = \frac{x}{\max(\text{abs}(x))}$$

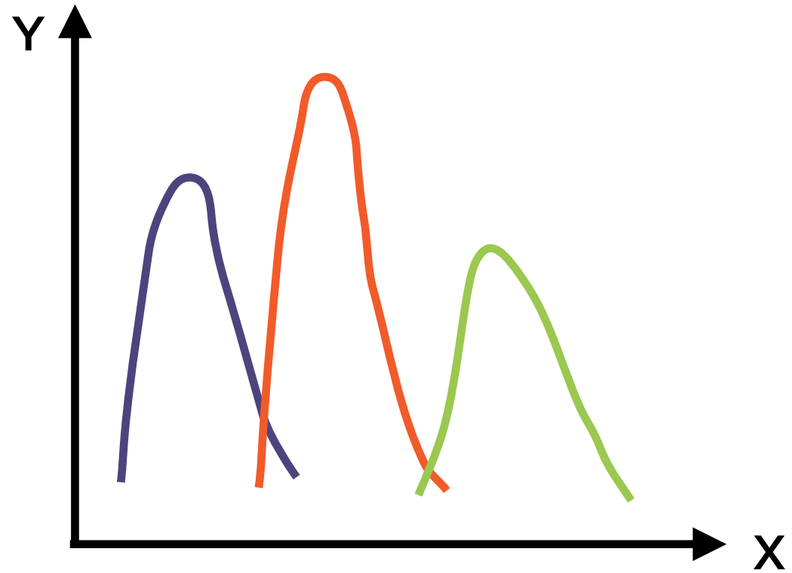
Scales features such that values lie in the range [-1, 1], the maximum absolute value will be 1

MinMaxScaler

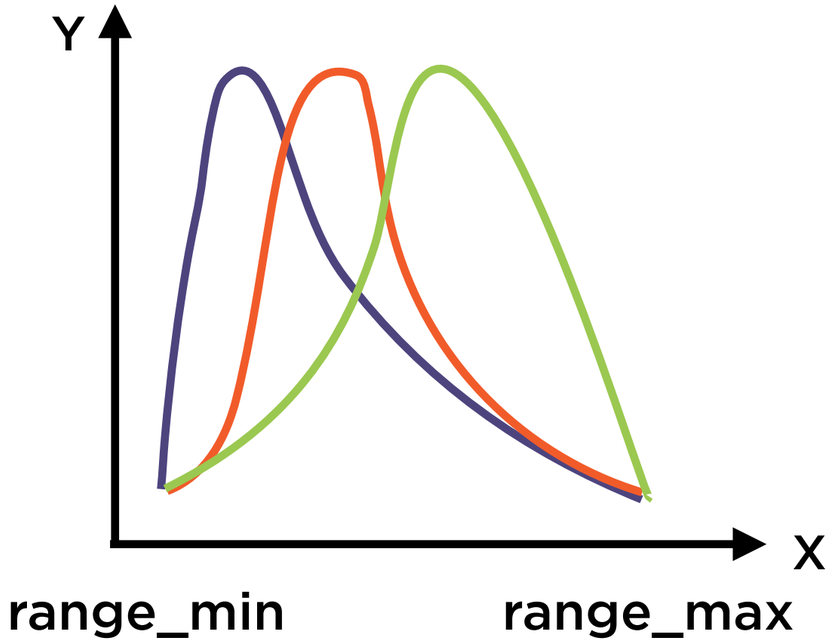
$$z = \frac{x_i - \min(x) * (\text{range\_max} - \text{range\_min})}{\max(x) - \min(x)} + \text{range\_min}$$

Scale the range of features to lie between a certain user specified range

MinMaxScaler



Before



After

MinMaxScaler is very sensitive to outliers

Demo

Using the MaxAbsScaler to scale data

Demo

Using the MinMaxScaler to scale data

FunctionTransformer

FunctionTransformer

Used to define a user-defined function or function object to be applied as a data transformation step, such as in a pipeline

Demo

Using FunctionTransformers to specify custom transformations

PolynomialFeatures

PolynomialFeatures

Utility class for generating all polynomial combinations of features up to a specified degree.

```
y = Wx + b
```

$$f(x) = Wx + b$$

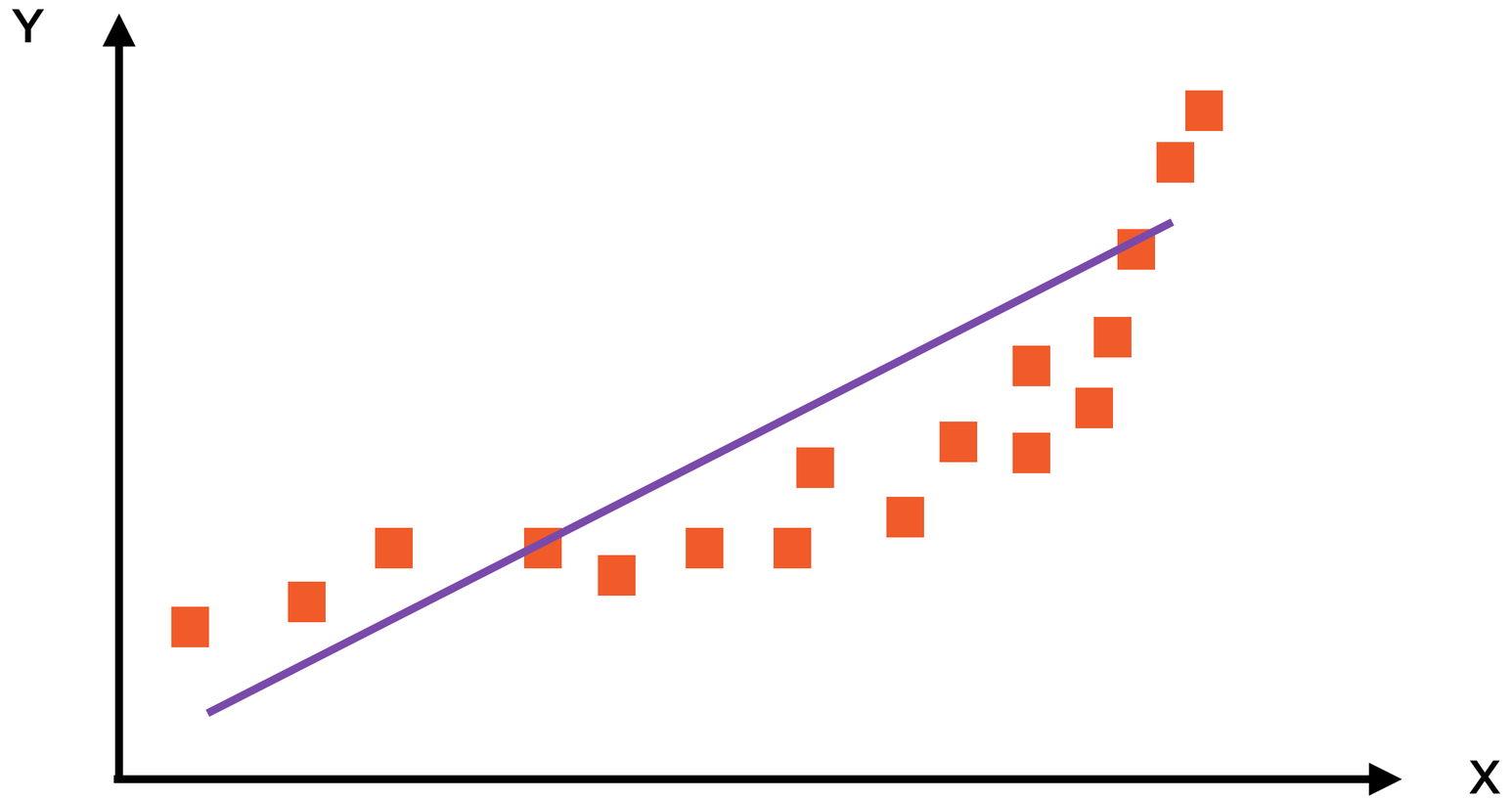
Relationship between y and x is a polynomial of degree 1

$$y = Vx^2 + Wx + b$$

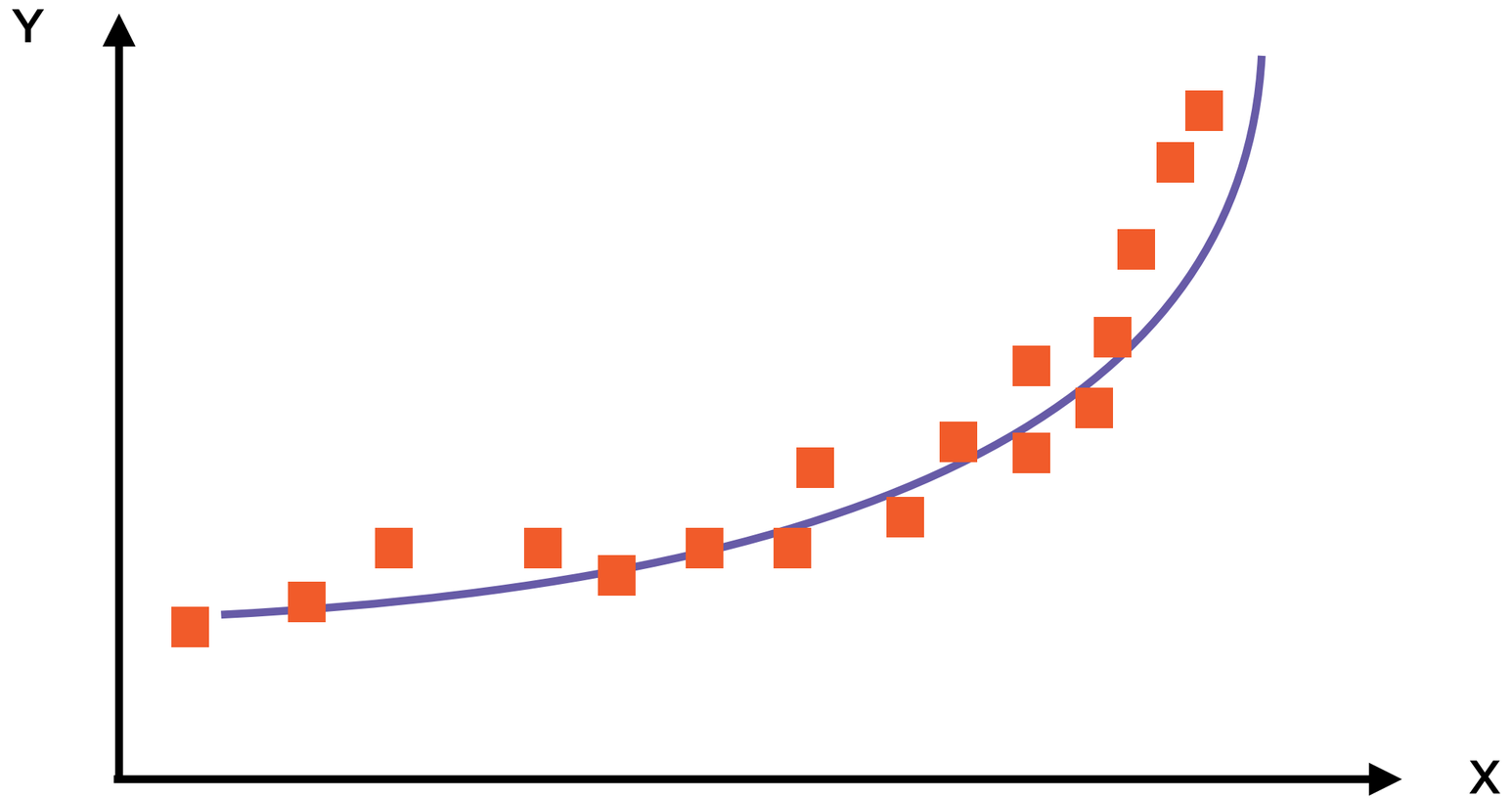
$$f(x) = Vx^2 + Wx + b$$

Now relationship between y and x is a polynomial of degree 2

Linear Fit Performs Poorly



Quadratic Fit Performs Well



Generate polynomials of a certain degree of all input features

Fit a simpler model on this polynomial data

Demo

Using PolynomialFeatures to generate polynomial inputs

Helps fit a simpler machine learning model

PowerTransformer

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

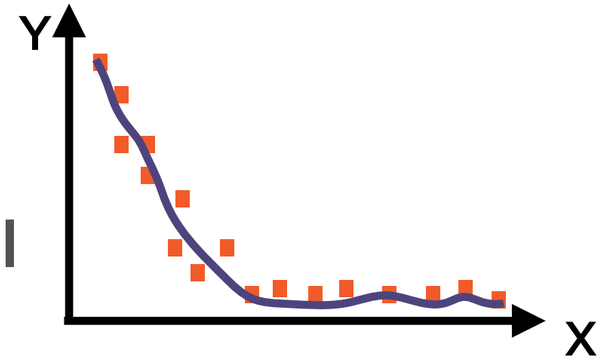
Requires strictly positive input data

Yeo-Johnson transform

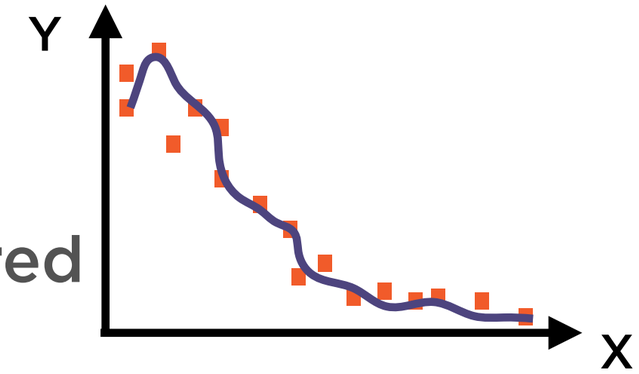
Supports both positive or negative data

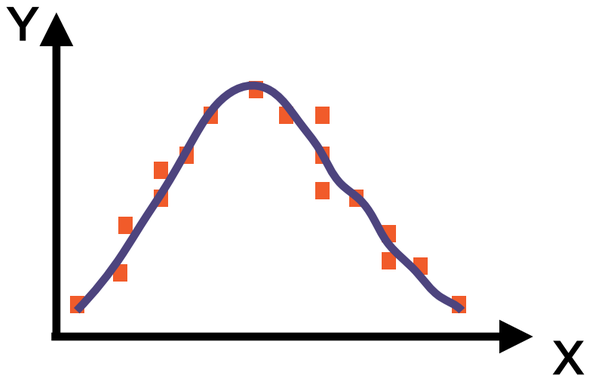
PowerTransformer

Lognormal

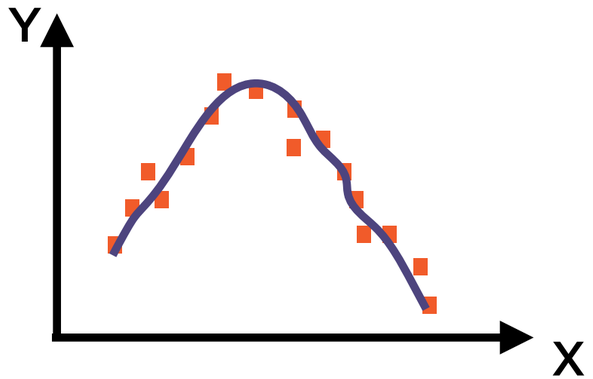


Chi-squared





Box-Cox



Yeo-Johnson

Demo

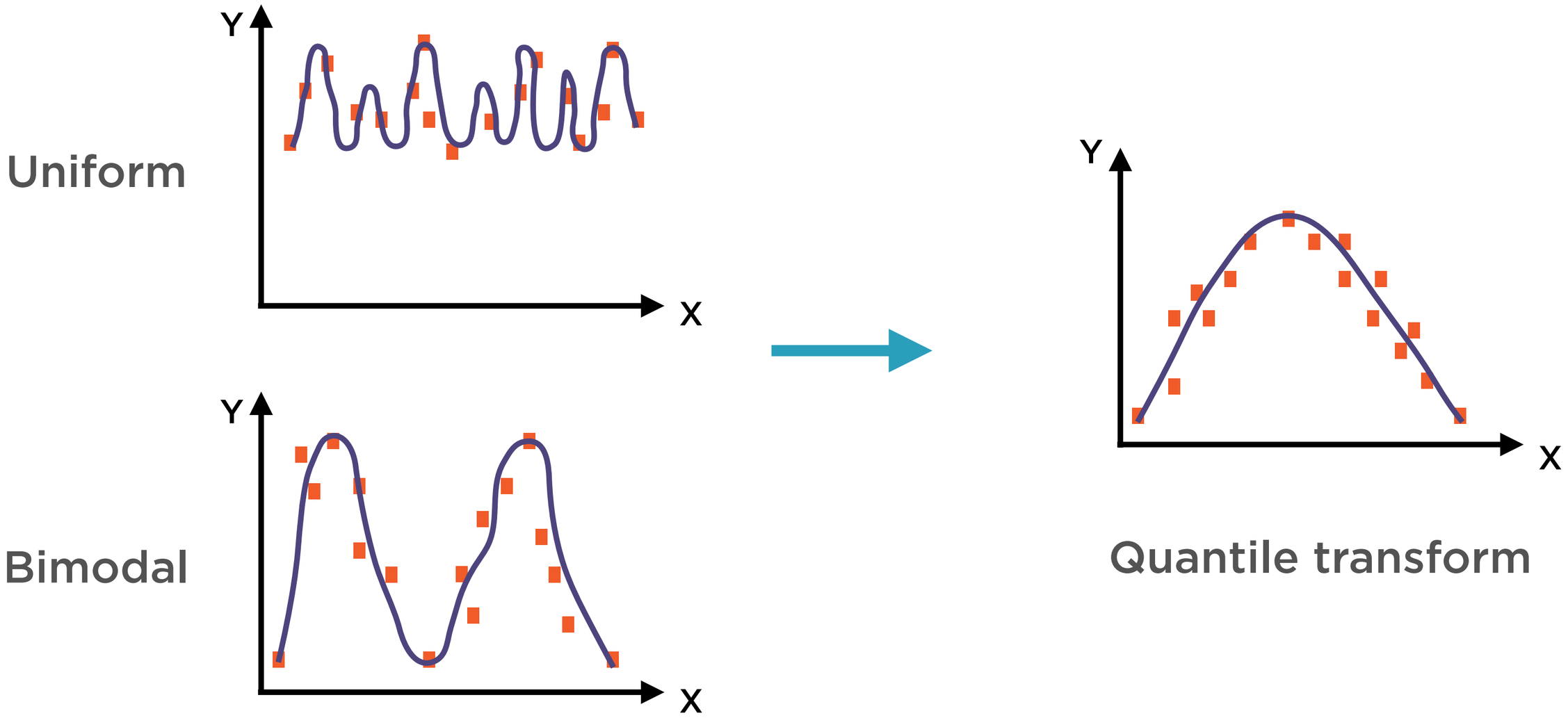
Using the PowerTransformer to transform data to normal form

QuantileTransformer

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

Using the QuantileTransformer to transform bimodal data to normal form

Summary

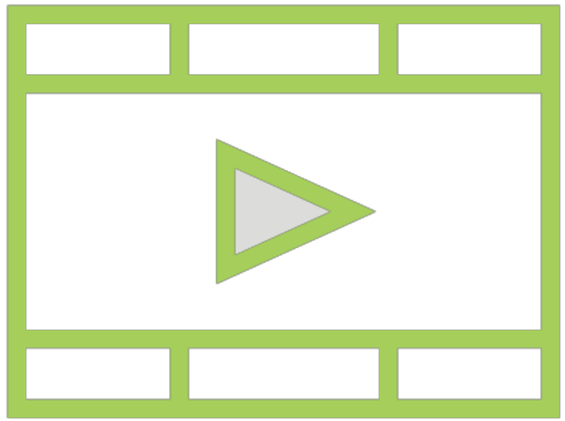
Quantization (binning) of counts and values

MinMax scaling, MaxAbs scaling

Generating polynomial features

Power transformers and quantile transformers

Related Courses



Building Features from Image Data

Building Features from Text Data

Reducing Complexity in Data