9.3. Wstępne przetwarzanie danych: dyskretyzacja i wybór atrybutów

iBigWorld:
Innovations for Big Data in a Real World

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jaka jest różnica między czyszczeniem a przekształcaniem danych?

Czyszczenie danych to proces usuwania danych, które nie należą do Twojego zbioru danych.

Przekształcenie danych to proces konwersji danych z jednego formatu lub struktury na inny.

Przekształcenie można również nazwać manipulacją danymi lub transformacją danych, transformacją i mapowaniem danych z jednej „surowej” formy danych do innego formatu do przechowywania i analizy

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dyskretyzacja

Dyskretyzacja przekształca zmienne ciągłe za pomocą dyskretne przedziały, podczas gdy normalizacja dostosowuje tylko rozkłady

- **Binarizer**
- **Bucketizer**
- **Discrete Cosine Transform**
- **Normalizer**
- **StandardScaler**
- **MinMaxScaler**
- **ElementwiseProduct**

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

MinMaxScaler

normalizuje każdą funkcję do określonego zakresu przy użyciu dwóch parametrów: dolnej granicy i górnej granicy

$$X^*_{\text{mm}} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Z-score standardization

bardzo powszechny w świecie analizy statystycznej, działa poprzez pobranie różnicy między wartością pola a średnią pola i skalowanie tej różnicy przez odchylenie standardowe wartości pola

$$\text{Z-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

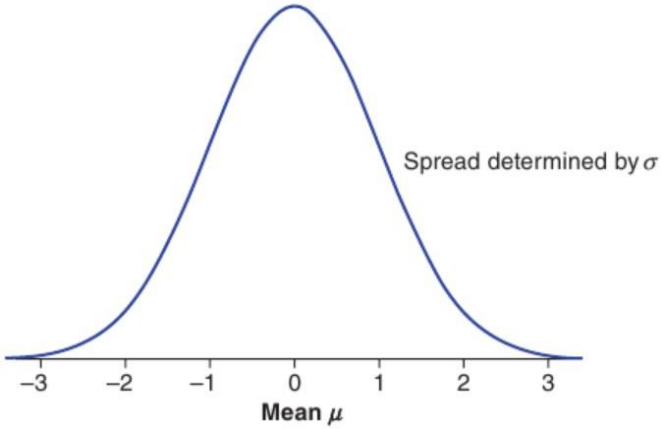
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

StandardScaler

normalizuje każdą funkcję tak, aby była zgodna z rozkładem normalnym.
Rozkład normalny to ciągły rozkład prawdopodobieństwa znany jako krzywa dzwonowa.

Jest symetryczne i wyśrodkowane na średniej, a jego rozrzut określa odchylenie standardowe.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

StandardScaler: ogólne przekształcenia

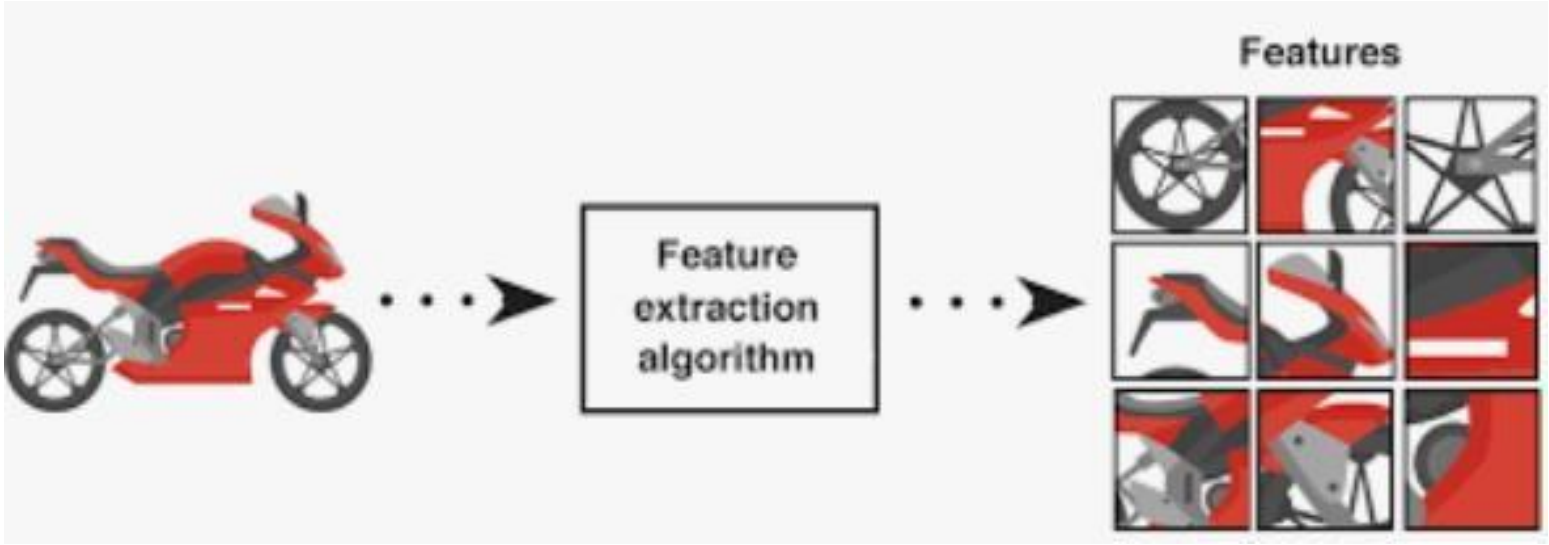
1. ln(data)
2. square root of data
3. inverse square root of data

Aby przetestować normalność, wykreślamy normalny wykres prawdopodobieństwa, który przedstawia kwantyle określonego rozkładu względem kwantylów standardowego rozkładu normalnego.

Co-funded by the Erasmus+ Programme of the European Union

Identyfikacja atrybutów (cech)

Metody wyodrębniania cech łączą oryginalny zestaw cech w celu uzyskania nowego zestawu mniej redundantnych zmiennych



Co-funded by the Erasmus+ Programme of the European Union

Metody doboru cech

Rozszerzenie wielomianowe rozszerza zbiór cech w przestrzeń wielomianową **VectorAssembler** łączy zestaw funkcji w jedną kolumnę wektorów

- **Single Value Decomposition (SVD)** jest metodą faktoryzacji macierzy, która przekształca macierz rzeczywistą/złożoną M (mxn) na macierz faktoryzowaną A .

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody doboru cech

- **Principal component analysis (PCA)** próbuje znaleźć taki zwrot, aby zbiór cech możliwie skorelowanych zamieniał się w zbiór cech nieskorelowanych liniowo.

Kolumny używane w tej transformacji ortogonalnej nazywane są składnikami głównymi.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dziękuję za uwagę!

Co-funded by the
Erasmus+ Programme
of the European Union