



University
of Bielsko-Biala



9.3. Wstępne przetwarzanie danych: dyskretyzacja i wybór atrybutów

iBigWorld:

Innovations for Big Data in a Real World



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

Dyskretyzacja

Dyskretyzacja

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

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

MinMaxScaler

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

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

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

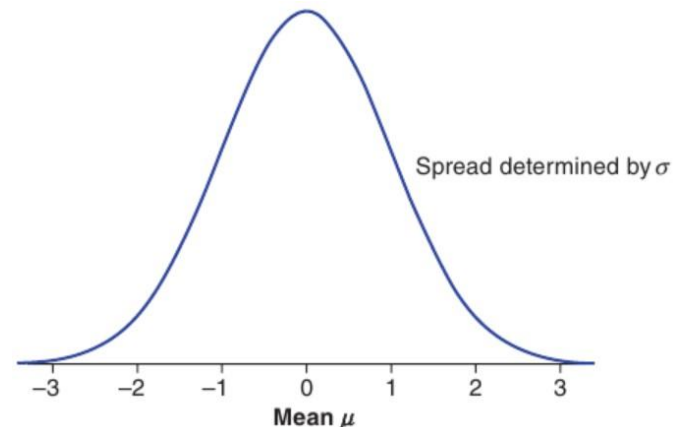
$$\text{Z-score} = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

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.



StandardScaler: ogólne

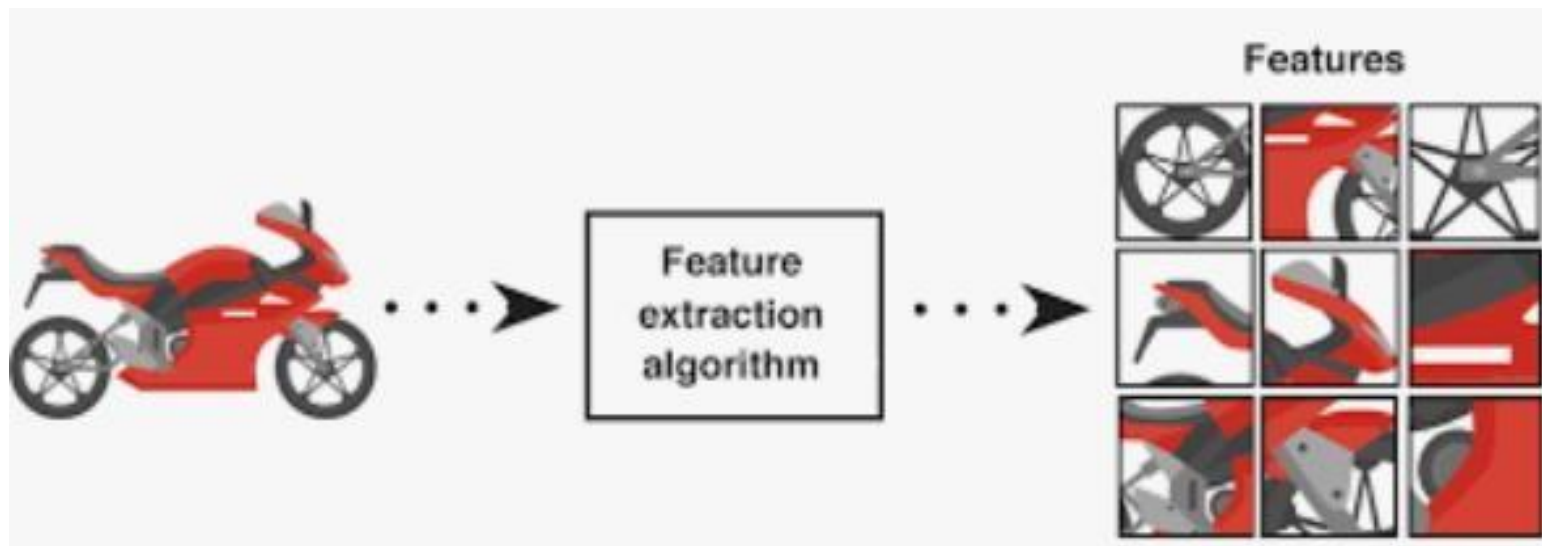
przekształcenia

1. $\ln(\text{data})$
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.

Identyfikacja atrybutów (cech)

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



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 ($m \times n$) na macierz faktoryzowaną A .

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.

Dziękuję za uwagę!

