DATA EXPLORATION

JENS BAETENS

WAT IS HET?

Wordt ook Exploratory Data Analysis of EDA genoemd

Beter begrip van de karakteristieken van de dataset

Waarom?

- Welk model is het best geschikt?
- Herkennen van patronen die niet door tools herkend worden.

EERSTE STAP – WAT ZIT ER IN DE DATASET

Hoeveel rijen (observations) en kolommen (features) zijn er?

Wat voor data zit er in elke kolom

- Categorieke data of numerieke?
- Discrete of Continue?

.info()
.shape

Let op categorieke data na ordinal encoding

TECHNIEKEN – UNIEKE WAARDEN

Het aantal verschillende waarden per kolom .nunique()

Kan gebruikt worden voor kolommen die een categorie bevatten

- Geeft het aantal elementen in elke categorie weer

Kan voorgesteld worden in een barplot

- 1 bar per kolom met categorieke data

TECHNIEKEN – FREQUENTIE

Geef weer hoe frequent een waarde voorkomt in een kolom

Kan gebruikt worden voor kolommen die een categorie bevatten

Kan voorgesteld worden in een barplot

- 1 plot per kolom
- 1 bar per unieke waarde

TECHNIEKEN – STATISTISCHE WAARDEN

Een aantal interessante waarden berekenen en vergelijken:

- Gemiddelden (E[X])
- Minimum / Maximum
- Variantie (Informatie over de spreiding)
  = $E[(X - E[X])^2]$

.describe()

- Mediaan / IQR beter als er veel outliers/extreme waarden zijn
  - Outlier als waarde kleiner is dan 25% kwartiel – 1.5 IQR
  - Extreem als waarde kleiner is dan 25% kwartiel – 3.5 IQR

75% + 1,5 IQR
75% + 3,5 IQR

Toepasbaar op numerieke kolommen

TECHNIEKEN – HISTOGRAM

↳ hoe is de verdeling ↗ uniform ↗ normaal ↘ ...

Geeft informatie over in welk bereik de meeste waarden vallen.

Aantal bins heeft een grote impact

Een histogram per kolom met numerieke data

feature

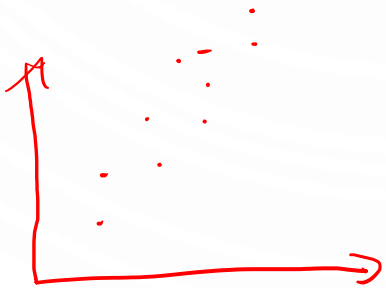
ZOEKEN NAAR VERBANDEN TUSSEN FEATURES

Met behulp van een scatterplot zoeken naar features die met elkaar verband kunnen houden.

Voor numerieke waarden

Alle combinaties afzoeken kan veel werk zijn

.pairplot() (later)



≈ lineair Verband

TECHNIEKEN – CORRELATION HEAT MAP

verband

Geeft de samenhang tussen twee elementen weer

Kans als A hoog is dat dan ook B hoog is: Positieve Correlatie

Kans als A hoog is dat B dan laag is: Negatieve Correlatie

Wordt berekend als: $\frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$

anders

Heatmap met zowel op X als Y as de numerieke kolommen

Bekijk de correlatie van 2 kolommen in meer detail met een scatterplot

.corr() → matrix te berekenen | .matshow()

TECHNIEKEN – PEARSON CORRELATION AND TREND

Plot een aantal interessante combinaties uit de heatmap als scatterplot.

Bijvoorbeeld de combinaties met een sterke negatieve of positieve correlatie

TECHNIEKEN – CRAMER-V CORRELATION

Correlation heat map voor kolommen met categorieke data.

Cramer's V correlatie = sqrt(phi / min(r-1, k-1))

Waar phi = $\sum_{i,j} \frac{\Pr[A=i,B=j]^2}{\Pr[A=i]*\Pr[B=j]} - 1$

niet nodig om te kennen.

De correlatie is

- 0 als de kolommen onafhankelijk zijn
- 1 als de kolommen volledig samenhangen

Correlatie kan in meer detail bekeken worden met een bubble plot

- size bubble is het aantal keer het voorkomt

↳ scatter plot + size dots = correlatie

TECHNIEKEN – IMPORTANT FEATURES

grootste correlatie

Important features zijn de features die een grote impact hebben op de gewenste feature. (label / target)

Kan uit de correlation heatmaps gehaald worden

Getoond als een bar-plot met op de x-as de kolommen en op de y-as de correlatie coëfficient

TECHNIEKEN – OUTLIER DETECTION

→ iets dat zeldzaam voorkomt

Wordt ook anomaly detection genoemd

Outliers komen overeen met zeldzame gevallen (positief of negatief)

Kan gedaan worden door

- standard deviation analysis (IQR)
- Isolation forest (Machine learning techniek)

) ML-technieken

Bubble shart met op de x-as alle numerieke kolommen

↳ outliers onder elkaar

TECHNIEKEN – OUTLIER ANALYSIS

Meer gedetaileerde overzicht van outliers en statistische waarden

Enkele kolom

- Box plot

Meerdere kolommen

- Scatter plot en outliers in aparte kleur
- Outliers moeten eerst gedetecteerd worden (op basis van statistische gegevens of ML-technieken zoals Standard Deviation Analysis of Isolation Forest

TECHNIEKEN – PARETO ANALYSIS

≈ outlier detection

Om te onderzoeken welke data belangrijk kan zijn.

Pareto 80-20 vaak gebruikt:

- De waarden kleiner dan 20% van het maximum zijn klein
- De waarden groter dan 80% van het maximum zijn groot

Afhankelijk van je vraag kan 1 of beide groepen genegeerd worden.