Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Dekomposition Grundlagen

CAS Practical Machine Learning

▸ Violeta Vogel, TI BFH

Agenda

- Was ist Dekomposition?
- Principal components analyses, PCA)
- Einsatz

Was ist Dekomposition?

- Dekomposition bezeichnet allgemein die Zerlegung eines komplexen Ganzen in seine einzelnen Bestandteile, Segmente oder Teilprobleme

Was ist Dekomposition?

- Informatik & Computational Thinking:
  - Hier ist die Dekomposition eine Kernmethode,
  - um ein grosses Problem in kleinere, handhabbare Teilaufgaben zu zerlegen.
  - In der Softwareentwicklung (z. B. bei der funktionalen Dekomposition) werden komplexe Systeme in einzelne Funktionen unterteilt, um die Entwicklung und Wartung zu erleichtern.

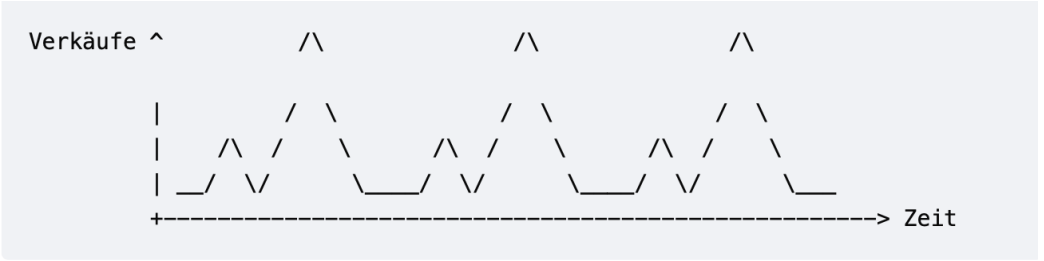
- Wirtschaft & Mathematik: Hier versteht man darunter die Analyse von Daten oder Prozessen durch Aufteilung in einzelne Komponenten (z. B. die Zeitreihenanalyse in Trend, Saison und Restrauschen).

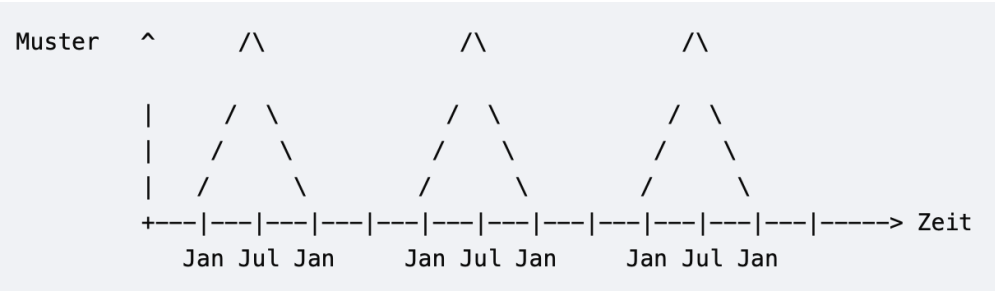
Dekomposition in den Daten

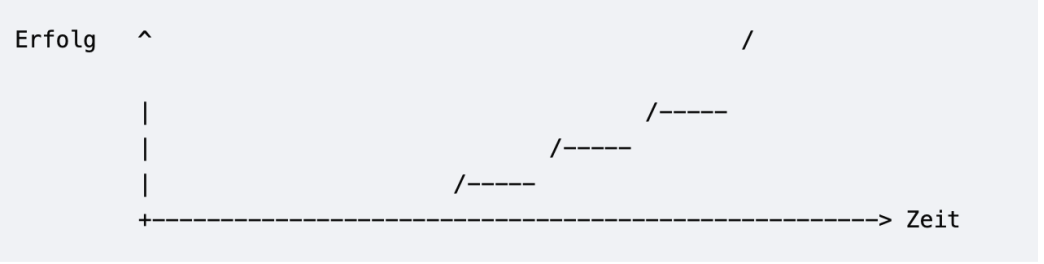
- In der Datenanalyse hilft die Dekomposition,
  - Saisonbereinigung versus Trend
  - Prognose (Forecasting): Aus Trend und Saisonalität kann man die Zukunft mathematisch berechnen.
  - Fehleranalyse: Ist das "Restrauschen" normal oder hat es plötzlich einen riesigen Ausschlag?

Dekomposition in Beispiel

- Eisverkauf über 3 Jahre:
- Saison
- Trend
- Zufall







Merkmals-Dekomposition

- Standardverfahren zum Diemensionsreduktion (Feature Reduktion):
  - Viele Features werden zu wenigen, aussagenkräftige Hauptfaktoren dekomponiert.
- Die wichtigsten Algorithmen:
  - PCA (Principlan Component Analysis): Er sucht nach den Richtungen (Achsen) in den Daten, die die größte Varianz (Informationsgehalt) aufweisen.

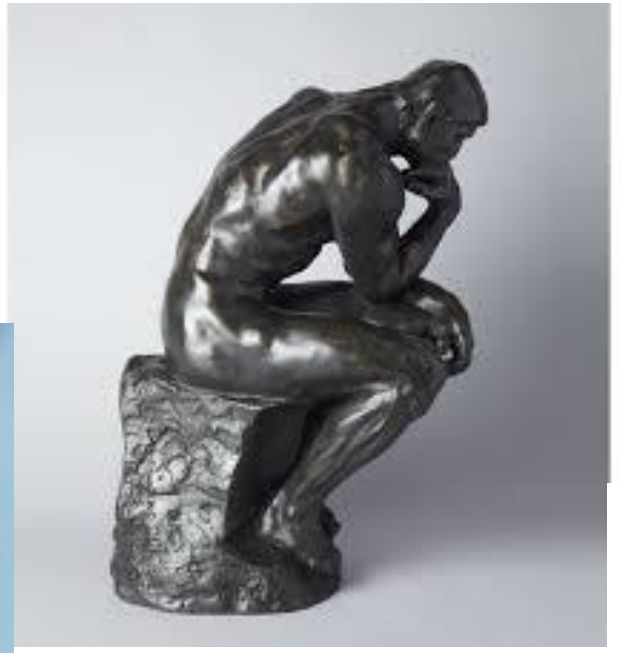
- t-SNE (t-distributed Stochastic Neighbor Embedding): Ein moderner Algorithmus, der besonders gut darin ist, nicht-lineare Strukturen in 2D oder 3D darzustellen.

  - Faktorenanalyse: geht davon aus, dass es verborgene (latente) Variablen gibt, die die beobachteten Daten beeinflussen

Merkmals-Dekomposition

Ziel	Algorthmus
Daten für KI komprimieren	PCA
Eine schöne Graphik mit Gruppen zeigen	T-SNE
Die Bedeutung der Daten verstehen	Faktorenanalyse

Merkmals-Dekomposition: PCA



Merkmals-Dekomposition: PCA

- Die PCA (Principal Component Analysis) oder Hauptkomponentenanalyse ist ein Verfahren:
  - um einen riesigen Datensatz mit vielen Informationen (Variablen) so zu vereinfachen,
  - dass nur die wichtigsten Merkmale übrig bleiben,
  - ohne dabei das "Gesamtbild" zu verlieren

Merkmals-Dekomposition: PSC

- Zentrierung: Man schiebt die Datenwolke so, dass ihr Mittelpunkt genau bei Null liegt.
- Berechnung der Hauptkomponenten:
  - Die 1. Hauptkomponente (PC1) ist die Linie, die so mitten durch die Daten gelegt wird, dass die Punkte so weit wie möglich darauf verteilt sind. Hier steckt die meiste Information drin.
  - Die 2. Hauptkomponente (PC2) steht im rechten Winkel (90◦) zur ersten und fängt den Rest der Information ein.
- Projektion:
  - Man bildet die Datenpunkte auf diese neuen Linien ab.
  - Jetzt gibt es statt 100 Variablen vielleicht nur noch 2 oder 3 Hauptkomponenten, die aber 90% der Information enthalten.

Merkmals-Dekomposition: PCA

- Visualisierung:
  - Daten mit 10 oder 50 Merkmalen kann man nicht zeichnen.
  - Nach der PCA hat man 2 Hauptkomponenten und
  - man kannt sie in einem einfachen Koordinatensystem als Punkte darstellen.

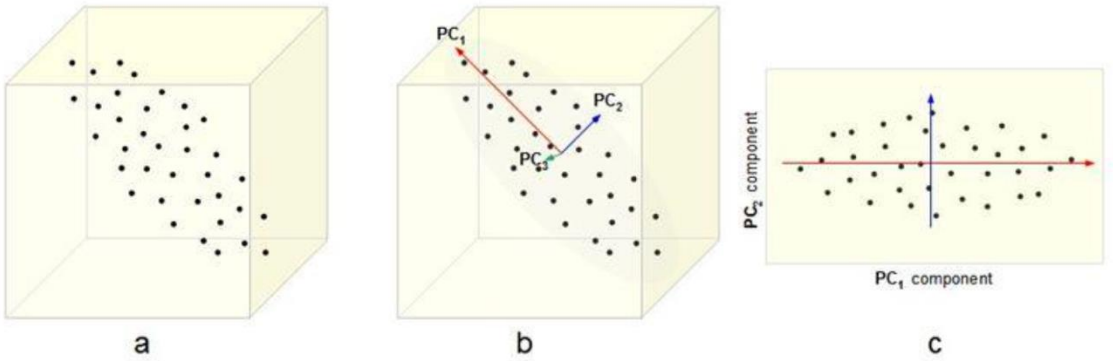
- Geschwindigkeit: KI-Modelle lernen viel schneller, wenn sie statt 1.000 nur noch 10 komprimierte Merkmale verarbeiten müssen.

- Rauschunterdrückung: Kleine, unwichtige Schwankungen (Rauschen) landen meist in den hinteren Hauptkomponenten und werden einfach weggeworfen.

Merkmals-Dekomposition: PCA

Visualisierung des oben genannten Prinzips:

- a: Ausgangslage: dreidimensionales Dataset in perspektivischer Darstellung
- b: erste Hauptkomponente (PC1, roter Pfeil): maximale Ausdehnung des Punkteschwarms
- zweite Hauptkomponente (PC2, blauer Pfeil): maximale Ausdehnung senkrecht auf PC1
- dritte Hauptkomponente (PC3): Rest, senkrecht auf PC1 und PC2
- c: Rotation: PC1 wird zur neuen x-Achse, PC2 zur neuen y-Achse
- im Hintergrund wirken Matrixoperationen, welche geschlossene Lösungen darstellen und recht schnell in der Ausführung sind



Merkmals-Dekomposition: PCA

- zur Anwendung von PCA mit Python wird ebenfalls eine Methode aus scikit-learn eingesetzt: `sklearn.decomposition.PCA`
- da es sich dabei um eine Machine Learning Methode handelt, müssen die Daten entsprechend aufbereitet sein, d.h.
  - keine Missing Values
  - nur numerische Daten
  - Bereinigung anderer möglicher Anomalien
- ausserdem wird das Target "y" vom Rest der Daten abgetrennt: "features - target - split"
  - Features $\rightarrow$ "X"
  - Target $\rightarrow$ "y"
- und die Features müssen standardisiert werden

Merkmals-Dekomposition: PCA

Vorgehen mit `sklearn.decomposition.PCA`

- der Umgang mit Trainerklassen erfolgt immer nach demselben Muster
  1. importieren der Klasse aus sklern
  2. instanziieren eines Trainer-Objektes
  3. anwenden des Trainer-Objektes auf die Daten

```
from sklearn.decomposition import PCA ## import trainer class
model = PCA()                         ## instantiate trainer object
pred = model.fit_transform(X)         ## train and apply trainer on data
print(pred[:3, :])                    ## check prediction (result, optional)
```

```
[[ 0.82454003 -1.45559011 -1.0976675   0.91524634  0.08531455 -0.15409804
   0.04340854 -0.21781905  0.01493519 -0.01620926]
 [-0.7627286  -0.72786538 -0.32087404 -0.13931963 -0.77314314  0.13132969
   0.59687245 -0.0447055   0.04996993  0.14532455]
 :
```

Merkmals-Dekomposition: PCA

- das Ergebnis ("pred") ist eine Matrix (numpy.ndarray) mit denselben Dimensionen wie die Feature-Matrix und enthält die Koordinaten (Hauptkomponenten) der rotierten Daten
- übliche Visualisierung
  - Scatterplot von "PC 1" vs "PC 2"
  - optional: einfärben nach einem kategorialen Merkmal (hier "y")
- eine vertieftere Analyse der erkennbaren Muster ist allerdings eine Tätigkeit der Datenanalyse und unterbleibt an dieser Stelle