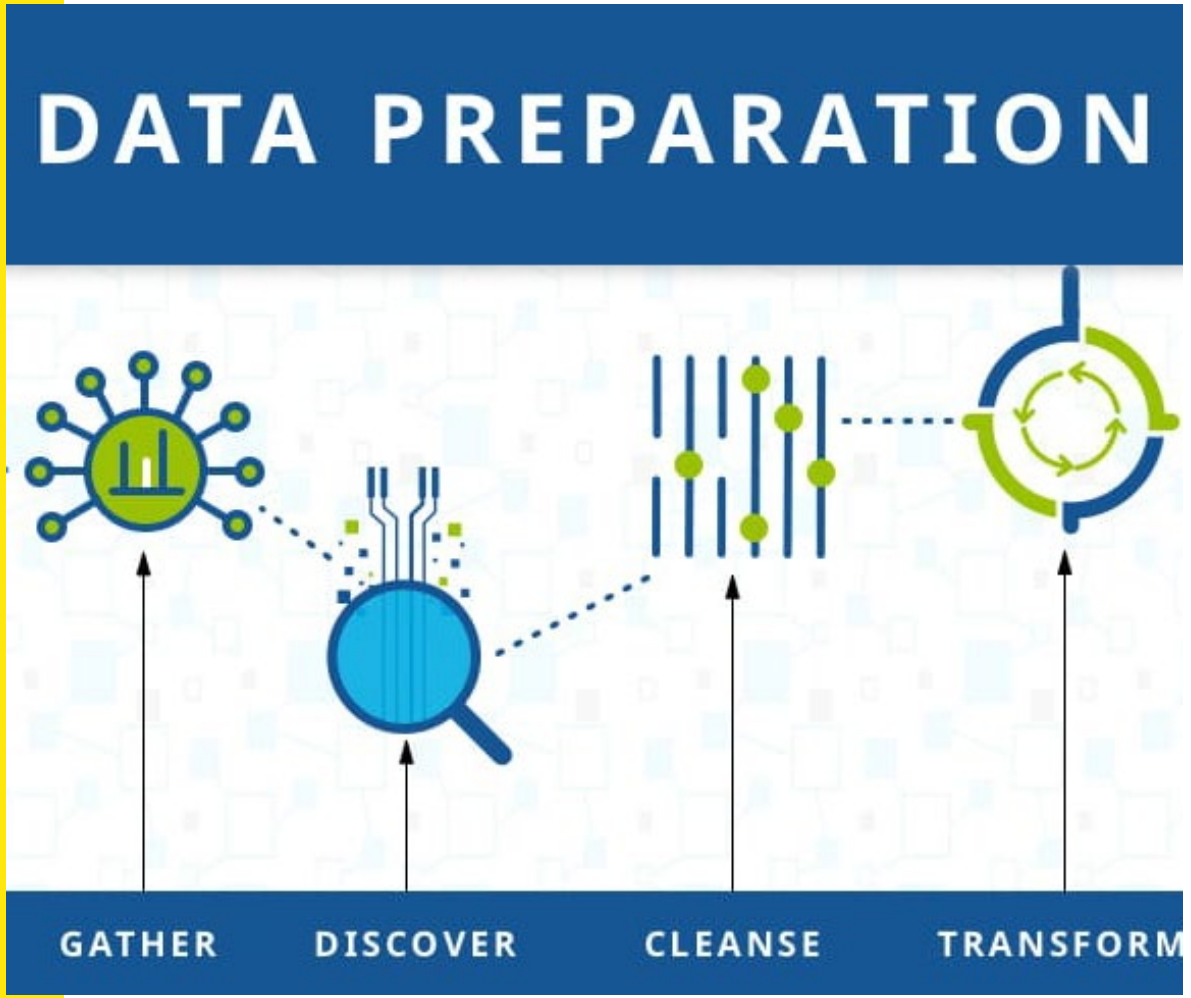
Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Maschinelles Lernen - Datenaufbereitung



– Hans Friedrich Witschel, Andreas Martin

Datenaufbereitung – unser «Running Example»

- Die Firma TeleFlow bietet 1-Jahres-Mobilfunkverträge
- Aufgabe: vorhersagen, welche Kunden ihren Vertrag *nicht* erneuern werden

contractId	Service
7590-VHVEG	OnlineBackup
5575-GNVDE	DeviceProtection
5575-GNVDE	OnlineSecurity
3668-QPYBK	OnlineBackup
3668-QPYBK	OnlineSecurity
7795-CFOCW	TechSupport
7795-CFOCW	DeviceProtection
7795-CFOCW	OnlineSecurity
9305-CDSKC	MultipleLines
9305-CDSKC	DeviceProtection
9305-CDSKC	StreamingMovies
9305-CDSKC	StreamingTV
1452-KIOVK	MultipleLines
1452-KIOVK	OnlineBackup

services.csv

contractId	ticketDate
7795-CFOCW	18.01.2022
7795-CFOCW	09.08.2023
7795-CFOCW	25.01.2022
7892-POOKP	17.02.2022
7892-POOKP	28.05.2022
0280-XJGEX	13.03.2023
0280-XJGEX	21.03.2023
0280-XJGEX	10.03.2023
0280-XJGEX	24.03.2023
9959-WOFKT	08.05.2021
9959-WOFKT	19.01.2020
9959-WOFKT	24.08.2022
9959-WOFKT	11.09.2022
6322-HRPFA	19.07.2020

tickets.csv

contractID	gender	SeniorCitize	Partner	Dependents	StartDate	PhoneServic	InternetServ	PaperlessBil	PaymentMe	MonthlyCha	CancellationReceived
7590-VHVEG	Female	0	Yes	No	31.01.2024	No	DSL	Yes	Electronic ch	29.85	
5575-GNVDE	Male	0	No	No	16.05.2021	Yes	DSL	No	Mailed check	56.95	
3668-QPYBK	Male	0	No	No	24.11.2023	Yes	DSL	Yes	Mailed check	53.85	23.01.2024
7795-CFOCW	Male	0	No	No	20.06.2020	No	DSL	No	Bank transfe	42.3	
9237-HQITU	Female	0	No	No	22.12.2022	Yes	Fiber optic	Yes	Electronic ch	70.7	20.02.2023
9305-CDSKC	Female	0	No	No	10.12.2021	Yes	Fiber optic	Yes	Electronic ch	99.65	07.08.2022
1452-KIOVK	Male	0	No	Yes	11.05.2022	Yes	Fiber optic	Yes	Credit card (	89.1	
6713-OKOM	Female	0	No	No	06.05.2023	No	DSL	No	Mailed check	29.75	
7892-POOKP	Female	0	Yes	No	08.02.2020	Yes	Fiber optic	Yes	Electronic ch	104.8	28.05.2022
6388-TABGU	Male	0	No	Yes	27.01.2019	Yes	DSL	No	Bank transfe	56.15	
9763-GRSKD	Male	0	Yes	Yes	05.02.2023	Yes	DSL	Yes	Mailed check	49.95	
7469-LKBCI	Male	0	No	No	07.11.2022	Yes	No	No	Credit card (	18.95	
8091-TTVAX	Male	0	Yes	No	27.05.2019	Yes	Fiber optic	No	Credit card (	100.35	
0280-XJGEX	Male	0	No	No	21.03.2019	Yes	Fiber optic	Yes	Bank transfe	103.7	30.03.2023
5129-JLPIS	Male	0	No	No	10.02.2022	Yes	Fiber optic	Yes	Electronic ch	105.5	
3655-SNQYZ	Female	0	Yes	Yes	01.07.2018	Yes	Fiber optic	No	Credit card (	113.25	

contracts.csv

A Transformation gemäss Formalisierung

- Aufgaben:

1. Instanzen konstruieren
2. Klassenattribut konstruieren
3. Attribute konstruieren

1. Instanzen konstruieren

- Typische Herausforderungen:

Kombination:
Instanzen existieren als Objekte in einer Datenbank / Tabelle, müssen aber mit Daten aus anderen Quellen angereichert werden

Beispiel Kreditvergabe:
starte mit Antragsformular, reichere mit Informationen aus Kundendatenbank an

Lösung: Joins (siehe später)

«Snapshots»:
Instanzen = Zustände eines Systems zu einem Zeitpunkt

Beispiel Teleflow:

??

Lösung: z.B. Instanzen mittels Zeitstempeln aufteilen

2. Klassenattribut konstruieren

- Typische Herausforderungen:

«Manuell»:
Wert des Klassenattributs ergibt sich nicht aus einem Ereignis, das sowieso aufgezeichnet wird

Beispiel Email-Klassifikation:
ob eine Email interessant ist oder nicht, muss ein Mensch entscheiden

Lösung: manuell klassifizieren, evtl. Mit Lernkurve kombinieren (siehe Kapitel “Evaluation” später), um Zeitpunkt für Abbruch festzulegen

«Komplex»:
Wert des Klassenattributs ergibt sich implizit aus Werten anderer Attribute

Beispiel TeleFlow:

??

Lösung: Berechnungen implementieren…

Feature Engineering

- Feature Selection
- Feature Transformation
- Feature Creation (Encoding, Binning)
- Feature Extraction (Automated in Deep Learning)

3. Attribute konstruieren / selektieren

- Typische Herausforderungen

«Transformation»:
Attribute müssen transformiert werden, um aus ihnen lernen zu können

Beispiel: Datumsangaben sind als Attribute meist ungeeignet (warum nochmal?). Aber man kann viel Interessantes aus ihnen ableiten, z.B. Dauer bis…, Saison, Ferienzeit, …

Lösung: Berechnungen implementieren…

«Feature Creation»:
Attribute können aus Informationen in anderen Datenquellen erstellt warden (siehe “Kombination” oben)

Beispiel TeleFlow:

??

Lösung: über Joins…

3. Attribute konstruieren / selektieren (2)

«Aggregation»:
Attribute müssen geeignet zusammengefasst werden, oft im Sinne einer “Historie”

Beispiel: TeleFlow

??

Lösung: Aggregation (z.B. mittels “Group By” und Aggregationsfunktion in SQL)

«Das n:m-Problem»:
Einer Instanz können mehrere verschiedene Werte eines Attributs zugeordnet sein

Beispiel: TeleFlow

??

Lösung: “Pivotieren”

3. Attribute konstruieren / selektieren (3)

«Feature Selection»:
Manche Attribute sind offensichtlich ungeeignet

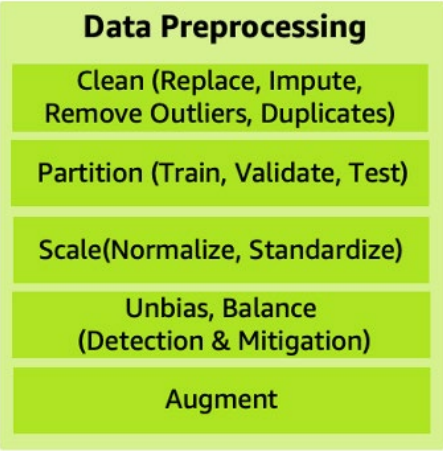
Beispiel: IDs, Datumsangaben, Attribute, die zur Vorhersagezeit nicht bekannt sind

Lösung: ungeeignete Attribute identifizieren und entfernen; aber: wann immer unklar ist, ob ein Attribut hilft, lieber behalten!

B Weitere Datenaufbereitung

- Herausforderungen:

1. Umgang mit fehlenden Werten
2. Einfache Typkonversion
3. Diskretisierung / One-Hot-Encoding
4. Normalisierung / Skalierung
5. Ausreisser + inkonsistente Werte entfernen
6. Duplikate entfernen



Fehlende Werte

- Umgang mit fehlenden Werten
  a. Ignorieren: Sonderwert «unknown» (kategorisch) oder «NaN» (numerisch) einfügen und bei Analysen ignorieren (nicht von allen Algorithmen unterstützt)
  b. Löschen 1: Instanzen mit fehlenden Attributwerten löschen (Zeile löschen)
  c. Löschen 2: Attribut mit fehlenden Werten aus allen Datenobjekten entfernen (Spalten löschen)
  d. Berechnen («impute») 1: Mittelwert / Median / Modus einfügen
  e. Berechnen («impute») 2: Wert aus ähnlichster Instanz des Datensatzes kopieren

Fehlende Werte – Beispiel

Kunden-ID	Datum letzte Reparatur	Anzahl Reparaturen (letzte 3 Jahre)	Region	Letztgekauftes Fahrrad (Kategorie)	Response
213232	May 23	1	city	Trekking	No
123244		5	city	Racing	Yes
546657	Nov 21	0	city	Single Speed	No
764566	Oct 20	0	rural	Single Speed	No
453232	Dec 21	0	rural	Mountain	No
785423	Apr 23	3	rural	Trekking	Yes
132567		2	city	Mountain	Yes
456467	Apr 22	1	rural	Mountain	No
342256	Nov 20	0	city	Single Speed	No
798888		4	rural	Trekking	Yes

- Beobachtungen:
  a. einige Werte fehlen in der zweiten Spalte
  b. die dritte Spalte (# Reparaturen) korreliert ziemlich gut mit dem Klassenattribut
  c. es besteht evtl. eine gewisse Korrelation zwischen der zweiten und dritten Spalte

- Frage: Was passiert, wenn Instanzen (=Zeilen) mit fehlenden Werten entfernt werden? Oder wenn wir die zweite Spalte entfernen?

Weitere Aufbereitungen

«Cleaning»:
Duplikate, Ausreisser, inkonsistente Werte finden und bereinigen

Beispiel: vergessenes Dezimaltrennzeichen bei einzelnen Preisen (→ Ausreisser), Verwendung unterschiedlicher Bezeichner, Datumsformate etc. (→ inkonsistente Werte)

«Einfache Typkonversion»:
Achtung: beim Laden von csv-Dateien werden manchmal kategoriale Attribute als numerisch interpretiert

Beispiel: Monat wird mit den Zahlen 1-12 kodiert und als numerisch interpretiert → es wird teils damit "gerechnet", was meist nicht korrekt ist

Komplexere Typkonversion

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

«Diskretisierung»:
Konversion numerisch → kategorial

Beispiel: ein Entscheidungsbaum-Algorithmus akzeptiert nur kategoriale Attribute → "Alter" wird diskretisiert in "jung", "mittelalt" und "alt"

Herausforderung: Kategorien definieren ("Bins"), kann auf verschiedene Weise passieren, auch "überwacht"

«One-Hot-Encoding»:
Konversion kategorial→ numerisch

Beispiel: ein neuronales Netz akzeptiert nur numerische Attribute → One-Hot-Encoding als "Trick" (siehe Bild oben)

Normalisierung und Skalierung

Ziel: Skaleneffekte und Probleme durch Ausreisser vermeiden

- **Skalierung** (Ausreisser): Transformation der Daten, sodass sie eine Normalverteilung mit einem Mittelwert von 0 und einer Standardabweichung von 1 aufweisen.

$$x_{neu} = \frac{x_{alt} - \bar{x}}{\sigma}$$

- **Normalisierung** (Skaleneffekte): Transformation in das Intervall [0,1], z.B. min-max

$$x_{neu} = \frac{x_{alt} - min}{max - min}$$