Applied Analytical Data Science

Teil 3: Preprocressing I

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

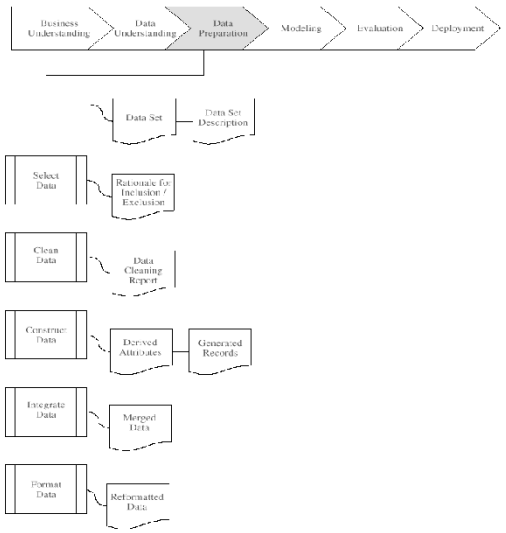
Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Datenaufbereitung im Kontext des CRISP-DM Models
- Ziel der Datenaufbereitung
- Daten Selektion
- Daten Cleaning
  - Kodierungs- & Typ-Korrekturen
  - Behandlung fehlender Werte (Missing Values)
    - Nicht anwendbare Attribute
    - Unbekannte Werte
  - Behandlung von Fehlern (Noise & Outlier)
  - Skalierung der Daten: Normalisierung & Gewichtung

Datenaufbereitung im CRISP-DM Model



Datenaufbereitung im CRISP-DM Model

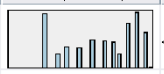
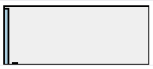
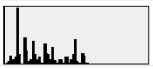
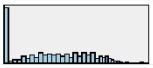
Bereits bekannt sind:

- Aus der Business Understanding Phase:
  - Business und Data Mining Ziele, Erfolgskriterien
  - Data Mining Tool
- Aus der Data Understanding Phase:
  - Welche Daten sind verfügbar
  - Mit welchem Inhalt (Data Description Report)
  - Anfragen, Visualisierungen und Reporting zu den Daten (Data Exploration Report)
  - Qualität der verfügbaren Daten (Data Quality Report)

Data Description Report Beispiel

Variable	Description
ODATEDW	Origin Date. Date of donor's first gift to PVA format YYMM.
OSOURCE	Origin Source (Only 1rst 3 bytes are used) Defaulted to 00000 for conversion, Code indicating which mailing list the donor was originally acquired from A nominal or symbolic field.
TCODE	Donor title code 000 = _ 001 = MR. . . . 132 = YOUR IMPERIAL MAJEST 210 = PROF.
STATE	State abbreviation (a nominal/symbolic field)
ZIP	Zipcode (a nominal/symbolic field)
MAILCODE	Mail Code: " "= Address is OK B = Bad Address
PVASTATE	EPVA State or PVA State Indicates whether the donor lives in a state served by the organization's EPVA chapter P = PVA State E = EPVA State (Northeastern US)
DOB	Date of birth (YYMM, Year/Month format.)

Visualisierung und Reporting Beispiel

Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ODATEDW		Range	8306	9701	9141.363	343.455	-0.267	--	95412
OSOURCE	Too many values	Discrete	--	--	--	--	--	--	94484
TCODE		Range	0	72002	54.223	953.844	34.211	--	95412
STATE		Discrete	--	--	--	--	--	57	95412
ZIP	Too many values	Discrete	--	--	--	--	--	--	95412
MAILCODE		Discrete	--	--	--	--	--	2	1399
PVASTATE		Discrete	--	--	--	--	--	3	1458
DOB		Range	0	9710	2723.603	2132.241	0.163	--	95412

Visualisierung und Reporting Beispiel

Role	Name	Type	Statistics	Range	
regular	ODATEDW	numeric	avg = 9141.363 +/- 343.455	[8306.000 ; 9701.000]	0
regular	OSOURCE	text	mode = MBC (4539), least = BEL (1)	MBC (4539), SYN (3563), AML (3430), BHG (3324), IMP (2986), ARG (2409),	928
regular	TCODE	numeric	avg = 54.223 +/- 953.844	[0.000 ; 72002.000]	0
regular	STATE	text	mode = CA (17343), least = DC (1)	CA (17343), FL (8376), TX (7535), IL (6420), MI (5654), NC (4160), WA (357	0
regular	ZIP	text	mode = 85351 (61), least = 51033 (	85351 (61), 92653 (59), 85710 (54), 95608 (50), 60619 (45), 89117 (45)	0
regular	MAILCODE	text	mode = B (1399), least = (0)	(0), B (1399)	94013
regular	PVASTATE	text	mode = P (1453), least = (0)	(0), E (5), P (1453)	93954
regular	DOB	numeric	avg = 2723.603 +/- 2132.241	[0.000 ; 9710.000]	0
regular	NOEXCH	numeric	avg = 0.003 +/- 0.055	[0.000 ; 1.000]	42
regular	RECINHSE	text	mode = X (6703), least = (0)	(0), X (6703)	88709
regular	RECP3	text	mode = X (2017), least = (0)	(0), X (2017)	93395
regular	RECPGVG	text	mode = X (114), least = (0)	(0), X (114)	95298
regular	RECSWEEP	text	mode = X (1617), least = (0)	(0), X (1617)	93795
regular	MDMAUD	text	mode = XXXX (95118), least = I5CM	C1CM (65), D1CM (20), I1CM (37), L1CM (44), C2CM (24), D2CM (28), I2CM	0
regular	DOMAIN	text	mode = R2 (13623), least = (0)	(0), C1 (6145), R1 (1358), S1 (11503), T1 (4982), U1 (4510), C2 (8264), R	2316
regular	CLUSTER	numeric	avg = 27.923 +/- 14.450	[1.000 ; 53.000]	2316
regular	AGE	numeric	avg = 61.612 +/- 16.664	[1.000 ; 98.000]	23665
regular	AGEFLAG	text	mode = E (57344), least = (0)	(0), E (57344), I (8520)	29548
regular	HOMEOWNR	text	mode = H (52354), least = (0)	(0), H (52354), U (20830)	22228
regular	CHILD03	text	mode = M (869), least = (0)	(0), B (40), F (237), M (869)	94266
regular	CHILD07	text	mode = M (1061), least = (0)	(0), B (97), F (408), M (1061)	93846
regular	CHILD12	text	mode = M (1149), least = (0)	(0), B (142), F (520), M (1149)	93601
regular	CHILD18	text	mode = M (1442), least = (0)	(0), B (263), F (1142), M (1442)	92565

Ziel der Datenaufbereitung

- Bereitstellung aller **relevanten** Daten in der **nützlichsten** Form für das DM-tool
  - Selektion der relevanten Daten, ausreichender Qualität
  - Erhöhung der Datenqualität
  - Erstellung abgeleiteter Attribute die nützlicher sind
  - Erstellung abgeleiteter Datensätze die nützlich/notwendig sind
  - Erfüllen aller Formatbeschränkungen des DM-Tools
- Jegliche Manipulation der (Lern) Daten muss automatisiert sein, damit sie auch auf Test- und Anwendungsdaten angewendet werden kann.

Formatbeschränkungen von DM-Tools

- Skalare und ordinale Attribute müssen numerisch sein
- Nominale Attribute müssen Char oder Boolean sein
- Keine unbekannten Werte für alle/bestimmte Attributtypen
- Nur numerische oder nur nicht-numerische Attribute
- Nicht mehr als X verschiedene Werte für nominale Attribute
- Vergleichbare Skalen für numerische Attribute
- Keine Schlüssel (ähnlichen) Attribute
- Ein einfaches, Trenner-separiertes File (Trenner muss nicht in Werten sein)
- Beispielmenge muss balanciert sein

Heutiges Programm

- Datenaufbereitung im Kontext des CRISP-DM Models
- Ziel der Datenaufbereitung
- Daten Selektion
- Daten Cleaning
  - Kodierungs- & Typ-Korrekturen
  - Behandlung fehlender Werte (Missing Values)
    - Nicht anwendbare Attribute
    - Unbekannte Werte
  - Behandlung von Fehlern (Noise & Outlier)
  - Skalierung der Daten: Normalisierung & Gewichtung

Daten Selektion

- Weglassen von irrelevanten Daten
  - Irrelevante Attribute, z.B. TCODE
  ⇒Data Description Report
  - Weglassen von irrelevanten Datensätzen
  ⇒Data Collection Report
- Weglassen von Daten zu schlechter Qualität
  - Attribute mit vielen fehlenden Werten, z.B. PVASTATE
  ⇒Vorsicht: Correlation mit dem Ziel prüfen (95:5 in der Übung)
  - Datensätze mit zuvielen fehlenden Werten, oder Messfehlern
  ⇒Data Exploration & Quality Report

Heutiges Programm

- Datenaufbereitung im Kontext des CRISP-DM Models
- Ziel der Datenaufbereitung
- Daten Selektion
- Daten Cleaning
  - Kodierungs- & Typ-Korrekturen
  - Behandlung fehlender Werte (Missing Values)
    - Nicht anwendbare Attribute
    - Unbekannte Werte
  - Behandlung von Fehlern (Noise & Outlier)
  - Skalierung der Daten: Normalisierung & Gewichtung

Codierungs & Typ korrekturen

- Die meisten DM Tools (incl. SPSS, Clementine) setzten
  - Numerische Codierung = skalares (range) Attribut
  - Character Codierung = Nominales (diskret) Attribut

  ⇒Durch Vergleich Data Description und Data Exploration Report kann man diese Fälle (z.B.: TCODE, ZIP, …) finden und korrigieren.
- Unbekannte/Fehlende Werte sind sehr unterschiedlich repräsentiert (z.B.: 0, ””, ” ”, ”X”, …. oder korrekt $null$)

  ⇒Sie müssen Vereinheitlicht werden, und der Data Description Report muss entsprechend aktualisiert werden.

Fehlende Daten

- Nichtverfügbarkeit
  - bei vielen Tupeln sind bestimmte Attribute nicht gefüllt,
  - unbekannt vs. nicht anwendbar
- Informationsgehalt nicht anwendbarer Werte
  - unbekannte und nicht anwendbare Werte sind unterschiedlich zu behandeln
- Ursachen unbekannter Werte
  - technische Fehlfunktionen
  - Löschungen aufgrund (vermeintlicher) Inkonsistenzen mit anderen Daten
  - Nichteingabe (aufgrund fehlender Relevanz zum Erfassungszeitpunkt)
  - nicht registrierte Änderungen
- Erfordernis der nachträglichen Ermittlung bzw. Erschliessung

Massnahmen bei fehlenden Daten

- Ersetze fehlenden Wert durch globale Konstante (z.B. unknown, $null$)
  - Data Ming Tool muss Daten mit unbekannten Werte akzeptieren
- Ignoriere Tupel
  - wenn Wert zum Zielattribut fehlt
  - falls viele Attributwerte fehlen
- Ignoriere Attribut
  - wenn kaum Werte für dieses Attribut vorhanden sind.
  - Problematisch bei ungleich verteiltem Ziel (Korrelation testen)
- Ergänze Wert manuell
  - In der Regel: zu Schwierig, zuviel Zeitaufwand, zu hohe Kosten
- Schätzen unbekannter Werte

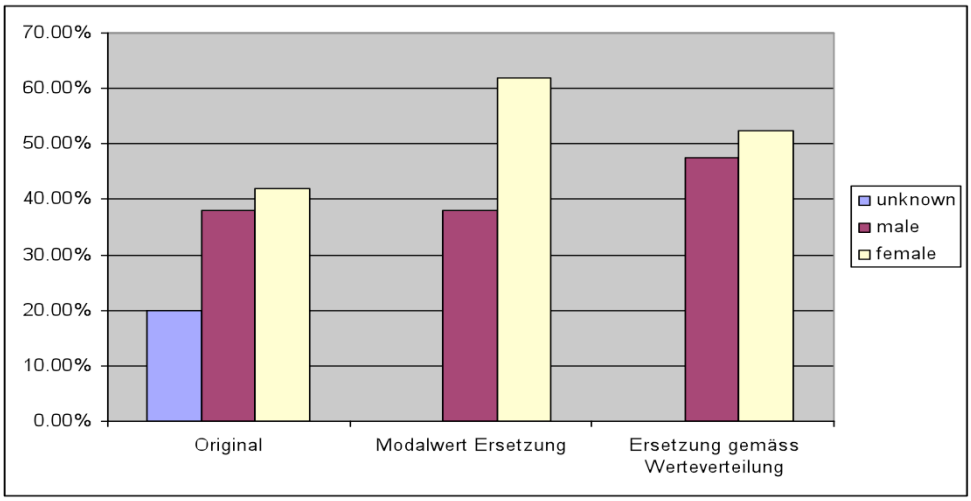
Schätzen unbekannter Werte

- Verwende einen Mittelwert (Mean, Median, Modal) des Attributes
  z.B.: durchschnittliches Einkommen oder Alter
- Verwende den Mittelwert (Mean, Median, Modal) des Attributes bezogen auf alle Tupel, die zu einer gleichen Klasse gehören
  z.B.: Durchschnittseinkommen von Kunden nach Risikogruppe
- Erzeuge einen zufälligen Wert gemäss Werteverteilung des Attributes
- Verwende den wahrscheinlichsten Wert inferiert durch Regression bzw, Klassifikation, d.h. Vorhersage aufgrund der Korrelation mit anderen Attributen

Grundsätzliches Problem:

Beeinflussung der Daten insbesondere des betroffenen Attributes und seiner Beziehung zu den anderen Attributen

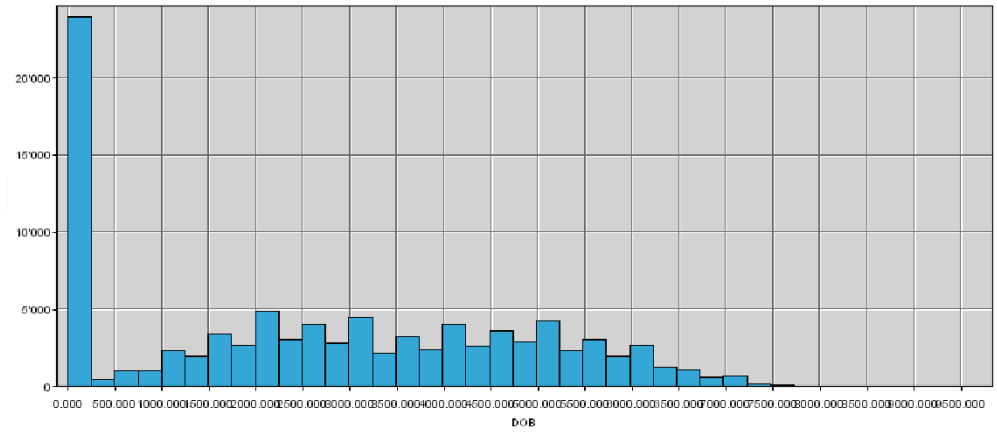
Einfluss von Ersetzungsmassnahmen



Einfluss von Ersetzungsmassnahmen

Position	Original sample	Position 11 missing	Preserve mean as estimate	Preserve variance as estimate
1	0.0886	0.0886	0.0886	0.0886
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	?	**0.3731**	**0.6622**
Mean	0.4023	0.3731	0.3731	0.3994
Standard deviation	0.2785	0.2753	0.2612	0.2753
Size of error in the estimate			0.3208	0.0317

Einfluss von Ersetzungsmassnahmen (DOB)



Schätzen unbekannter Werte mit DM-Tools

- Mit Hilfe von DM-tools können Funktionen zur Vorhersage fehlender Werte gelernt werden.
- Klassifikationslerner zur Vorhersage nominaler Attribute
- Regressionslerner zur Vorhersage skalarer Attribute
- Die Datensätze mit gefülltem Attribut dienen als Lern- und Testdaten
  - Das Attribut mit fehlenden Werten ist das Zielattribut
  - Andere dafür relevante Attribute dienen als Eingabeattribute
- Die Datensätze mit fehlendem Attributwert sind die Anwendungsdaten, bei denen der Wert vorhergesagt wird.

Heutiges Programm

- Datenaufbereitung im Kontext des CRISP-DM Models
- Ziel der Datenaufbereitung
- Daten Selektion
- Daten Cleaning
  - Kodierungs- & Typ-Korrekturen
  - Behandlung fehlender Werte (Missing Values)
    - Nicht anwendbare Attribute
    - Unbekannte Werte
  - Behandlung von Fehlern (Noise & Outlier)
  - Skalierung der Daten: Normalisierung & Gewichtung

Noise

- Noise: Zufallsfehler oder Abweichungen einer Variable
- Risiko der unerwünschten Beinflussung der Mining-Verfahren:
  - fehlerhafte “Interpretation” einzelner Datensätze
  - Eigenschaften der Datenmenge vs. Eigenschaft einzelner Datensätze
- Ursachen für fehlerhafte Attributwerte:
  - fehlerhafte Werkzeuge zur Datenerhebung/-extraktion
  - Dateneingabefehler
  - Datenübertragungsprobleme
  - technische Einschränkungen
  - Inkonsistenzen bei Namenskonventionen

Massnahmen gegen Noise

Grundidee: Entfernung von Noise durch Anpassung von Werten an ihre „Umgebung“

- Binning
  - Sortieren der Daten und Partitionierung in “bins” (gleicher Tiefe)
  - Glättung durch Durchschnitt, Median oder Grenzen des “bins”
- Clustering
  - Entdeckung und Entfernen von “Ausreissern”
- Semi-automatische Behandlung
  - Entdeckung verdächtiger Werte und manuelle Überprüfung
- Regression
  - Glätten durch Anpassung der Daten an eine Regressions-Funktion

Einfache Diskretisierung: Binning

- Partitionierung nach Distanz
  - unterteile Wertebereich in *N* Intervalle gleicher Grösse
  - wenn *A* und *B* die kleinsten bzw. grössten Attributwerte sind, ist die Intervallbreite: $W = (B\text{-}A)/N$.
  - sehr naheliegend
  - Ausreisser dominieren die Darstellung

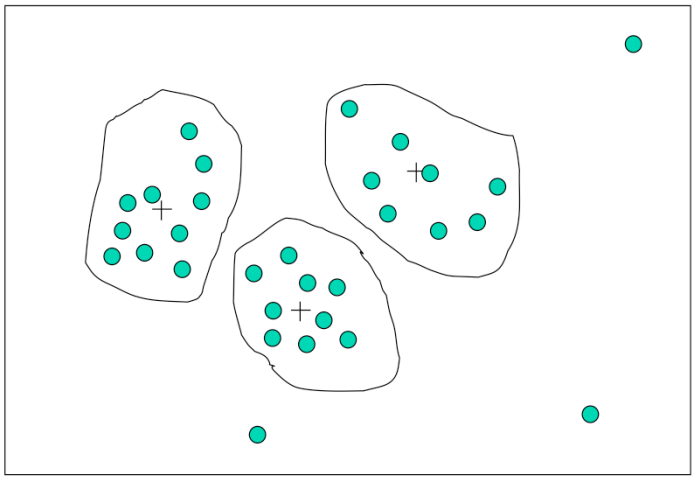
- Partitionierung nach Häufigkeit
  - unterteile den Wertebereich in *N* Intervalle mit jeweils ungefähr der gleichen Anzahl auftretender Werte,
  - gute Skalierbarkeit

Binning-Methoden zur Glättung

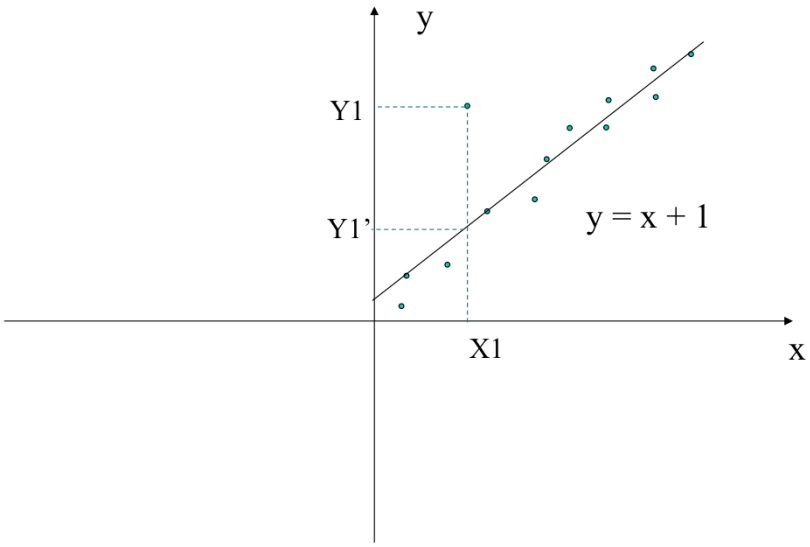
Beispiel: Sortierte Daten (z.B. Preise)
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partitionierung in bins (gleicher Tiefe):
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Glättung durch Mittelwerte:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Glättung durch Grenzen
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Cluster-Analyse



Regression



Heutiges Programm

- Datenaufbereitung im Kontext des CRISP-DM Models
- Ziel der Datenaufbereitung
- Daten Selektion
- Daten Cleaning
  - Kodierungs- & Typ-Korrekturen
  - Behandlung fehlender Werte (Missing Values)
    - Nicht anwendbare Attribute
    - Unbekannte Werte
  - Behandlung von Fehlern (Noise & Outlier)
  - Skalierung der Daten: Normalisierung & Gewichtung

Normalisierung & Gewichtung

- Um ungewollte Gewichtungen bei skalaren Attributen in Abstandsmassen zu vermeiden, sollten sie Normalisiert werden.
- Skalierung von Werten auf einen kleinen, spezifizierten Wertebereich
  - min-max Normalisierung
  - z-score Normalisierung
  - Normalisierung durch Dezimal-Skalierung
- Möglicherweise (Verteilung) Log-Skalierung (v‘ = log(v)) zuvor.
- Möglicherweise anschliessend eine gewollte Gewichtung.

Normalisierung

- min-max Normalisierung

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score Normalisierung

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- Normalisierung durch Dezimalskalierung

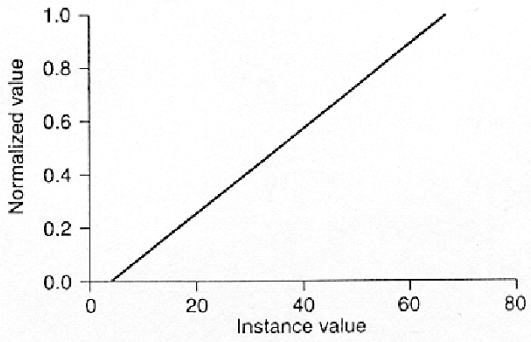
$$v' = \frac{v}{10^j}$$ wobei j der kleinste Integer ist, für den Max(|v'|) < 1

Beispiel: Lineare Skalierung

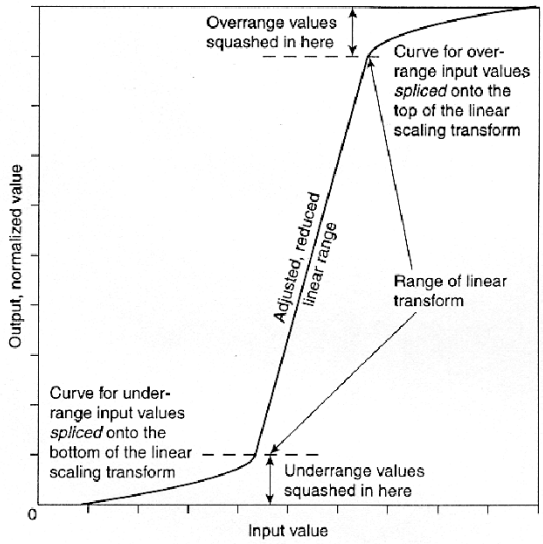
Number	Normalized
40.12	0.56
31.29	0.42
43.85	0.63
40.60	0.57
54.21	0.80
46.94	0.68
28.06	0.36
60.97	0.91
57.39	0.85
66.09	1.00
6.45	0.00
9.53	0.05
52.83	0.78
47.44	0.69
21.00	0.24
43.31	0.62
14.40	0.13
37.42	0.52
16.47	0.17
61.51	0.92

Maximum	66.09
Minimum	6.45
Difference	59.64

$$v_{norm} = \frac{v_i - \min(v_1 \ldots v_n)}{\max(v_1 \ldots v_n) - \min(v_1 \ldots v_n)}$$



Soft-Max-Skalierung



Normalisierung für obere Ausreisser:
1-1/v

Hilfsmittel: Logistik-Funktion

Normalisierung für untere Ausreisser:
1/(1+(Min-v))

Fazit

- Die Datenaufbereitung dient der Bereitstellung aller **relevanten** Daten in der **nützlichsten** Form für das DM-tool
- Jegliche Manipulation der (Lern) Daten muss automatisiert sein, damit sie später auch auf Test- und Anwendungsdaten angewendet werden kann.
- Data Cleaning soll die Qualität der Daten verbessern, mindestens muss es die Daten bereinigen, so dass die Eingaberestriktionen des gewählten DM-Tools erfüllt werden (z.B. keine fehlenden Werte erlaubt).
- Normalisierung der skalaren Daten hilft ungewollte Gewichtungen der Attribute zu vermeiden.

Literatur

- Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.