

Big Data Anwendungen

Aufgabenblatt 1

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
 - Aufgabe 1 – Eigenschaften von Attributen
 - Aufgabe 2 – Verständnisfragen
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Allgemeines zu Aufgabenblättern

- Aufgaben orientieren sich an Klausuraufgaben
 - Umfang vergleichbar mit Klausuraufgaben
 - Schwierigkeitsgrad vergleichbar mit Klausuraufgaben
 - Fokus analog Klausuraufgaben
- Punktzahlen zu den Übungsaufgaben
 - Punkte analog Punkten in Klausuraufgaben
 - Jeder Punkt entspricht einer Minute Bearbeitungszeit
- Inhalt der Aufgabenblätter 1 bis 6
 - Aufgabenblätter orientieren sich an Kapiteln der Vorlesung
 - In Semestern vor 2020: Nur Inhalt in drei Aufgabenblättern á 4 Stunden
 - Ab SS2020: Unterteilung in 1,5 Stundenblöcke (zur besseren Verdaulichkeit während Corona)
- Inhalt der späteren Aufgabenblätter
 - Einführung in Tools zur Durchführung entsprechender Analysen am PC
 - Aufgaben zur selbständigen Bearbeitung

Aufgabe 1 (a) - Aufgabenstellung

- Gegeben seien die folgenden Werte und ihre absolute Häufigkeit

Wert	Absolute Häufigkeit
1	2
2	4
3	1
4	1

- Bestimmen Sie Varianz und Standardabweichung.

(4 Punkte)

Aufgabe 1 (a) – Lösung (Varianz)

- Vorbereitung: Ermittlung des Mittelwerts

- Allgemein: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$

- Hier: $\bar{x} = \frac{1 \cdot 2 + 2 \cdot 4 + 3 \cdot 1 + 4 \cdot 1}{8} = \frac{17}{8} = 2,125$

Wert	Abs. #
1	2
2	4
3	1
4	1

- Varianz:

- Allgemein: $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$

- Hier:

$$s^2 = \frac{1}{8} (2 \cdot [1 - 2,125]^2 + 4 \cdot [2 - 2,125]^2 + [3 - 2,125]^2 + [4 - 2,125]^2)$$

$$s^2 = \frac{1}{8} (2 \cdot [-1,125]^2 + 4 \cdot [-0,125]^2 + [0,875]^2 + [1,875]^2)$$

$$s^2 \approx \frac{1}{8} (2 \cdot 1,266 + 4 \cdot 0,016 + 0,766 + 3,516)$$

$$s^2 \approx \frac{1}{8} (6,875) \approx 0,859$$

Aufgabe 1 (a) – Lösung (Standardabweichung)

- Bisher:

- Mittelwert: $\bar{x} = 2,125$
- Varianz: $s^2 \approx 0,859$

- Standardabweichung:

- Allgemein: $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{s^2}$
- Hier: $s = \sqrt{0,859} = 0,927$

- Warum diese Teilaufgabe?

- Berechnung von Standardabweichung trivial
- Standardabweichung bildet Grundlage für Schiefe und Wölbung

Wert	Abs. #
1	2
2	4
3	1
4	1

Aufgabe 1 (b) - Aufgabenstellung

- Gegeben seien die folgenden Werte und ihre absolute Häufigkeit

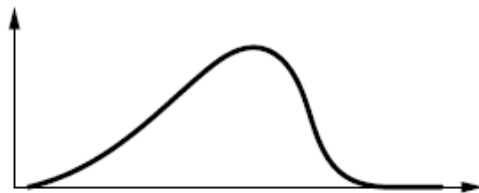
Wert	Absolute Häufigkeit
1	2
2	4
3	1
4	1

- Bestimmen Sie die Schiefe.

(4 Punkte)

Schiefe – Idee

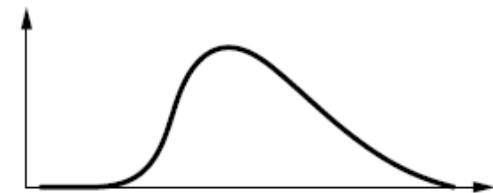
- Themenkomplex
 - Lageparameter
 - Eigenschaften der Verteilung der Daten
- Schiefe (graphisch)



$\alpha_3 < 0$: rechtssteil



$\alpha_3 = 0$: symmetrisch



$\alpha_3 > 0$: linkssteil

- Ziel (also)
Sind die Daten eher „links“ (Richtung Minimum) oder ... eher „rechts“ (Richtung Maximum) der Verteilung
- Alternative Betrachtung
Ermitteln eines Histogramms und „Draufgucken“

Aufgabe 1 (b) – Lösung (Schiefe)

- Bisher:

- Mittelwert: $\bar{x} = 2,125$
- Standardabweichung: $s \approx 0,927$

Wert	Abs. #
1	2
2	4
3	1
4	1

- Schiefe:

- Allgemein: $\alpha_3 = \frac{1}{n \cdot s^3} \sum_{i=1}^n (x_i - \bar{x})^3 = \frac{1}{n} \sum_{i=1}^n z_i^3$ mit $z_i = \frac{x_i - \bar{x}}{s}$

- Hier:

$$\alpha_3 = \frac{1}{8 \cdot (0,927)^3} (2 \cdot [1 - 2,125]^3 + 4 \cdot [2 - 2,125]^3 + [3 - 2,125]^3 + [4 - 2,125]^3)$$

$$\alpha_3 = \frac{1}{8 \cdot (0,927)^3} (2 \cdot [-1,125]^3 + 4 \cdot [-0,125]^3 + [0,875]^3 + [1,875]^3)$$

$$\alpha_3 \approx \frac{1}{8 \cdot (0,927)^3} (2 \cdot [-2,848] + 4 \cdot [-0,008] + [0,670] + [6,592])$$

$$\alpha_3 \approx \frac{1}{8 \cdot (0,927)^3} \cdot 4,406 \approx 0,691$$

⇒ Die Verteilung der Datenpunkte ist „linkssteil“ (da $\alpha_3 > 0$)

Aufgabe 1 (c) - Aufgabenstellung

- Gegeben seien die folgenden Werte und ihre absolute Häufigkeit

Wert	Absolute Häufigkeit
1	2
2	4
3	1
4	1

- Bestimmen Sie die empirische Wölbung.

(4 Punkte)

Empirische Wölbung – Idee

- Themenkomplex (analog Schiefe)
 - Lageparameter
 - Eigenschaften der Verteilung der Daten
- Empirische Wölbung (graphisch)



$\gamma < 0$: platykurtisch,
flachgipflig



$\gamma = 0$: mesokurtisch,
normalgipflig



$\gamma > 0$: leptokurtisch,
steilgipflig

- Ziel (also)
Sind die Daten eher „weit“ oder eher „dicht“ um den Mittelwert der Verteilung
- Alternative Betrachtung (vgl. Schiefe)
Ermitteln eines Histogramms und „Draufgucken“

Aufgabe 1 (c) – Lösung (empirische Wölbung)

- Bisher:
 - Mittelwert: $\bar{x} = 2,125$
 - Standardabweichung: $s \approx 0,927$
- Empirische Wölbung:
 - Allgemein:

Wert	Abs. #
1	2
2	4
3	1
4	1

$$\gamma = \alpha_4 - 3$$

wobei $\alpha_4 = \frac{1}{n \cdot s^4} \sum_{i=1}^n (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^n z_i^4$ mit $z_i = \frac{x_i - \bar{x}}{s}$

- Hier:

$$\alpha_4 = \frac{1}{8 \cdot (0,927)^4} (2 \cdot [1 - 2,125]^4 + 4 \cdot [2 - 2,125]^4 + [3 - 2,125]^4 + [4 - 2,125]^4)$$

$$\alpha_4 = \frac{1}{8 \cdot (0,927)^4} (2 \cdot [-1,125]^4 + 4 \cdot [-0,125]^4 + [0,875]^4 + [1,875]^4)$$

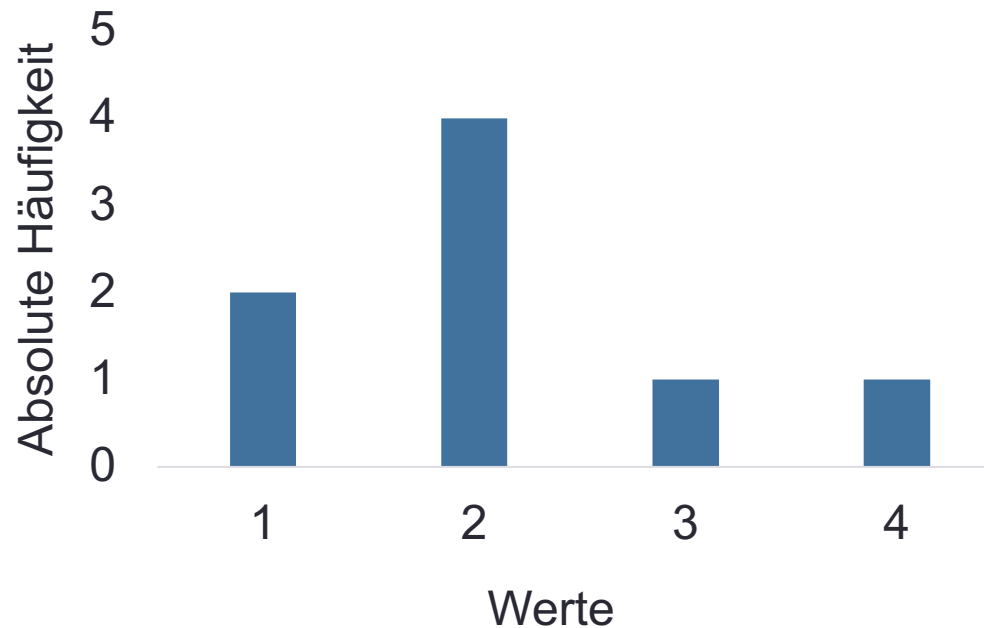
$$\alpha_4 \approx \frac{1}{8 \cdot (0,927)^4} (2 \cdot 3,204 + 4 \cdot 0,001 + 0,586 + 12,360) \approx 2,734$$

$$\gamma = \alpha_4 - 3 \approx 2,734 - 3 = -0,266$$

⇒ Die Verteilung der Datenpunkte ist „platykurtisch“ (da $\gamma < 0$)

Exkurs: Plausibilität der Ergebnisse

- Visualisierung der Verteilung



- Verteilung ist offensichtlich
 - linkssteil (Datenpunkte sind eher links der „Mitte“) und
 - platykurtisch (Verteilung ist innerhalb des Wertebereichs „flach“)
- D.h. die numerischen Ergebnisse decken sich mit Visualisierung

Aufgabe 1 (d) - Aufgabenstellung

- Gegeben seien die folgenden Werte und ihre absolute Häufigkeit

Wert	Absolute Häufigkeit
1	2
2	4
3	1
4	1

- Standardisieren Sie die obigen Werte.

(4 Punkte)

Aufgabe 1 (d) – Lösung (Standardisierung)

- Bisher:
 - Mittelwert: $\bar{x} = 2,125$
 - Standardabweichung: $s \approx 0,927$

- Standardisierung:

- Allgemein: $z_i = \frac{x_i - \bar{x}}{s}$

- Hier:

Wert	Abs. #	Transformation	Standardisierter Wert
1	2	$\frac{1-2,125}{0,927}$	-1,214
2	4	$\frac{2-2,125}{0,927}$	-0,135
3	1	$\frac{3-2,125}{0,927}$	0,944
4	1	$\frac{4-2,125}{0,927}$	2,023

- Nutzen der Standardisierung
 - Vergleichbarer Wertebereich aller Attribute
 - Richtung der Abweichung vom Mittelwert ablesbar (positive bzw. negative Werte)
 - Wenig sensitiv bezüglich Ausreißern als Min-Max-Skalierung

Aufgabe 1 (e) - Aufgabenstellung

- Gegeben seien die folgenden Werte und ihre absolute Häufigkeit

Wert	Absolute Häufigkeit
1	2
2	4
3	1
4	1

- Min-Max skalieren Sie die obigen Werte.

(4 Punkte)

Aufgabe 1 (e) – Lösung (Min-Max-Skalierung)

- Min-Max-Skalierung:

- Allgemein:
$$z_i = \frac{x_i - \min_{\forall j} x_j}{\max_{\forall j} x_j - \min_{\forall j} x_j}$$

- Hier:

- $\min_{\forall j} x_j = 1$ und $\max_{\forall j} x_j = 4$

- damit:

Wert	Abs. #	Transformation	Standardisierter Wert
1	2	$\frac{1-1}{4-1}$	0
2	4	$\frac{2-1}{4-1}$	$\frac{1}{3}$
3	1	$\frac{3-1}{4-1}$	$\frac{2}{3}$
4	1	$\frac{4-1}{4-1}$	1

- Nutzen der Min-Max-Skalierung

- Wertebereich sicher zwischen 0 und 1 (aber empfindlich bzgl. Ausreißern)
 - Transformation ist leicht verständlich

Agenda

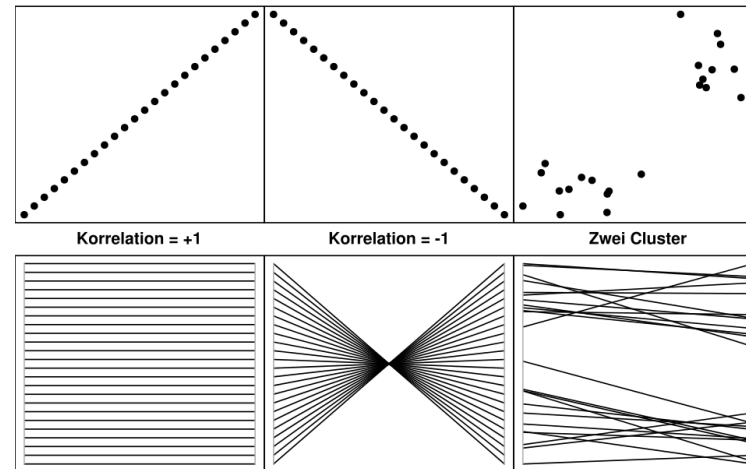
- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
 - Aufgabe 1 – Eigenschaften von Attributen
 - Aufgabe 2 – Verständnisfragen
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 2 (a) - Aufgabenstellung

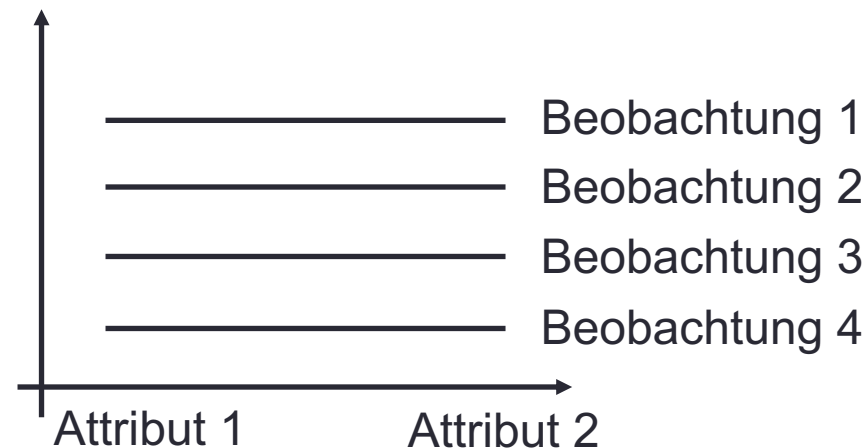
- Visualisieren Sie ein Parallelkoordinaten-Diagramm für 2 korrelierte Attribute mit 4 beliebigen Beobachtungen. **(2 Punkte)**

Aufgabe 2 (a) – Lösung (Parallelkoordinaten)

- Kurze Wiederholung



- Lösung

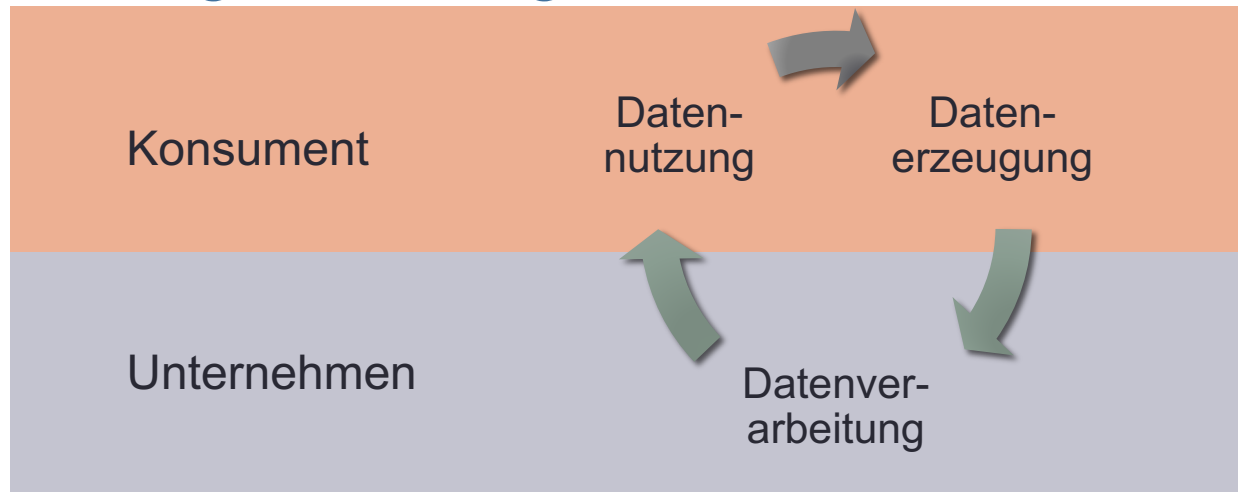


Aufgabe 2 (b) - Aufgabenstellung

- Erläutern Sie anhand der Datenerzeugung, Datenverarbeitung und Datennutzung wie Big Data Probleme entstehen und grenzen Sie dies von traditionellem Data Mining ab.

(6 Punkte)

Wiederholung Vorlesung: Primärkreislauf



- Datenerzeugung
 - Benutzung von Produkten, Services, Kommunikationsmedien
 - „Online sein“
- Datenverarbeitung
 - Schaffung neuer Produkte
 - Schaffung technischer Innovationen
- Datennutzung
 - Einsatz neuer Produkte und Innovationen liefert neue Daten
 - Datenkreislauf wird beschleunigt ⇒ Big Data

Aufgabe 2 (b) – Lösung (Primärkreislauf)

- Entstehung Big Data Probleme
 - Durch Primärkreislauf Erhöhung des Datenaufkommen
 - Beispiel Google Maps:
 - Datenerzeugung: Sammlung von Bewegungsdaten der Nutzer
 - Datenverarbeitung: Verbesserung Verkehrsprognose durch Nutzerdaten
 - Datennutzung: Verarbeitet Nutzerdaten für andere Nutzern interessant
 - Konsequenz:
Durch Primärkreislauf Einsatz der Lösung interessant
(Bsp. Google Maps: Qualität übersteigt traditionelle Stauprognose)
- Abgrenzung zu traditionellem Data Mining
 - Traditionell lediglich Verarbeitung der Daten
(kein Zurückspielen der Daten an die Nutzer)
 - Dadurch kein Anstieg an beobachteten Daten

Aufgabe 2 (c) - Aufgabenstellung

- Nennen und beschreiben Sie kurz typische Fehler in Daten. **(5 Punkte)**

Wiederholung Vorlesung: Datenfehler

Guthaben	Telefon	CC seit	CC histor	Stand	Besitz	Alter	seit	
0ja		6	kritisch_o_offene_Zahlung	W	Wohnung	67	4	
1nein		48	alles_bezahlt	0W	Wohnung	22	2	
nan	nein	12	kritisch_o_offene_Zahlung	20.96	0M_single	Wohnung	49	3
0nein		42	alles_bezahlt	78.82	0M_single	Versicherung	45	3
0nein		37	alles_bezahlt	3.7	0M_single	unbekannt	53	2
nan	ja	35	alles_bezahlt	35	M_single	unbekannt	35	2
nan	1ja	48	alles_bezahlt	48	0M_single	Auto	35	2
nan	nein	12	alles_bezahlt	30.59	3M_geschieden	Wohnung	61	3
1nein		30	kritisch_o_offene_Zahlung	52.34	0M_verheiratet	Auto	28	0
1nein		12	alles_bezahlt	12.95	0W_verheiratet	Auto	25	1
0nein		48	alles_bezahlt	43.08	0W_verheiratet	Versicherung	24	1
1ja		12	alles_bezahlt	15.67	0W_verheiratet	Auto	22	2
0nein		24	kritisch_o_offene_Zahlung	11.99	0M_single	Auto	60	4
nein		15	alles_bezahlt	14.03	0W_verheiratet	Auto	3	2
		24	alles_bezahlt	12.82	1W_verheiratet	Auto		2

Ordinale Attribute in kategorischen versteckt

Fehlende Werte unterschiedlich kodiert (nan oder „ „)

Binäre Werte unterschiedlich kodiert ([0, 1] oder [ja, nein])

Zwei Attribute in einem Feld

Dauer in Monaten oder Jahren

Aufgabe 2 (c) – Lösung (Unreine Daten)

- Fehlende Werte
Für einzelne Beobachtungen ist Ausprägung des Attributs unbekannt
- Unterschiedliche Kodierung identischer Werte
 - Abbildung in unterschiedlichen Skalen / Datentypen
 - Abbildung einzelner Attributwerte über unterschiedliche Attributwerte
- Kombination mehrere Attribute zu einem Attribut
 - Unterschiedliche Informationen zu einer Zeichenkette verknüpft
 - Kodierung von Attributkombinationen zu einem Wert
- Ordnung innerhalb von Attributen durch Kategorien verloren
 - Attribute werden auf sprechende Beschreibung abgebildet
 - Ordnung der Attribute geht verloren

Aufgabe 2 (d) - Aufgabenstellung

- Nennen Sie die Ihnen bekannten Skalentypen und diskutieren Sie, wie diese für Entscheidungsbäume genutzt werden können. **(6 Punkte)**

Wiederholung Vorlesung: Skalentypen

Skalentyp	Wertebereich	Mögliche Operationen	Beispiele
Nominale Größen	diskret, endlich	Gleichheit	Geschlecht Augenfarbe
Ordinale Größen	diskret, endlich, Ordnung	Gleichheit, größer / kleiner als	Prüfungsnoten Schulabschluß
Intervallgrößen	kontinuierlich bzw. ganzzahlig, unendlich	Gleichheit, größer / kleiner als Differenz	Datum Temperatur
Ratiogrößen	kontinuierlich bzw. ganzzahlig, unendlich	Gleichheit größer / kleiner als Differenz Verhältnis	Abstand Alter

Aufgabe 2 (d) – Lösung (Skalentypen)

- Problem bei Entscheidungsbäumen
Unterteilung der Daten in Gruppen bei Wahl der Splits...
... (je nach Algorithmus) durch Prüfung auf Ungleichheit zu einem Wert
- Lösung für die Skalentypen
 - Nominale und ordinale Größen
 - Bei wenigen Ausprägungen direkt nutzbar
 - Bei vielen Ausprägungen Kombination von ähnlichen Ausprägungen
 - Intervallgrößen und Ratiogrößen
 - Binning: Kombination in gleichgroße oder gleichhäufige Intervalle
 - Identifikation interessanter Wertebereiche
(z.B. Temperatur: Wasser koch, Wasser ist flüssig, Wasser ist gefroren)

Aufgabe 2 (e) - Aufgabenstellung

- Beschreiben Sie kurz die Grundidee von Biased Sampling und Stratified Sampling und geben Sie je ein Szenario an, für das die Sampling Methoden besonders geeignet sind. **(6 Punkte)**

Wiederholung Vorlesung: Stichprobe

- Idee
 - Algorithmen zur Datenanalyse oft ressourcenintensiv...
... mit dem Ziel „charakteristische“ Eigenschaften zu identifizieren
 - Ziehen einer repräsentativen Stichprobe
- Ansätze
 - Biased Sampling
 - Jüngere Beobachtungen werden eher für Stichprobe gezogen
 - Vorteile
 - Sample repräsentiert aktuellen Kunden, ... gut
 - Historische Daten werden nach und nach verdrängt
 - Stratified Sampling
 - Separierung der Daten in Klassen und ziehen für Klassen einzeln
 - Vorteile
 - Analysen auch auf Klassen mit wenigen Beobachtungen möglich (z.B. SPAM Klassifikation von E-Mails)
 - Einfach zu realisieren

Aufgabe 2 (e) – Lösung (Sampling)

- Biased Sampling
 - Beschreibung (siehe letzte Folie)
 - Besonders geeignet, wenn sich Beobachtungen über Zeit ändern
Beispiel: Vorhersage der Nutzung von Handytarifen
 - Vor 20 Jahren: Nur Telefon und SMS
 - Vor 10 Jahren: Primär Telefon und einzelne Kunden Daten
 - Heute: Erste Kunden telefonieren bereits primär über Datentarif
- Stratified Sampling
 - Beschreibung (siehe letzte Folie)
 - Anzahl der Beobachtungen pro Klasse wird angeglichen
Beispiel: Vorhersage von SPAM E-Mails
 - 95% des E-Mail Aufkommens ist SPAM
 - 05% des E-Mail Aufkommens ist relevantLernt Klassifikator auf unstratifizierten Daten...
... Erkennung relevanter E-Mails kaum möglich

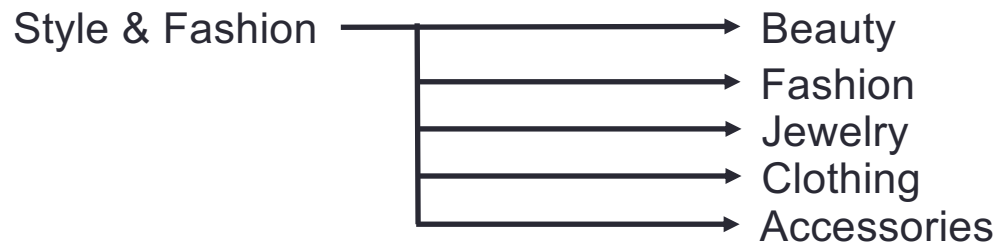
Aufgabe 2 (f) - Aufgabenstellung

- Beschreiben Sie Struktur und Nutzen von Taxonomien für die Datenqualität.
(4 Punkte)

Aufgabe 2 (f) – Lösung (Taxonomien)

- Idee
 - Aufbau von Bäumen zur Sicherstellung einer gemeinsamen Sprache
 - Obere Ebenen sind Abstraktionen der unteren Ebenen
 - Namen der einzelnen Ausprägungen sind standardisiert

- Beispiel (vgl. Vorlesung)



- Nutzen
 - Verschiedene Anbieter nutzen gleiche Taxonomie...
... Daten über Kunden gemäß Taxonomie austauschbar
 - Anreicherung der eigenen Daten mit fremden Daten möglich

Aufgabe 2 (g) - Aufgabenstellung

- Nennen Sie zwei andere Klassifikationsverfahren neben den Entscheidungsbäumen und diskutieren Sie kurz die Vorteile der Verfahren gegenüber Entscheidungsbäumen. **(4 Punkte)**

Aufgabe 2 (g) – Lösung (Klassifikationsverfahren)

- Neuronale Netzwerke
 - Unempfindlich bei verrauschten Daten
 - Auch für Regressionsprobleme geeignet
- Support Vector Machine
 - Unempfindlich bei verrauschten Daten
 - Auch für Regressionsprobleme geeignet
- Bayes Klassifikator
 - Mathematisch ableitbar
 - Kein Trainieren des Modells nötig

Aufgabe 2 (h) - Aufgabenstellung

- Nennen Sie vier verschiedene Stoppkriterien für Entscheidungsbäume.
(4 Punkte)

Aufgabe 2 (h) – Lösung (Stoppkriterien)

- Minimale Zahl von Beobachtungen pro Knoten
Split wird nur durchgeführt, wenn mehr als n Beobachtungen...
... durch den Knoten repräsentiert werden
- Minimaler Anteil falsch klassifizierter Beobachtungen pro Knoten
Split nur dann, wenn mind. $x\%$ der Beobachtungen im Knoten nicht...
... der Mehrheitsklasse angehören
- Maximale Baumtiefe
Split nur dann, wenn ein Knoten vorgegebene Tiefe t
... (Anzahl Knoten zur Wurzel) noch nicht erreicht hat
- Maximale Knotenanzahl pro Baum
Split nur dann, wenn Baum insgesamt weniger als n Knoten enthält