Big Data Anwendungen

Aufgabenblatt 3

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation

- Aufgabenblatt 3 – Recommender Systems
  - Aufgabe 1 – Association Rules
  - Aufgabe 2 – Recommender Systems
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 1 (a) - Aufgabenstellung

- Gegeben seien folgende Daten:

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Leiten Sie mit Hilfe des Apriori Algorithmus die Frequent Itemsets ab, die einen Support von mindestens 1/2 besitzen. **(10 Punkte)**

Wiederholung Vorlesung: A-Priori Algorithmus

- Finden aller Itemsets mit ausreichendem Support:

- Beginn mit einelementigen Sets (1)-Sets:
  - einfaches Abzählen

- Berechnung der k-Sets aus den (k-1)-Sets:
  - Join-Step: Ermittlung von Kandidaten;
    Aus A-Priori Eigenschaft:
    Alle (k-1)-elementigen Teilmengen
    eines k-Sets sind (k-1)-Sets,
  - Prune-Step: Löschen aller Kandidaten,
    die eine „unzulässige“
    (k-1)-elementige Teilmenge haben.
  - Support Counting, d. h. Abzählen,
    wie häufig die Kandidaten wirklich sind.

Aufgabe 1 (a) – Lösung (A-Priori Algorithmus)

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Vorgehen:
  - Ermittlung 1-elementige Sets
  - Ableiten der k-elementigen aus k-1
- Minimaler Support: $\frac{1}{2}$

  ⇒ Item ist Frequent, wenn drei oder mehr Vorkommen
- Ableiten der 1-elementigen Frequent Itemsets
  - Kandidaten

Frequent Itemset	Anzahl
Brötchen	3
Brezeln	3
Brot	4
Salat	6
Kuchen	3
Die Bild	2

  - Streichen der Kandidaten mit zu geringem Support

Aufgabe 1 (a) – Lösung (A-Priori Algorithmus)

- Bisher
  - Support: $\frac{1}{2}$
  - Item Sets:

1-elementige Sets	Anzahl
Brötchen	3
Brezeln	3
Brot	4
Salat	6
Kuchen	3

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Join Step: 2-elementige Kandidatenen
- Prune Step: keine Streichung
- Support Counting

Frequent Itemset	Anzahl
Brötchen, Brezeln	2
Brötchen, Brot	2
Brötchen, Salat	3
Brötchen, Kuchen	1
Brezeln, Brot	2
Brezeln, Salat	3
Brezeln, Kuchen	2
Brot, Salat	4
Brot, Kuchen	2
Salat, Kuchen	3

Aufgabe 1 (a) – Lösung (A-Priori Algorithmus)

- Bisher
  - Support: $\frac{1}{2}$
  - Item Sets:

Frequent Itemset	Anzahl
Brötchen, Salat	3
Brezeln, Salat	3
Brot, Salat	4
Salat, Kuchen	3

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Join Step:
  3-elementige Kandidatenen

- Prune Step:
  Alle streichen

Frequent Itemset	Anzahl
Brötchen, Brezeln, Salat	2
Brötchen, Brot, Salat	2
Brötchen, Salat, Kuchen	1
Brezeln, Brot, Salat	2
Brezeln, Salat, Kuchen	2
Brot, Salat, Kuchen	2

- Support Counting:
  Nicht mehr nötig

Aufgabe 1 (b) - Aufgabenstellung

- Gegeben seien folgende Daten:

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Identifizieren Sie auf der Basis der Frequent Itemsets aus (a) alle Association Rules mit minimaler Confidence von 3/4. **(5 Punkte)**

Wiederholung Vorlesung: Confidence

- Alternative Namen
  - Genauigkeit
  - „Überraschungsmass“
- Idee
  Wenn eine Transaktion X enthält, dann auch Y (mit gegebener Genauigkeit)
- Formal:
  $conf(X \rightarrow Y) = P(Y|X) = \frac{|X \cap Y|}{|X|} = \frac{supp(X \rightarrow Y)}{supp(X)}$
- Beispiel:
  - Wenn Windeln gekauft wurden, wurde in 100% aller Fälle auch Bier gekauft
  - Confidence = 100%
- Ziel: Finden von Regeln mit
  - … hohem Support (support > minSup) und …
  - … hoher Confidence (confidence > minConf)

Aufgabe 1 (b) – Lösung (A-Priori Algorithmus)

- Bisher
  - Identifikation Frequent Itemsets

Frequent Itemset
Brötchen, Salat
Brezeln, Salat
Brot, Salat
Salat, Kuchen

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Jetzt
  - Identifikation der Kandidaten
  - Ermitteln der Confidence

- Ermitteln der Kandidaten
- Ermittlung Confidence
- Streichen der Kandidaten mit Confidence < 3/4

Association Rule	Confidence
Brötchen ⇒ Salat	3/3
Brezeln ⇒ Salat	3/3
Brot ⇒ Salat	4/4
Kuchen ⇒ Salat	3/3
Salat ⇒ Brötchen	3/6
Salat ⇒ Brezeln	3/6
Salat ⇒ Brot	4/6
Salat ⇒ Kuchen	3/6

Aufgabe 1 (c) - Aufgabenstellung

- Gegeben seien folgende Daten:

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

- Diskutieren Sie die Unterschiede der in der Vorlesung besprochenen Algorithmen zur Identifikation von Frequent Itemsets. **(5 Punkte)**

Wiederholung Vorlesung: FP-Growth

- Nachteile A priori Algorithmus
  - Anzahl möglicher Kandidaten kann sehr groß sein (insbesondere die mit ein und zwei Elementen)
  - Hohe Anzahl von kompletten „Datenscans“ (für Big Data also nur bedingt geeignet)

- Idee
  - Stufe 1: Ableiten des FP-Trees (Ableiten häufiger Itemsets in einem Baum)
  - Stufe 2: Ableiten der Frequent Itemsets (unter Einsatz des Baums anstelle von „Datenscans“)

- Vorteil
  - Datenbank muss nur zwei Mal komplett durchlaufen werden
  - Algorithmus ist bedeutend schneller als Apriori (bei identischen Ergebnissen)

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation

- Aufgabenblatt 3 – Recommender Systems
  - Aufgabe 1 – Association Rules
  - Aufgabe 2 – Recommender Systems
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 2 (a) - Aufgabenstellung

- Gegeben seien folgende Bewertungen von Personen für Produkte:

Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Mittels Collaborative Filtering, soll Karl ein Produkt, das er noch nicht bewertet hat, empfohlen werden. Entscheiden Sie, ob dabei auf Bewertungen von Eugen oder August zurückgegriffen wird, wenn nur die ähnlichste andere Person Berücksichtigung findet. **(10 Punkte)**

Wiederholung Vorlesung: Collaborative Filtering

- Anwendung des Pearson Korrelationsmaßes
  - $Pearson(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{S}(x_i - \hat{x}) \cdot (y_i - \hat{y})}{\sqrt{\sum_{i=1}^{S}(x_i - \hat{x})^2} \cdot \sqrt{\sum_{i=1}^{S}(y_i - \hat{y})^2}}$

    mit $\bar{X} = (x_1 \dots x_n)$ bzw. $\bar{Y} = (y_1 \dots y_n)$ sind Vektoren der User...
    ... und $\hat{x} = \sum_{i=1}^{S} \frac{x_i}{S}$ bzw $\hat{y} = \sum_{i=1}^{S} \frac{y_i}{S}$ die Mittelwerte der Vektoren

- Pearson Korrelationsmaß muss für betrachteten User...
  ... mit allen anderen Usern ermittelt werden

- Top k User mit höchstem Index bei Vergleich mit betrachteten User...
  ... bilden Grundlage für Empfehlungen

- Nachteil des Verfahrens
  - Bewertungen der Nutzer oft nicht gleich verteilt
    (Pessimisten vs. Optimisten)
  - Anpassung der Skalen notwendig

Aufgabe 2 (a) – Lösung (Collaborative Filtering)

Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Betrachtung Karl / Eugen
- Ermittlung $\overline{x}$, $\overline{y}$
  - $\overline{x} = \frac{3+1+4}{3} = \frac{8}{3}$
  - $\overline{y} = \frac{1+1+2}{3} = \frac{4}{3}$
- $\text{Pearson}_{\text{Karl,Eugen}} = \frac{\sum(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2}\cdot\sqrt{\sum(y_i-\overline{y})^2}}$

$$= \frac{\left(3-\frac{8}{3}\right)\left(1-\frac{4}{3}\right)+\left(1-\frac{8}{3}\right)\left(1-\frac{4}{3}\right)+\left(4-\frac{8}{3}\right)\left(2-\frac{4}{3}\right)}{\sqrt{\left(3-\frac{8}{3}\right)^2+\left(1-\frac{8}{3}\right)^2+\left(4-\frac{8}{3}\right)^2}\sqrt{\left(1-\frac{4}{3}\right)^2+\left(1-\frac{4}{3}\right)^2+\left(2-\frac{4}{3}\right)^2}}$$

$$= \frac{\frac{1}{3}\left(-\frac{1}{3}\right)+\left(-\frac{5}{3}\right)\left(-\frac{1}{3}\right)+\left(-\frac{4}{3}\right)\left(-\frac{2}{3}\right)}{\sqrt{\left(-\frac{1}{3}\right)^2+\left(-\frac{5}{3}\right)^2+\left(-\frac{4}{3}\right)^2}\sqrt{\left(-\frac{1}{3}\right)^2+\left(-\frac{1}{3}\right)^2+\left(-\frac{2}{3}\right)^2}}$$

$$= \frac{-\frac{1}{9}+\frac{5}{9}+\frac{8}{9}}{\sqrt{\frac{42}{9}}\sqrt{\frac{6}{9}}} = \frac{\frac{12}{9}}{\sqrt{\frac{28}{9}}} = \frac{\frac{4}{3}}{\sqrt{\frac{28}{9}}} \approx 0{,}756$$

Aufgabe 2 (a) – Lösung (Collaborative Filtering)

- Betrachtung Karl / August

Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Ermittlung $\overline{x}$, $\overline{y}$
  - $\overline{x} = \frac{3+1+4}{3} = \frac{8}{3}$
  - $\overline{y} = \frac{5+4+5}{3} = \frac{14}{3}$

- $\text{Pearson}_{\text{Karl,August}} = \frac{\sum(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2}\cdot\sqrt{\sum(y_i-\overline{y})^2}}$

$$= \frac{\left(3-\frac{8}{3}\right)\left(5-\frac{14}{3}\right)+\left(1-\frac{8}{3}\right)\left(4-\frac{14}{3}\right)+\left(4-\frac{8}{3}\right)\left(5-\frac{14}{3}\right)}{\sqrt{\left(3-\frac{8}{3}\right)^2+\left(1-\frac{8}{3}\right)^2+\left(4-\frac{8}{3}\right)^2}\sqrt{\left(5-\frac{14}{3}\right)^2+\left(4-\frac{14}{3}\right)^2+\left(5-\frac{14}{3}\right)^2}}$$

$$= \frac{\frac{1}{3}\cdot\frac{1}{3}+\left(-\frac{5}{3}\right)\left(-\frac{2}{3}\right)+\frac{4}{3}\cdot\frac{1}{3}}{\sqrt{\left(\frac{1}{3}\right)^2+\left(-\frac{5}{3}\right)^2+\left(\frac{4}{3}\right)^2}\sqrt{\left(\frac{1}{3}\right)^2+\left(-\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2}}$$

$$= \frac{\frac{1}{9}+\frac{10}{9}+\frac{4}{9}}{\sqrt{\frac{42}{9}}\sqrt{\frac{6}{9}}} = \frac{\frac{15}{9}}{\sqrt{\frac{28}{9}}} = \frac{\frac{5}{3}}{\sqrt{\frac{28}{9}}} = 0{,}945$$

Aufgabe 2 (a) – Lösung (Collaborative Filtering)

- Bisher
  - $Pearson_{Karl,Eugen} = 0{,}756$
  - $Pearson_{Karl,August} = 0{,}945$

Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Ergebnis
  - $Pearson_{Karl,August} > Pearson_{Karl,Eugen}$
    ⇒ Empfehlung von Produkten auf Basis von Augusts Konsum

- Addendum: Welche Artikel empfehlen?
  - Bei den gegebenen Daten würde 156921 empfohlen

Aufgabe 2 (b) - Aufgabenstellung

- Gegeben seien folgende Bewertungen von Personen für Produkte:

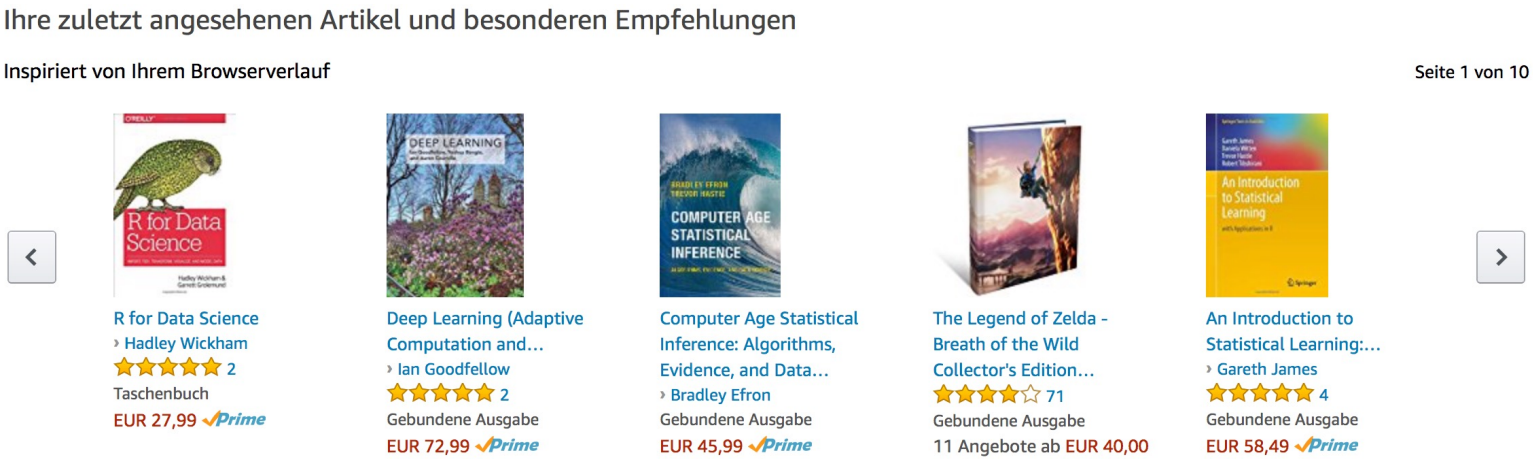
Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Diskutieren Sie wofür Collaborative Filtering Algorithmen in der Praxis eingesetzt werden können. **(3 Punkte)**

Wiederholung Vorlesung: Recommender Systems

- Assoziationsregeln revisited
  - Identifikation von Implikationen: „Wer A kauft, kauft auch B“
  - Assoziationsregeln entgegen aktuellem Trend in Big Data: Finden Regeln für alle, statt individuelle Empfehlungen (Generalisierung statt Individualisierung)

- Modernerer Ansatz: Recommender Systems
  - Identifikation ähnlicher Kunden
  - Empfehlung von Artikeln auf Basis ähnlicher Kunden



Aufgabe 2 (b) – Lösung (Einsatzgebiete)

- Einsatz von Collaborative Filtering
  - Ableiten von Kundenverhalten auf Basis von anderen Kunden
  - Ableiten von Zielgruppen für Artikel

- Bekannte Beispiele
  - Amazon „Ähnliche Kunden kauften auch“,
  - Spotify/Netflix „Angebot weiterer Musiktitel“

Aufgabe 2 (c) - Aufgabenstellung

- Gegeben seien folgende Bewertungen von Personen für Produkte:

Person	123456	214121	938123	102311	192312	156921
Karl	3	1	1	4		
Eugen	1		1	2	5	
August	5	4		5		5

- Erläutern Sie kurz den Unterschied zwischen Collaborative Filtering und Content-Based Recommendations. **(2 Punkte)**

Aufgabe 2 (c) – Lösung (Abgrenzung)

- Collaborative Filtering
  - Beobachtung des Verhaltens einzelner Nutzer
  - Identifikation ähnlichen Verhaltens von Nutzern
  - Empfehlungen auf Basis ähnlichem Verhalten

- Content-Based Recommendations
  - Identifikation von Eigenschaften der konsumierten Produkte
  - Ableiten von Nutzerprofilen auf Basis der konsumierten Produkte
  - Empfehlung von Produkten auf Basis deren Eigenschaften

- Unterschied
  - Fokus auf ähnlichen Kunden vs. ähnlichen Artikeleigenschaften
  - Content-Based Recommendations hilft bei guten Artikeleigenschaften
  - Collaborative Filtering bei vielen ähnlichen Kunden

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation

- Aufgabenblatt 3 – Recommender Systems
  - Aufgabe 1 – Association Rules
  - Aufgabe 2 – Recommender Systems
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 3 (a) - Aufgabenstellung

- Geben Sie an, wie Assoziation Rules helfen können, um Prospekte für Supermärkte gewinnbringend aufzubauen. **(5 Punkte)**

Aufgabe 3 (a) – Lösung (Abgrenzung)

- Verschiedene Möglichkeiten
  - Anbieten des Antecedent im Prospekt…
    … Consequent im Markt (überteuert) neben Antecedent
  - Verkauf von Antecedent und Consequent (nur) im Bundel
  - Positionierung von Antecedent und Consequent nahe beeinander (Steigerung der Kundenzufriedenheit)
  - Positionierung von Antecedent und Consequent in unterschiedlichen Bereichen
    (Steigerung des Umsatzes – Kunde muss beide Bereiche besuchen)

- Anmerkung
  - Hohe Zahl von Möglichkeiten (ohne konkrete Empfehlung)…
    … ist zentrale Schwäche des Verfahrens

Aufgabe 3 (b) - Aufgabenstellung

- Nennen Sie die zwei unterschiedlichen Ansätze für Recommender Systeme und diskutieren Sie deren Vor- und Nachteile. **(5 Punkte)**

Aufgabe 3 (b) – Lösung (Abgrenzung)

- Assoziation Rules
  - Inhalt des Warenkorbs wird berücksichtigt
  - Effiziente Umsetzung möglich (FP Growth)

- Content Based Filtering
  - Eigenschaften der Nutzerpräferenzen werden berücksichtigt
  - Benutzerverhalten muss nicht erhoben werden

- Collaborative Filtering
  - Vergleich der Ähnlichkeit von Nutzern über Eigenschaften des Konsums
  - Beobachtungen insbesondere bei langer Kundenbeziehung interessant

Aufgabe 3 (c) - Aufgabenstellung

- Erläutern Sie, wie Association Rules für die Lösung von Klassifikations-problemen genutzt werden können. **(5 Punkte)**

Vorlesung: Assoziationsregeln zur Klassifikation

- Idee
  - Antecedents sind Eingangsparameter
  - Consequent ist vorhergesagte Klasse

- Eine Regel passt:
  $\Rightarrow$ Klassifikation eindeutig (mit Konfidenz der Regel)

- Keine Regel passt:
  $\Rightarrow$ Mehrheitsklasse bzw. unklassifiziert

- Mehrere Regeln passen:
  - Berücksichtigung der Regel mit höchster Konfidenz
    - Regel entscheidet
  - Berücksichtigung der k Regeln mit höchster Konfidenz (oder auch aller Regeln)
    - Häufigste auftretende Klasse
    - Klasse mit höchster durchschnittlicher Konfidenz der Regeln
  - …