Multivariate Verfahren

Assoziationsregeln

Annika Hoyer

Sommersemester 2020

Assoziationsregeln - Inhalt

AssoziationsregeIn

Problemstellung

- Betrachte Merkmalsvektor $\mathbf{X} = (X_1, \ldots, X_p)^\top$ und finde die Werte/Kombinationen der Variablen, die in den Daten am häufigsten vorkommen
- Beispiel: Warenkorb-Analyse
  - Artikel, die Kunden eines Supermarktes zusammen einkaufen
  - ”In 45% der Fälle, in denen Lachs gekauft wird, wird auch Weißwein gekauft. Diese beiden Produkte kommen in 2% aller Transaktionen vor.”
- Typische Anwendungsgebiete:
  - Einzel- und Versandhandel
  - Tourismus

Problemstellung

Motivation als ”Market Basket Analysis”:

Die Beobachtungen sind Kauftransaktionen. Die Variablen sind binär kodiert, $X_j \in \{0, 1\}$, und entsprechen allen Items, die verkauft werden. Für jede Beobachtung i gilt:

$$x_{ij} = \begin{cases} 1, & \text{falls das } j\text{-te Item gekauft wurde,} \\ 0, & \text{falls das } j\text{-te Item nicht gekauft wurde.} \end{cases}$$

Problemstellung

Für Assoziationsanalysen sind folgende Parameter relevant:

- Konfidenz der Regel, d.h. Stärke der Korrelation (*in 45% der Fälle*)
- Support der Regel, d.h. Häufigkeit des gemeinsamen Auftretens (*in 2% aller Transaktionen*)

Generellere Formulierung

Finde eine Menge an prototypischen Werten $\nu_1, \ldots, \nu_L$ des Merkmalvektores $\mathbf{X}$, so dass die Wahrscheinlichkeit $P(\nu_\ell)$ groß ist.

- Einfacher Schätzer: Anteil der Beobachtungen, für die $\mathbf{X} = \nu_\ell$
- Insbesondere für kleine Anzahl an Merkmalen schwierig
- Verwendet werden für dieses Problem die Begriffe ”mode finding” oder ”bump hunting”

Modifizierte Zielsetzung

- Gegeben: Merkmale $X_1, \ldots, X_p$ mit zugehörigen Wertebereichen $S_1, \ldots, S_p$
- Idee: Finde Teilmengen $s_j \subseteq S_j$, so dass die Wahrscheinlichkeit

$$P\left(\bigcap_{j=1}^{p}(X_j \in s_j)\right)$$

  groß ist
- Mögliche Vereinfachungen:
  - Betrachte nur einzelne Werte, d.h. $s_j = \nu_{0j}$.
  - Betrachte die Menge aller Werte, d.h. $s_j = S_j$.

Beispiel: Supermarkt

Einkaufs-Transaktion	gekaufte Artikel (Items)
t_1	Saft, Cola, Bier
t_2	Saft, Cola, Wein
t_3	Saft, Wasser
t_4	Cola, Bier, Saft
t_5	Saft, Cola, Bier, Wein
t_6	Wasser

Frage: Welche Assoziationen lassen sich aus diesen Daten ableiten?
Beispiel: Saft $\rightarrow$ Cola

Dummy-Variablen

Umstrukturierung des Datensatzes:

- Betrachte binäre Zufallsvariablen $Z_1, \ldots, Z_K$, die für jeden Wert $\nu_{\ell j}$ aus den Merkmalen $X_1, \ldots, X_p$ gebildet werden
- Anzahl der Dummy-Variablen berechnet sich als

$$K = \sum_{j=1}^{p} |S_j| \,,$$

  wobei $|S_j|$ der Anzahl an unterschiedlichen Werten von X_j entspricht
- Für die binären Variablen gilt:

$$Z_k = \begin{cases} 1, & \text{falls Merkmal entsprechenden Wert annimmt,} \\ 0, & \text{sonst.} \end{cases}$$

Beispiel: Supermarkt

Transaktion	Z_1=Saft	Z_2=Cola	Z_3=Bier	Z_4=Wein	Z_5=Wasser
t_1	1	1	1	0	0
t_2	1	1	0	1	0
t_3	1	0	0	0	1
t_4	1	1	1	0	0
t_5	1	1	1	1	0
t_6	0	0	0	0	1

Dummy-Variablen

- Finde die Itemmenge $\mathcal{K} \subset \{1, \ldots, K\}$, so dass

$$P\left(\bigcap_{k \in \mathcal{K}} (Z_k = 1)\right) = P\left(\prod_{k \in \mathcal{K}} (Z_k = 1)\right)$$

  groß ist
- Schätzer (mit Daten z_{ik}, $i = 1, \ldots, n$):

$$T(\mathcal{K}) = \hat{P}\left(\prod_{k \in \mathcal{K}} (Z_k = 1)\right) = \frac{1}{n} \sum_{i=1}^{n} \prod_{k \in \mathcal{K}} z_{ik}$$

- $T(\mathcal{K})$ wird als Support der Itemmenge $\mathcal{K}$ bezeichnet

Assoziationsanalyse

Assoziationsanalyse

- Spezifiziere untere Grenze t, die der Support mindestens annehmen soll
- Bestimme aus Variablen $Z_1, \ldots, Z_K$ alle Itemmengen $\mathcal{K}_s$, für die gilt

$$T(\mathcal{K}_s) > t$$

- Beachte: Um eine rechnerisch machbare Lösung zu erhalten, sollte t so gewählt werden, dass die Menge $\{\mathcal{K}_s | T(\mathcal{K}_s) > t\}$ im Vergleich zur Zahl aller möglichen Itemmengen (2^K) relativ klein ist

Bildung von Assoziationsregeln

- Zerlegung der Variablen Z_k, $k \in \mathcal{K}$ in zwei disjunkte Teilmengen $A \cup B = \mathcal{K}$
- Es gilt:

$$A \Rightarrow B$$

- Bestimmung der Eigenschaften von Assoziationsregeln durch:
  - Support: $T(A \Rightarrow B)$
  - Confidence: $C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$
  - Lift: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}$

Eigenschaften von Assoziationsregeln

Support

$$T(A \Rightarrow B) = \hat{P}(A \cap B)$$

$\rightarrow$ relative Häufigkeit der Beispiele, in denen die Regel anwendbar ist

Confidence

$$C(A \Rightarrow B) = \hat{P}(B|A)$$

$\rightarrow$ relative Häufigkeit der Beispiele, in denen die Regel richtig ist

Eigenschaften von Assoziationsregeln

Lift

$$L(A \Rightarrow B) = \frac{\hat{P}(B|A)}{\hat{P}(B)}$$

$\rightarrow$ Wie hoch übertrifft der Konfidenzwert der Regel den Erwartungswert?

- $L(A \Rightarrow B) > 1 \rightarrow A$ und B sind positiv korreliert
- $L(A \Rightarrow B) < 1 \rightarrow A$ und B sind negativ korreliert
- $L(A \Rightarrow B) = 1 \rightarrow A$ und B sind unabhängig

Beispiel: Supermarkt

Transaktion	Z_1=Saft	Z_2=Cola	Z_3=Bier	Z_4=Wein	Z_5=Wasser
t_1	1	1	1	0	0
t_2	1	1	0	1	0
t_3	1	0	0	0	1
t_4	1	1	1	0	0
t_5	1	1	1	1	0
t_6	0	0	0	0	1

Betrachte die Regel Saft (A) $\Rightarrow$ Cola (B).

- $T(A \Rightarrow B) = \hat{P}(A \cap B) = 4/6 = 0.67$
- $C(A \Rightarrow B) = \hat{P}(B|A) = \frac{\hat{P}(A \cap B)}{P(A)} = \frac{4/6}{5/6} = 0.8$
- $L(A \Rightarrow B) = \frac{\hat{P}(B|A)}{\hat{P}(B)} = \frac{4/5}{4/6} = 1.2$

Der Apriori-Algorithmus

Grundidee

Nutze die Beziehung zwischen Teilmengen der Itemmengen:

$$\mathcal{K}_s \subset \mathcal{K}_\ell \Rightarrow T(\mathcal{K}_s) \geq T(\mathcal{K}_\ell)$$

1. Finde im ersten Schritt alle Itemmengen $\mathcal{K}_{s_1}$, die ein Element enthalten, mit $T(\mathcal{K}_{s_1}) > t$.
2. Finde im ν-ten Schritt alle Itemmengen $\mathcal{K}_{s_\nu}$, die ν Elemente enthalten, mit $T(\mathcal{K}_{s_\nu}) > t$. Hierbei sind nur Obermengen von $\mathcal{K}_{s_{\nu-1}}$ zu prüfen.
3. Wiederhole Schritt 2 bis keine Itemmenge mehr hinzukommt.

Beispiel: Apriori-Algorithmus mit $t = 50\%$

n	Itemmenge	Items mit Support $\geq 50\%$
1	{{Saft}, {Cola}, {Bier}, {Wein}, {Wasser}}	{{Saft}, {Cola}, {Bier}}
2	{{Saft, Cola}, {Saft, Bier}, {Cola, Bier}}	{{Saft, Cola}, {Saft, Bier}, {Cola, Bier}}
3	{{Saft, Cola, Bier}}	{{Saft, Cola, Bier}}
4	{}	{}

Erzeugung von Regeln mit minimaler Confidence von 75% am Beispiel {Saft, Cola}:

$$C(\text{Saft} \Rightarrow \text{Cola}) = \frac{2/3}{5/6} = 0.8$$

$$C(\text{Cola} \Rightarrow \text{Saft}) = \frac{2/3}{2/3} = 1.0$$

Beispiel: Apriori-Algorithmus mit $t = 50\%$

Regeln mit Support $\geq 50\%$	Support	Confidence
Saft $\Rightarrow$ Cola	66%	80%
Cola $\Rightarrow$ Saft	66%	100%
Cola $\Rightarrow$ Bier	50%	75%
Bier $\Rightarrow$ Cola	50%	100%
Saft $\Rightarrow$ Bier	50%	60%
Bier $\Rightarrow$ Saft	50%	100%
Saft, Bier $\Rightarrow$ Cola	50%	100%
Cola, Saft $\Rightarrow$ Bier	50%	75%
Bier, Cola $\Rightarrow$ Saft	50%	100%
Cola $\Rightarrow$ Saft, Bier	50%	75%
Bier $\Rightarrow$ Cola, Saft	50%	100%
Saft $\Rightarrow$ Cola, Bier	50%	60%

Praktische Vorgehensweise

- Ziel: Assoziationsregeln identifizieren, die hohe Werte für Support und Confidence aufweisen
- Apriori-Algorithmus liefert Itemmengen mit hohem Support (in Abhängigkeit von t)
- Definiere untere Grenze c und bestimme daraus die Menge der Assoziationsregeln

$$\{A \Rightarrow B | C(A \Rightarrow B) > c\}.$$

- Ergebnis ist Menge von Assoziationsregeln, die die Bedingungen $T(A \Rightarrow B) > t$ und $C(A \Rightarrow B) > c$ erfüllen