BAYESIAN NETWORKS

APPLICATIONS FOR CUSTOMER ANALYTICS

DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

APRIL 2024

ANA ALINA TUDORAN (AAT)
ASSOCIATE PROFESSOR

AGENDA

Applications of Bayesian nets in customer analytics

a. Customer retention
b. Targeted advertising
c. Product recommendation

CUSTOMER RETENTION

Gupta, S. and Kim, H.W. (2008). Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities, *European Journal of Operational Research*, 190, 818-833

ENHANCING CUSTOMER RETENTION

- **Cost benefits:** lowered costs and higher profits from repeat purchase
- **Online challenges:** difficult to foster customer retention in an online setting
- **Value of Virtual Communities** (VC): great value to online firms
- **Mechanism of of customer commitment** formation in a VC helps to retain the customers

THEORETICAL MODEL

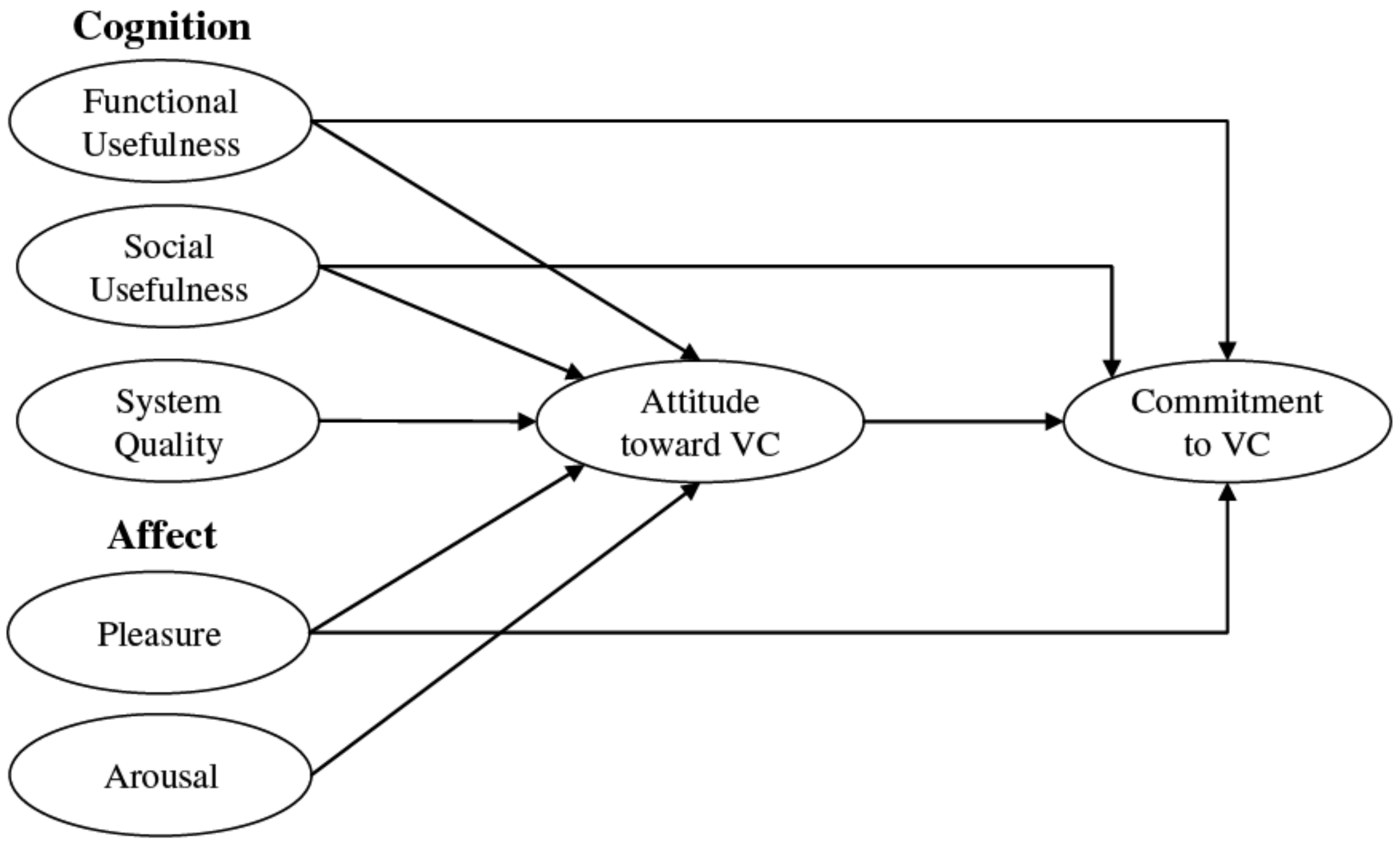


Fig. 1. A theoretical model for customer retention in a VC.

Source: Gupta and Kim (2008)

SEM RESULTS

Table 1
Factor determinacy results

	CFR	AVE
Functional usefulness	0.82	0.61
Social usefulness	0.89	0.72
System quality	0.89	0.68
Pleasure	0.86	0.68
Arousal	0.76	0.51
Attitude	0.89	0.68
Commitment	0.85	0.66

Table 2
Results of hypothesis testing using LISREL (SEM)

Dependent variable	Independent variable	Standard β	R^2
Commitment to VC	Attitude	0.28**	0.32
	Functional usefulness	0.22**	
	Social usefulness	ns	
	System quality	–	
	Pleasure	0.21*	
	Arousal	–	
Attitude toward VC	Functional usefulness	0.27***	0.55
	Social usefulness	ns	
	System quality	0.13*	
	Pleasure	0.42***	
	Arousal	ns	

AARHUS BSS

A BAYESIAN NET MODEL

(i) Functional usefulness $= f_1(u_1)$,
(ii) Pleasure $= f_2(u_2)$,
(iii) Attitude $= f_3$(Functional usefulness, Pleasure, u_3), and
(iv) Commitment $= f_4$(Functional usefulness, Pleasure, Attitude, u_4).

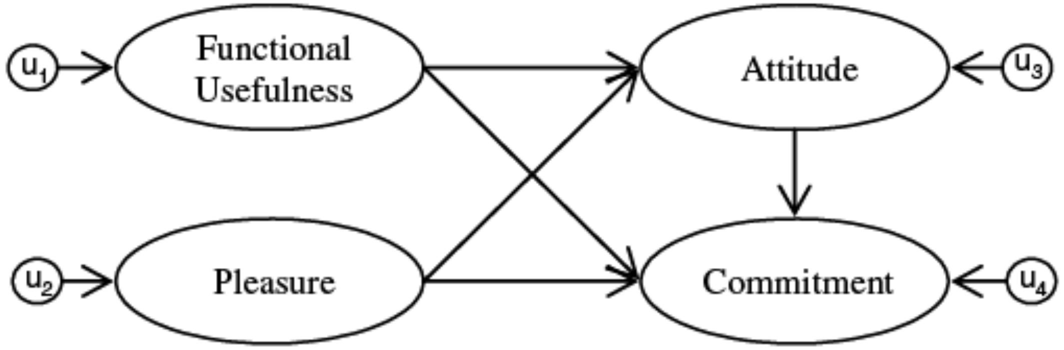


Fig. 2. A Bayesian network of customer retention in an online store using a VC. PLEA: pleasure, FUSE: functional usefulness, ATTI: attitude, COMM: commitment.

Source: Gupta and Kim (2008)

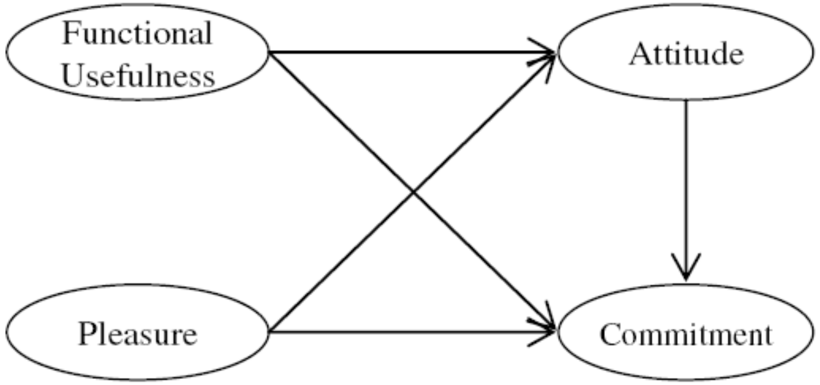
BUILDING AND VALIDATING A BAYES NET

- Structure learning
- Parameter learning

PARAMETERS

Functional Usefulness		
p(Low)	*p*(Med)	*p*(High)
0.02	0.26	0.72

FUSE	PLEA	Attitude		
		p (Low)	*p* (Med)	*p* (High)
Low	Low	0.99	0.00	0.00
Low	Med	0.00	0.67	0.33
Low	High	0.00	0.99	0.00
Med	Low	0.33	0.33	0.33
Med	Med	0.00	0.79	0.21
Med	High	0.00	0.40	0.60
High	Low	0.99	0.00	0.00
High	Med	0.00	0.47	0.53
High	High	0.00	0.09	0.91



Pleasure		
p(Low)	*p*(Med)	*p*(High)
0.01	0.55	0.44

FUSE	PLEA	ATTI	Commitment		
			p (Low)	*p* (Med)	*p* (High)
Low	Low	Low	0.00	1.00	0.00
Low	Low	Med	0.33	0.33	0.33
Low	Low	High	0.33	0.33	0.33
Low	Med	Low	0.33	0.33	0.33
Low	Med	Med	0.00	1.00	0.00
Low	Med	High	1.00	0.00	0.00
Low	High	Low	0.33	0.33	0.33
Low	High	Med	0.00	1.00	0.00
Low	High	High	0.33	0.33	0.33
Med	Low	Low	0.33	0.33	0.33
Med	Low	Med	0.33	0.33	0.33
Med	Low	High	0.33	0.33	0.33
Med	Med	Low	0.33	0.33	0.33
Med	Med	Med	0.00	0.98	0.02
Med	Med	High	0.00	0.83	0.17
Med	High	Low	0.33	0.33	0.33
Med	High	Med	0.00	0.33	0.67
Med	High	High	0.00	0.44	0.56
High	Low	Low	1.00	0.00	0.00
High	Low	Med	0.33	0.33	0.33
High	Low	High	0.33	0.33	0.33
High	Med	Low	0.33	0.33	0.33
High	Med	Med	0.00	0.84	0.16
High	Med	High	0.00	0.71	0.29
High	High	Low	0.33	0.33	0.33
High	High	Med	0.00	0.40	0.60
High	High	High	0.00	0.10	0.90

Source: Gupta and Kim (2008)

AARHUS BSS

USING A BN AS DECISION SUPPORT TOOL

- Forward inference
- Backward inference

FORWARD INFERENCE - PREDICTION

Table 5
Forward inference due to change in different states of functional usefulness

State	FUSE		ATTI		COMM	
	PCP	NCP	PCP	NCP	PCP	NCP
Low	**0.02**	**1.00**	0.01	0.01	0.01	0.18
Medium	0.26	0.00	0.39	0.81	0.54	0.81
High	0.72	0.00	0.60	0.18	0.45	0.00

ATTI: attitude, FUSE: functional usefulness, COMM: commitment, PCP: prior conditional probability, NCP: new conditional probability.

Source: Gupta and Kim (2008).

Assume a person joins a VC having a low functional usefulness perception. This information can be fed to the network as evidence to predict the conditional probability of the attitude and commitment.

BACKWARD INFERENCE - DIAGNOSTIC

Table 8
Backward inference of low commitment on three variables

State	Variables							
	COMM		ATTI		FUSE		PLEA	
	PCP	NCP	PCP	NCP	PCP	NCP	PCP	NCP
Low	**0.01**	**1.00**	0.01	0.59	0.02	0.36	0.01	0.64
Medium	0.54	0.00	0.39	0.02	0.26	0.07	0.55	0.36
High	0.45	0.00	0.60	0.39	0.72	0.57	0.44	0.00

COMM: commitment, ATTI: attitude, FUSE: functional usefulness, PLEA: pleasure, PCP: prior conditional probability, NCP: new conditional probability.

Source: Gupta and Kim (2008).

Assume that the online vendor observes decreasing commitment towards participation among its customers. This information can be fed to the network as evidence to diagnose the conditional probability of the attitude, functional usefulness and pleasure.

AARHUS BSS

CONCLUSION

- This application combines SEM and BN
- It considers the process from identification of causal relationships based on SEM
- It builts a Bayes net and uses it as a decision support tool for customer retention
- It presents how to support decision making regarding customer retention with prediction and diagnosis.

AARHUS
BSS

TARGETED ADVERTISING

Chapter 12: Targeted Advertising. In: Neapolitan & Jiang: Probabilistic methods for financial and marketing informatics, pp.373-382. San Francisco: Morgan Kaufmann

BACKGROUND

- One way to advertise a product is to simply try to reach as many potential customers as possible
- This method could be very costly for businesses because many of the individuals:
  - Will not buy the advertised product
  - Will be offended by receiving an unwanted advertisement or call, or
  - Will always buy
- The idea is to use BN to identify segments of customers that will most likely purchase when sending the ad (persuadable segments) and avoid sending the ad to the rest.

OBJECTIVE

- Our objective is to identify populations with positive Expected Lift in Profit P and send the ad only to those populations.

- To do so, for any given population Y, we need:

$$p(Buy_{Mailed}|Pop.Y)$$
$$p(Buy_{NOTMailed}|Pop.Y)$$

- These probabilities can be extracted from a BN trained on a historical data. We just need:
  - Target variable: Buying (Yes; No)
  - Indicator variable: Mailed (Yes; No)
  - The variables identifying the population of interest

AARHUS BSS

COST OF AD AND REVENUE

c - the **cost** of mailing the advertisement to a given person

r_u - the **income** obtained from a sale to an **unsolicited** customer

r_s - the **income** obtained from a sale to an **solicited** customer

$r_u \neq r_s$ because we may offer some discount in our ad

EXPECTED LIFT IN PROFIT (ELP)

We mail an ad to a population Y if:

$$EP_{Mail} = p(Buy_{Mailed}|Pop.Y) \cdot r_s - c > 0$$

… in addition, because some will buy anyway, we mail an ad if and only if:

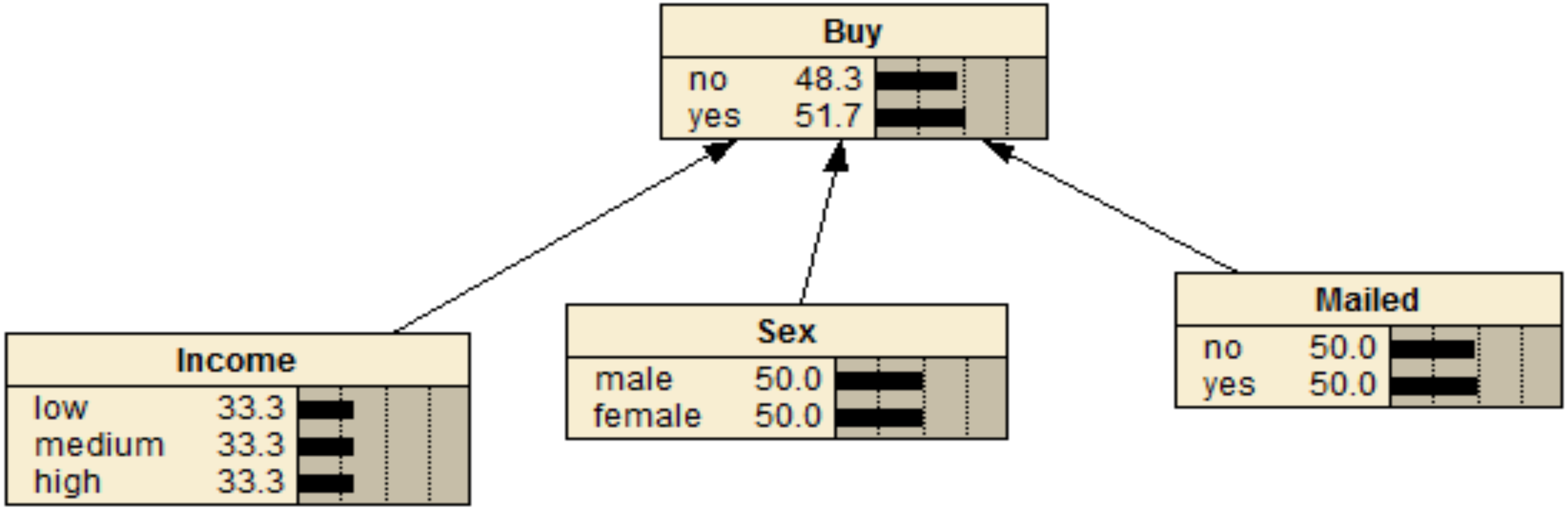
$$EP_{Mailed} > EP_{NOTMailed}$$

$$ELP : EP_{Mailed} - EP_{NOTMailed} > 0$$

$$p(Buy_{Mailed}|Pop.Y) \cdot r_s - p(Buy_{NOTMailed}|Pop.Y) \cdot r_u - cost > 0$$

AARHUS BSS

EXAMPLE



The figure below is a BN for targeted advertising.
The structure and parameters are based on the Class Probability Tree (p. 389 Neapolitan and Jiang)

EXAMPLE (CONT).

$$c = 0.5$$

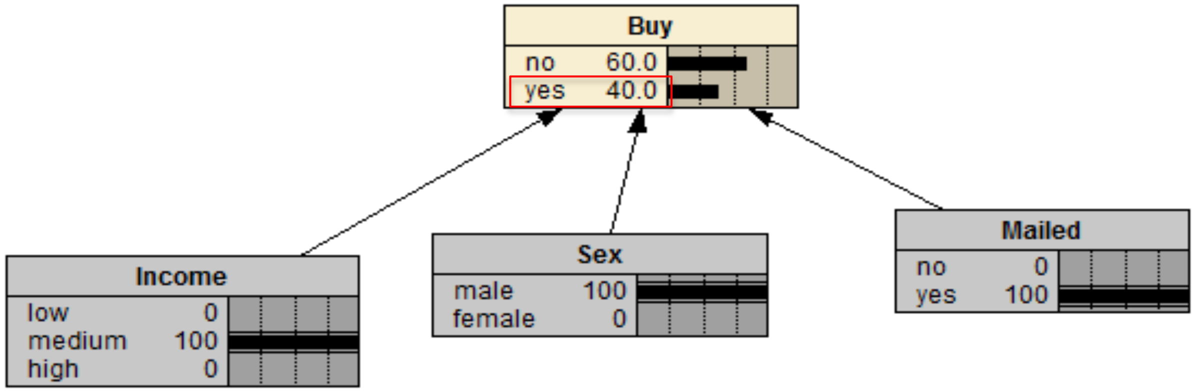
$$r_s = 8 \ (income \ with \ some \ discount \ is \ offered)$$

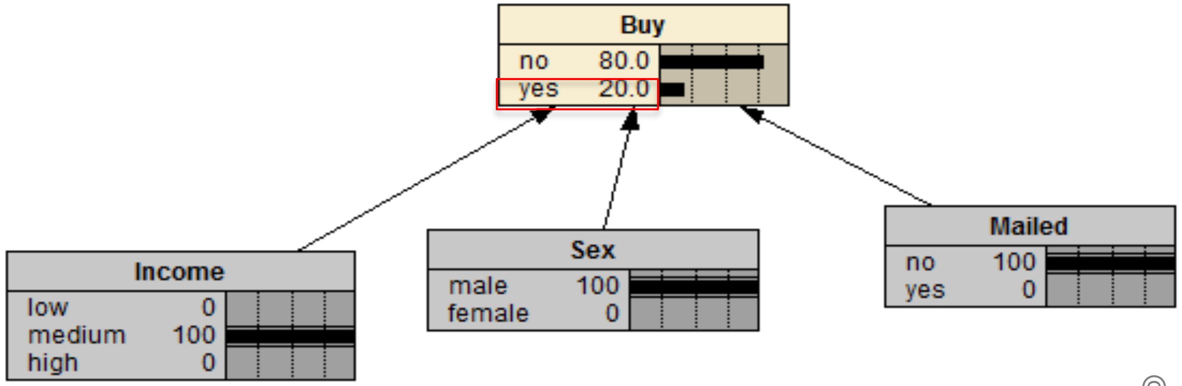
$$r_u = 10$$

Compute the ELP for the population consisting of individuals with
- medium income
- who are male

Should we mail an ad to this population?

EXAMPLE (CONT).





$$ELP = P(Buy = yes \mid Mailed = yes) r_s - P(Buy = yes \mid Mailed = no) r_u - c =$$
$$= 0.4 \times 8 - 0.2 \times 10 - 0.5 = 0.7$$

Since the ELP is positive, the recommendation is to mail to this population.

EXAMPLE (CONT).

$$c = 0.6$$
$$r_s = 7$$
$$r_u = 9$$

Compute the ELP for the population consisting of individuals with
- medium income
- who are female

Should we mail to this population?

$$ELP = P(Buy = yes \mid Mailed = yes) r_s - P(Buy = yes \mid Mailed = no) r_u - c =$$
$$= 0.7 \times 7 - 0.4 \times 9 - 0.6 = 0.7$$

Since the ELP is positive, we mail to this population.

EXAMPLE (CONT).

$c = 0.6$

$r_s = 7$

$r_u = 9$

Compute the ELP for the population consisting of individuals with
- low income.
Should we mail to this population?

$$ELP = P(Buy = yes \mid Mailed = yes) r_s - P(Buy = yes \mid Mailed = no) r_u - c =$$
$$= 0.6 \times 7 - 0.5 \times 9 - 0.6 = -1.8$$

Since the ELP is negative, we should not mail to this population.

CONCLUSION

- Using BN in this application allows to identify persuadable segments of individuals who would buy only if they are sent an ad.

- It avoids sending ads to those who:
- will never buy
- those who always buy (thus avoid wasting the ad), and
- those who are turned off by the advertisement when they receive it.

PRODUCT RECOMMENDATION

Chapter 11: Collaborative Filtering. In: Neapolitan & Jiang: Probabilistic methods for financial and marketing informatics, pp.373-382. San Francisco: Morgan Kaufmann

BACKGROUND

- **Collaborative Filtering** = the process of recommending items of interest to an individual, based on his/her interests or the interests of similar individuals.

- It is particularly effective for online stores

© Slides by AAT

AARHUS BSS

DATA TYPE

- **Explicit voting:**

  It learns individual´s preferences from individuals' **reported preferences** by asking the individuals explicitly to rank the items on some scale

- **Implicit voting:**

  It learns individual's preferences from individuals' past **behaviour**

EXPLICIT VOTING

- A data set with individuals ranking 4 products (X, Y, Z, W) based on their preferences on a scale from 1 to 5.

Person	X	Y	Z	W
Gloria	1	4	5	4
Juan	5	1	1	2
Amin	4	1	2	1
Sam	2	5	4	5
Judy	1	5	5	5
etc				

- Suppose a new customer (Joe) votes some of these products as follows:

	X	Y	Z	W
Joe	1	5	5	?

- Estimate if Joe would like product W, in order to recommend it (or not).

IMPLICIT VOTING

- A data set with individuals´s past behaviour (e.g. books visited, etc).

Person	X	Y	Z	W
Gloria	yes	no	yes	yes
Juan	no	yes	no	no
Amin	no	yes	yes	yes
Sam	yes	yes	yes	yes
Judy	yes	yes	yes	no
etc				

- Suppose a new customer (Joe) is currently visiting some of these products.

	X	Y	Z	W
Joe	yes	?	yes	?

- Estimate if Joe would visit book Y and W and recommed them (or not).

AARHUS BSS

MEMORY-BASED ALGORITHMS

CLASSICAL ALG. IN MACHINE LEARNING

- These methods find users that are similar to the active user (i.e. the user we want to make predictions for), and uses their preferences to predict ratings for the active user, **by using the entire dataset.**

- Common similarity measures:
  - Pearson corr.: how much two users vary together
  - Distance measures: Manhattan distance, Euclidian distance, etc.
  - Vector similarity: we can treat two users as vectors and take the cosine of the angle btw. two vectors

J.S. Breese, D.Heckerman, and C.Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence.

PEARSON CORRELATION (REVIEW)

Active user: $user_a$
Any other user in the dataset: $user_i$
Weight (similarity): **w ($user_a$, $user_i$)**

$$w(user_a, user_i) = \rho(V_a, V_i) = \frac{E\left(\left(V_a - \overline{V_a}\right)\left(V_i - \overline{V_i}\right)\right)}{\sigma_{V_a}\sigma_{V_i}} = \frac{\sum_j (v_{aj} - \overline{V_a}) \cdot (v_{ij} - \overline{V_i})}{\sqrt{\sum_j (v_{aj} - \overline{V_a})^2} \sqrt{\sum_j (v_{ij} - \overline{V_i})^2}}$$

AARHUS BSS

EXAMPLE

Items

Users	X	Y	Z	W
user$_1$	1	4	5	4
user$_2$	5	1	1	2
user$_3$	4	1	2	1
user$_4$	2	5	4	5
user$_5$	1	5	5	5
....				

	X	Y	Z	W
user$_a$	1	5	5	?

$Averages:$

$$\bar{V}_1 = \frac{1+4+5}{3} = 3.33$$

$$\bar{V}_2 = \frac{5+1+1}{3} = 2.33$$

...

$$\bar{V}_a = \frac{1+5+5}{3} = 3.67$$

$Weights(similarities)\text{e.g. } \rho:$

$$w(user_a, user_2) = -1.000$$

$$w(user_a, user_1) = .971$$

$$w(user_a, user_3) = -.945$$

$$w(user_a, user_4) = .945$$

$$w(user_a, user_5) = 1.000$$

Prediction (recommendation):

$$\hat{v}_{ak} = \bar{V}_a + \alpha \sum_{i=1}^{n} w(user_a, user_i)(v_{ik} - \bar{V}_i) = 3.67 + .206 \begin{bmatrix} .971(4-3.33) \\ -1(2-2.33) \\ -.945(1-2.33) \\ +.945(5-3.67) \\ +1(5-3.67) \end{bmatrix} = 4.66 =>$$

High preference for W

α – normalizing constant

$$\alpha = \frac{1}{\sum w}$$

NOTE

Advantages

- The quality of predictions are rather good.
- This is a **relatively simple** algorithm to implement for any situation.

Disadvantages

- It uses the entire database every time it makes a prediction, so it needs to be in memory => **extremely slow.**
- It can sometimes not make a prediction for certain active users/items. This can occur if the active user has no items in common with all people who have rated the target item.
- Overfitting the data: it takes all random variability in people's ratings as causation, which can be a real problem.

AARHUS BSS

MODEL-BASED ALG.

- Based only on selecting **a portion of the existing users/items** and use that as a "model" to make recommendations without having to use the complete dataset every time.
- Adv: speed and scalability

- Three possible approaches:
  1. Enhancement of memory-based alg.: calculate similarities using only k-most similar users or items – *these models were seen in the previous lectures with HJJ*
  2. As a linear algebra problem: use linear equations
  3. As a probability model: Bayesian nets – discussed here

AARHUS BSS

AS A PROBABILITY PROBLEM

$$\hat{v}_{ak} = E(V_{ak}) = \sum_{i=1}^{r} i \times P(V_{ak} = i \mid V_a)$$

where P are learned from data

AARHUS BSS

EXAMPLE

- Given $V_a = \{v_{a1} = 1, v_{a2} = 5, v_{a3} = 5\}$

- and the estimated the probabilities associated with each option:

$$P(V_{a4} = 1 | V_a) = .02$$
$$P(V_{a4} = 2 | V_a) = .03$$
$$P(V_{a4} = 3 | V_a) = .10$$
$$P(V_{a4} = 4 | V_a) = .4$$
$$P(V_{a4} = 5 | V_a) = .45$$

$$\hat{v}_{a4} = E(V_{a4}) = \sum_{i=1-5} i \times P(V_{a4} = i | V_a) = 1 \times 0.2 + \ldots + 5 \times 0.45 = 4.23$$

- How do we obtain the probabilities?

OPTION 1: A BN ALGORITHM

Learn the probabilistic relationships using a learning algorithm and select the best model

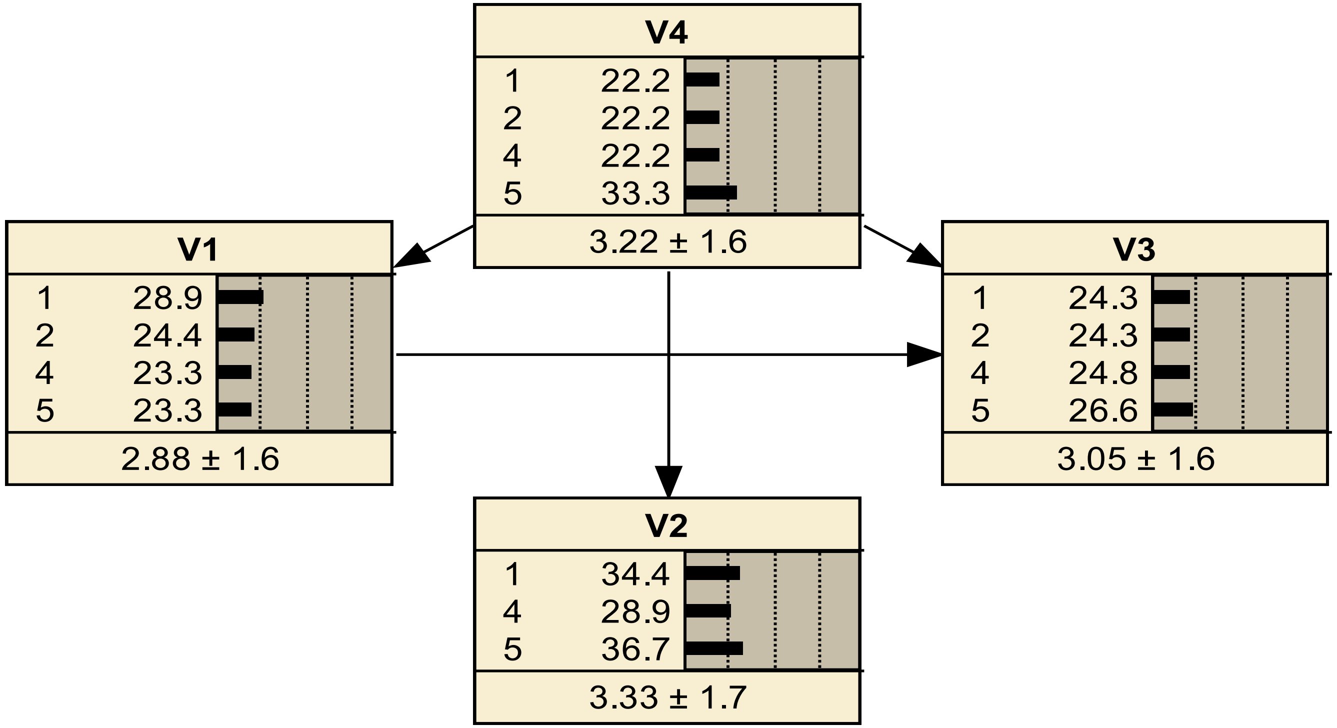
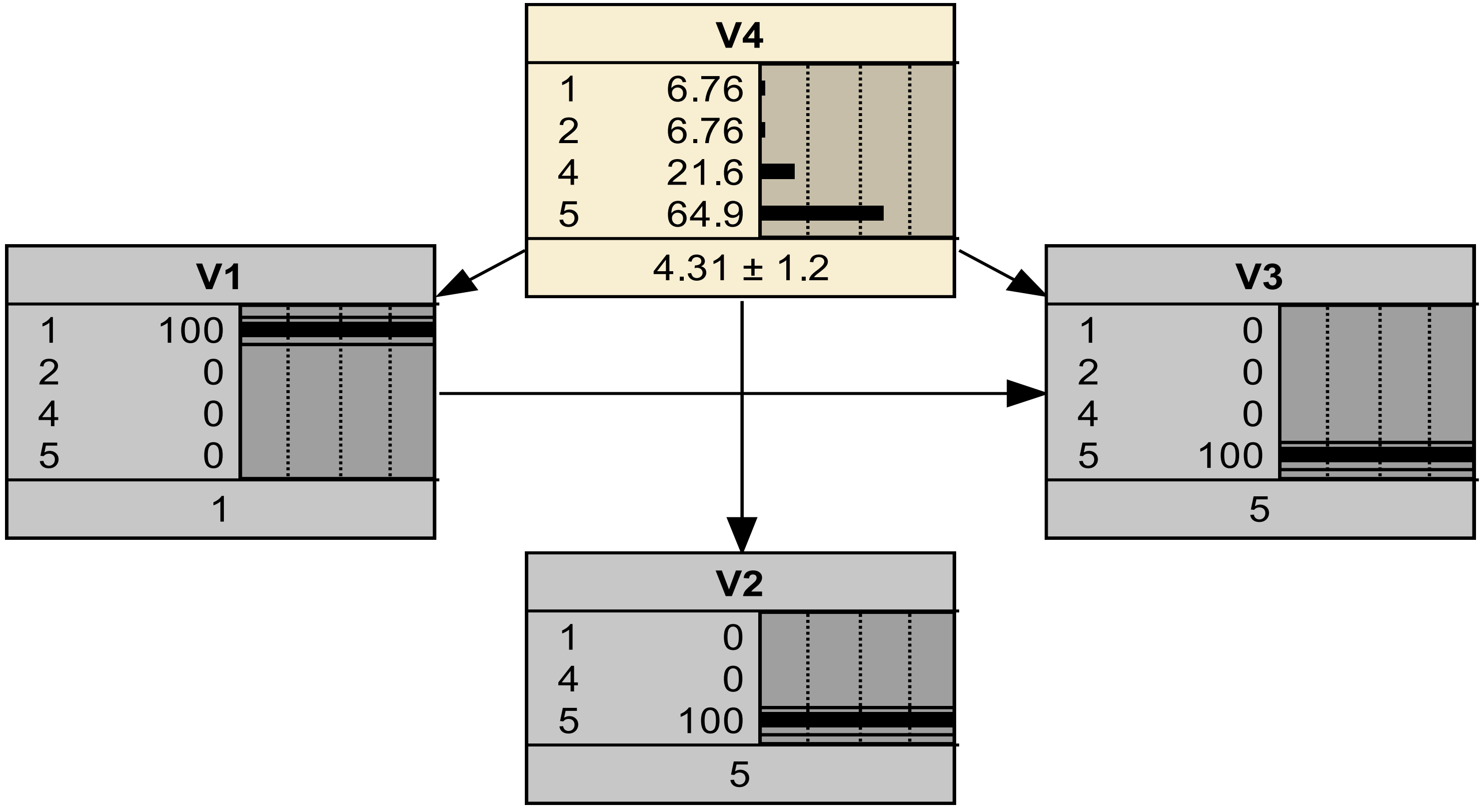


Figure 2: A TAN BN learned from data

(V4 was randomly set as target)

After that, we can do inference for a new customer given the evidence:



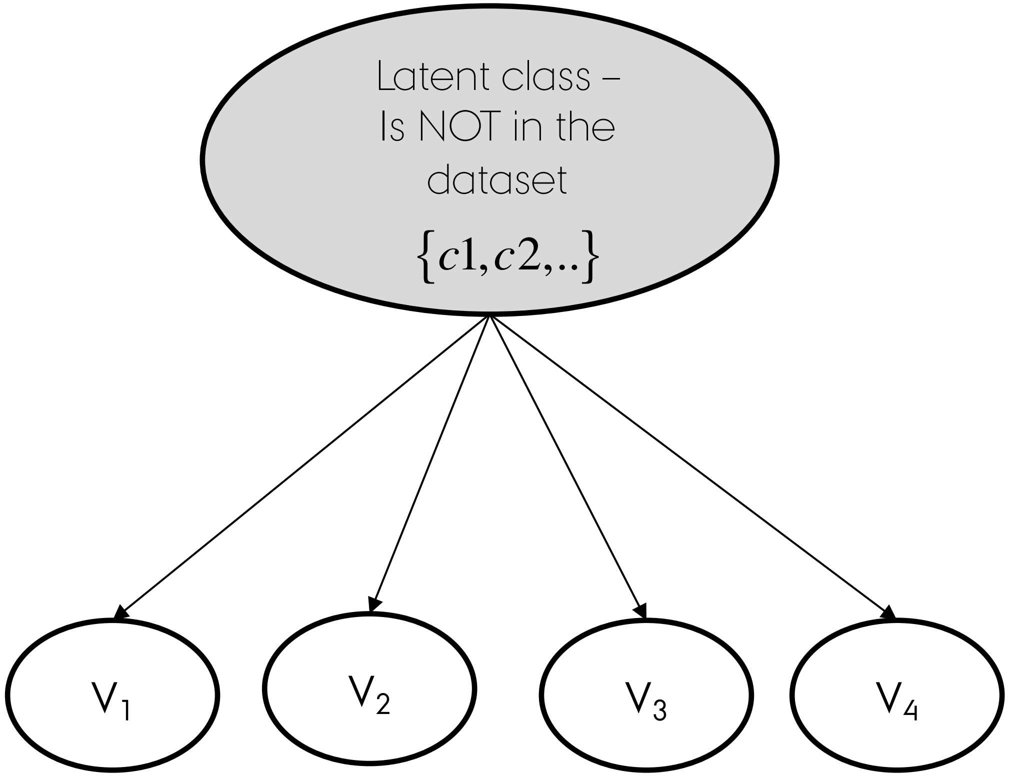
$$P(V_{a4}=i|V_a)\ \ for\ i=1...5\ \ given\ V_a=\{v_{a1}=1, v_{a2}=5, v_{a3}=5\}\ are:$$

$$Estimated\ \ preference\ is: 4.31$$

OPTION 2: THROUGH CLUSTER LEARNING

It works on the assumption that the active user belongs to a latent class (segment) that can accurately predict the ratings for the user on all items. Technically, in line with Markov property, LCA tries to assign groups such that, inside of a group the relations between the observable variables become non-significant ("zero correlated"), because the group membership explains any relationship between the variables.

E.g. V_1 and V_2 are conditionally independent given C=c1.



Observable categorical variables
(columns representing the products in the dataset)

EM ALGORITHM

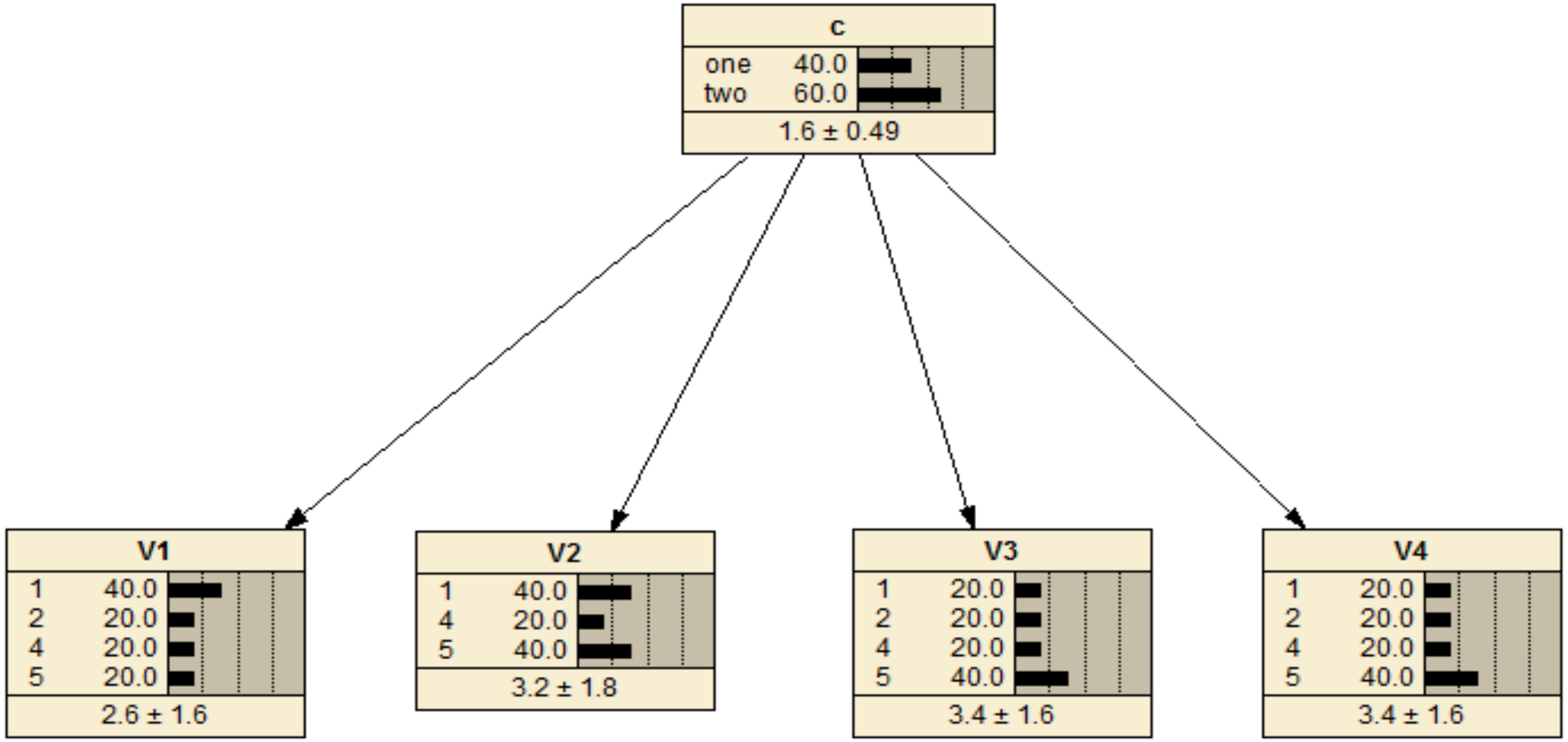
- In LCA, the units are not assigned absolutely to classes, but **probabilistically.** Thus, we get a probability value for each individual to be assigned to cluster 1, cluster 2, …, cluster k.

- LCA uses **EM algorithm** to estimate the parameters.
  - "Expectation step" : estimation of the the class-membership probabilities
  - "Maximization step": the estimates are altered to maximize the likelihood-function
  - Both steps are iterative and repeated until the algorithm finds the global maximum (the highest possible likelihood)

- Traditional procedures for cluster analysis (seen under Segmentation topic) use rules-of-thumb to determine the number of clusters. Comparatively, LCA is a statistical model, which uses statistics to determine the number of clusters - criteria like **BIC, AIC, loglik.**

EXAMPLE

A BN with 2 clusters in Netica



AARHUS BSS

MODEL SELECTION : HOW MANY CLUSTERS?

Criteria for model selection based on LL :

LogLik (LL) for 2 clusters = 3.03245 32 parameters
LogLik (LL) for 3 clusters = 2.44127, 48 parameters
LogLik (LL) for 4 clusters = 2.44129, 64 parameters

$$BIC = \ln(n)k - 2\ln(LL), \qquad n = sample\ size; \quad k = \#\, of\ \ parameters$$

BIC for 2 clusters = 49.28
BIC for 3 clusters = 75.46
BIC for 4 clusters = 101.21

The model with the lowest BIC is preferred (classical rule).

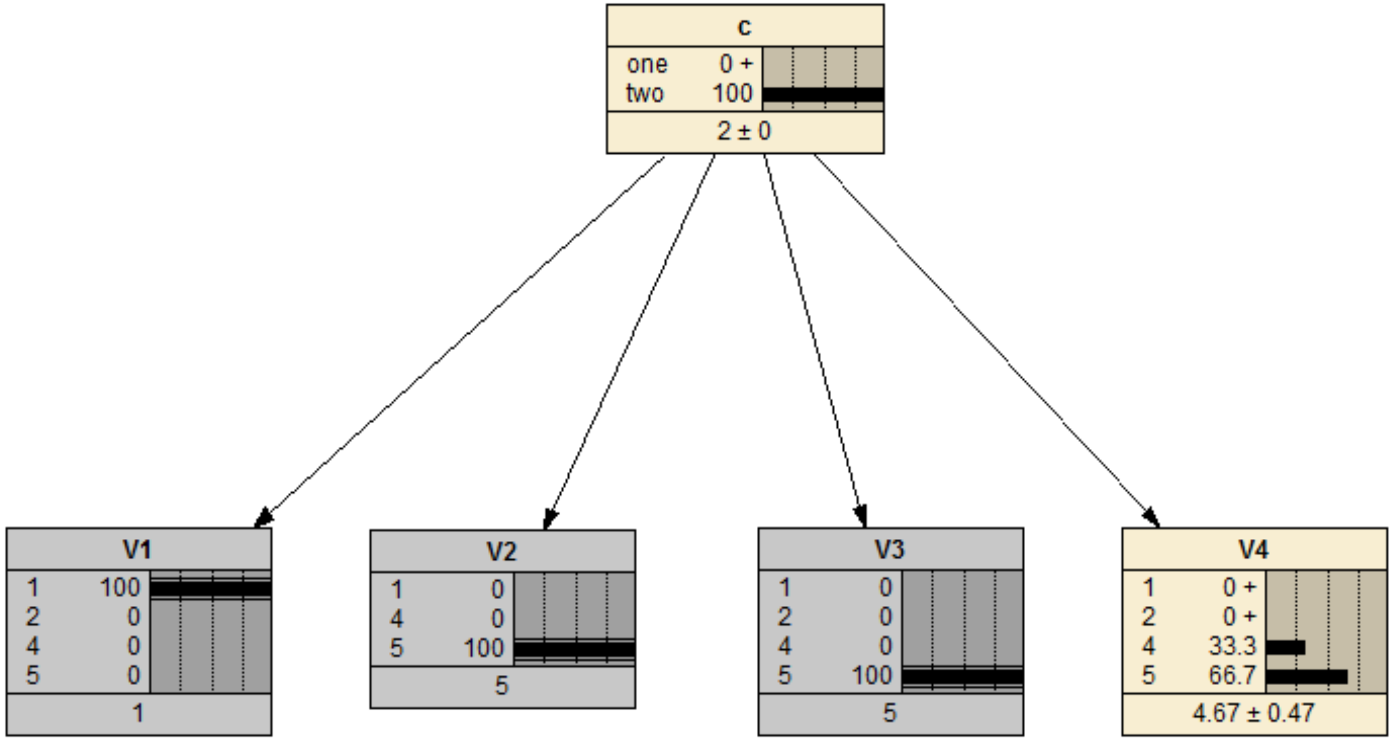
AARHUS BSS

USE

Once the model is selected, it can be used to:

1. Estimate the classes size and describe the classes characteristics
2. Predict the class for new customers based on some evidence
3. Predict preferences and make recommendations

EXAMPLE



$P(V_{a4} = i | V_a)$ *for* $i = 1 \ldots 5$ *given* $V_a = \{v_{a1} = 1, v_{a2} = 5, v_{a3} = 5\}$ *are*:

Estimated preference is: 4.67

CONCLUSION

- poLCA and flexmix libraries can learn a cluster model from a data set of customer preferences
- After that, we can do inference for a new customer (the active user) using a Bayesian network inference algorithm in bnlearn package
- Go to R applications.

REFERENCES

1. Gupta, S. and Kim, H.W. (2008). Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities, European Journal of Operational Research, 190, 818-833
2. Chapter 11: Collaborative Filtering. In: Neapolitan & Jiang: Probabilistic methods for financial and marketing informatics. San Francisco: Morgan Kaufmann
3. Chapter 12: Targeted Advertising. In: Neapolitan & Jiang: Probabilistic methods for financial and marketing informatics. San Francisco: Morgan Kaufmann

AARHUS
BSS