BABES-BOLYAI UNIVERSITY
FACULTATEA DE MATEMATICĂ ȘI
INFORAMTICĂ

# Bayesian Network

*Authors:*
Sergiu BREBAN Razvan SALAJAN

January 3, 2017

# Contents

# List of Figures

# Chapter 1

# Data Mining

## 1.1 Informatia inseamna putere

In ziua de azi, datorita serviciilor ieftine de stocare a datelor, stocarea de informatii devine o treaba usoara si deschide calea invatarii unor lucruri noi. Asadar obtinerea de noi cunostine se transforma in serviciul de extragere prin diferite metode de noi idei dintr-un ocean de informatii stocate care asteapta sa fie analizate. In zilele noastre majoritea actiunilor cotidiene sunt inregistrate si stocate. Datorita numarului mare de oameni si de actiuni pe care un om le face se transmit date intr-un volum greu de imaginat. Avand disponibil un volum atat de imens de date partea mai grea devine analiza, invatarea si corelarea acestora. Desi lucrul cu date a fost folosit de mult timp de catre economisti, statisticieni sau meteorologi e de remarcat cresterea recenta al oportunitatilor de gasire al modelelor in date. Bazat pe [3] numarul datelor la nivel global stocat in bazele de date se dubleaza la fiecare 20 de luni. Avand in vedere volumul imens de date cu care suntem inundati si eficientizarea calculatoarelor de a suporta cautari are ca urmare cresterea interesului in data mining. Astfel, data mining, reprezinta o solutie la a gasi noi concepte, noi idei care pot ajuta la rezolvarea unor probleme din diferite ramuri, de exemplu, intr-un context comercial, obtinandu-se un avantaj comercial.

## 1.2 Definitie

Data mining reprezinta procesul de analiza si cercetare asupra unui volum imens de date stocat in baze de date sau alte tipuri de stocare de informatii. Procesul are ca scop descoperirea de noi cunostinte precum: modele, asocieri, comportamente, anomalii, structuri semnificative care sa ofere descoperiri noi, inovatoare. Procesul poate fi privit ca o cutie inchisa care produce rezultatul dorit. La fel cum si [4] precizeaza, procesul de data mining poate avea si rezultate nedorite daca o metoda este aplicata intr-un context nepotrivit sau modelele sunt construite pe presupuneri eronate.

## 1.3 CISP

"Cross-industry standard practice"(CISP) reprezinta un proces standard de abordare al solutiilor de tip data mining. Acest standard a fost creat cu scopul de a fi liber si de a fi utilizat de catre toti cei care folosesc data mining pentru a rezolva diferite probleme. A fost creat, cumva natural, datorita faptului ca fiecare incerca sa isi fac

propriul proces, astfel creandu-se haos. Asadar, CISP, a luat nastere ca fiind un standard care nu e specific unei industrii, aplicatii, sau unui tool. Bazat pe standardul CISP, un proiect de tip data mining consta din 6 etape dupa cum se poate vedea si din Figure 1.1. Procesul definit e unul iterativ care se poate adapta; fiecare compo-

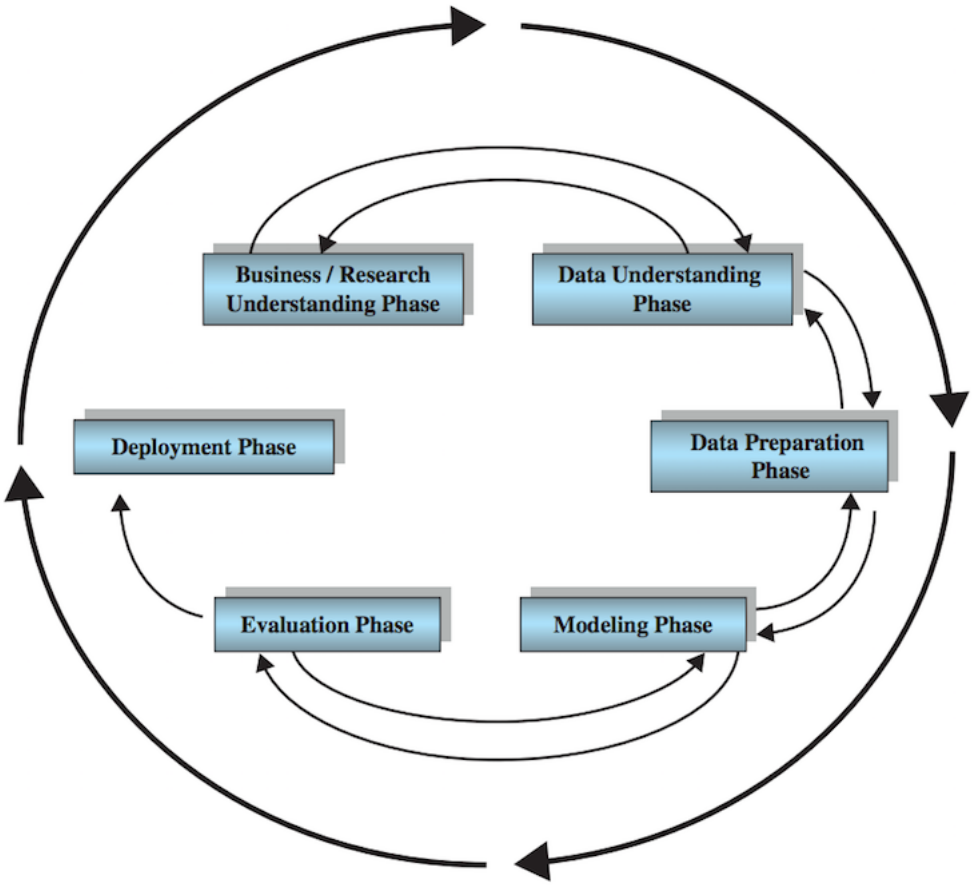


FIGURE 1.1: CISP

nenta depinde de componenta precedenta iar in caz de iregularitati procesul se poate intoarce inapoi la pasul precedent.

Data mining poate fi asociat cu expresii precum descoperirea de cunostinte in baze de date, (potrivit lui [2]), cu toate ca alti cercetatoti considera, data mining, ca fiind un pas important in descopeirirea de noi cunostinte. Se poate sintetiza procesul de descoperire a cunostintile din date in urmatorii pasi (bazat pe [2]): curatarea datelor, integrarea datelor, selectarea datelor, transformarea datelor, extragerea datelor, evaluarea modelelor si prezentarea cunostintelor. Acesti 7 pasi pot fi impartiti in 2 grupuri mai mari. Primii 4 pasi pot fi priviti ca construirea unor depozite de date si efectuare unor operatii pe acestea. Ultimii 3 pasi se transforma intr-un proces iterativ denumit data mining.

## 1.4 La ce se foloseste?

Cerintele in care data mining e folosit s-au grupat in 6 clase (potrivit cu [4]).

- Descrierea: reprezinta modul prin care se cauta cai care precizeaza unele "trenduri" sau modele care se observa in date.

- Estimarea: se refera la a estima valoarea numerica a unui atribut tinta avand la dispozitie variabile predictorii care pot fi de tipul numerice sau categoriale(?). De exemplu se pot face estimari despre cat v-a cheltui o familie cu 2 copii pentru inceperea unui nou an scolar. De asemenea, un alt exemplu il reprezinta prezenta la vot pentru alegerile curente ale cetatenilor cu varsta cuprinsa intre [18-24] cu drept de vot bazat pe alegerile din anii trecuti.

- Predictia: este similara cu estimarea si clasificarea cu precizarea ca rezultatul predictiei o sa se intample in viitor. Exemple de predictii ar fi: pretul unei actiuni la bursa peste cateva luni sau cine va castiga campionatul mondial de fotbal. Bazat pe [4], in alegerile prezidentiale din America din 2012, echipa de campanie a lui Barack Obama s-a folosit de data mining. Acestia au folosit data mining pentru a identifica tipul sustinatorilor lui Barack Obama si s-au asigurat ca acestia merg la vot si au mai folosit pentru a face predictii despre voturile pentru Barack Obama in fiecare judet. Un exemplu de predictie pentru judetul Hamilton din statul Ohio considerat "indecis", a fost de 56.4% iar rezulultul final a fost de 56.6%.

- Clasificarea: este asemantoare cu estimarea cu diferenta ca variabila tinta, care se doreste sa fie clasificata, are ca valori categorii. Modelul de data mining examineaza datele care contini valori pentru atributele predictorii si atributul tinta.

- Gruparea: se refera la procesul de a grupa date, observatii, cazuri in clase asemanatoare. O grupare reprezinta o colectie de obiecte care prezinta insusiri asemanatoare intre ele. Gruparea reprezinta o sarcina diferita conceptual fata de clasificare, estimare sau predictia. In procesul de grupare nu exista un atribut tinta care sa fie clasificat. In schimb, gruparea, incearca sa imparta datele in grupuri sau clase relativ omogene, astfel incat similaritatiile elementelor dintr-un grup sunt maximizate iar disimilaritatille cu elemente straine grupului sunt minimizate.

- Asocierea: consta in a determina ce atribute "merg impreuna". Are ca scop gasirea de legaturi intre atribute. Este folosita cu succes in analiza pietei cumparatorilor din magazine pentru a gasi ce produse sunt cumparate impreuna si ce produce nu sunt. Asocierile sunt de forma "daca antecedentul atunci consecinta".

# Chapter 2

# Rețeaua Bayesiană

## 2.1 Definiție

O rețea Bayesiană este un model probabilistic grafic care reprezintă un set de variabile aleatoare și dependețele condiționale dintre ele sub forma unui graf orientat aciclic (DAG).

Fiecare nod al grafului reprezintă o variabilă aleatoare, iar arcele dintre noduri reprezintă dependențele probabilistice dintre nodurile conectate.

Clasificatorul naiv Bayes presupune că toate variabilele sunt independent condiționate, în contrast cu rețelele Bayesiene, care permite condiționarea independentă aplicată la subseturi de variabile.

Aceste modele probabilistice, ca modelul naiv Bayes sau modelele logistice de regresie sunt diferite de alte modele de reprezentare, cum ar fi arborii de decizie, prin faptul că produc estimări probabilistice în loc de clasificări exacte.

Pentru fiecare clasă de valori, estimează probabilitatea ca o instanță dată să aparțină acelei clase. Aceste estimări probabilistice sunt mai utile decât simple predicții, deoarece pot fi clasate, iar costul acestora poate fi minimizat.

## 2.2 Deducția rețelei Bayesiene

Deducția rețelei Bayesiene se realizează în 3 pași:

- Deducția de variabile neobservate; rețeaua poate fi folosită pentru a oferi informații probabilistice despre relațiile dintre variabile.
- Învățarea probabilităților condiționale; pentru fiecare nod specificarea distribuției probabilităților condiționate de părinții nodului respectiv.
- Învățarea structurii rețelei și construirea grafului.

## 2.3 Învățarea rețelei Bayesiene

Algoritmul pentru construirea rețelei Bayesiene are două componente: o funcție pentru evaluarea rețelei în funcție de date și o metodă de a genera toate rețelele posibile și a o selecta pe cea mai bună.

După definirea structurii grafului care reprezintă rețeaua, calculul probabilităților condiționale este ușor de realizat, necesitând doar calculul frecvențelor relative a combinațiilor de atribute asociate din setul de date.

## 2.4 Reprezentarea datelor

Pentru reprezentarea datelor, un format foarte răspândit este formatul ARFF, datele fiind stocate într-un fișier cu formatul .arff. Acest format permite definirea atributelor și a valorilor posibile pentru fiecare atribut, cât și un set de instanțe, cu valori specificate pentru fiecare atribut.

Exemplu de fișier .arff:

```
@relation skiing
@attribute temperature { hot, mild, cool }
@attribute windy { true, false }
@attribute outlook { sunny, overcast }
@attribute snowCover { low, medium, high }
@attribute rainfall { sleet, rain, snow, none }
@attribute ski? { yes, no }
@data
hot, false, overcast, medium, none, no
hot, true, sunny, high, none, no
hot, true, sunny, high, sleet, no
mild, false, sunny, high, none, yes
mild, false, sunny, low, none, no
cool, true, sunny, medium, none, yes
cool, true, overcast, medium, snow, no
```

Acest set de date reprezintă starea unei pârtii de ski, caracterizată de 5 atribute, cu valorile corespunzătoare:

```
temperature { hot, mild, cool }
windy { true, false }
outlook { sunny, overcast }
snowCover { low, medium, high }
rainfall { sleet, rain, snow, none }
```

## 2.5 Algoritm de învățare a structurii

Odată reprezentate datele, avem nevoie de algoritmi specifici pentru a construi și inițializa structura rețelei Bayesiene.

Un exemplu de structură pentru setul de date despre starea pârtiei de ski poate fi observat în Figura 2.1.

Unul dintre cei mai eficienți algoritmi pentru construirea structurii este algorimul K2, iar pentru calcul costului unui graf în cadrul algorimului de căutare, funcția de scor K2, propusă de Cooper and Herskovits (1992). [1]

### 2.5.1 Algoritmul K2

Algoritmul K2 începe cu o ordine dată a atributelor (noduri) și încearcă să adauge o muchie de la nodurile procesate la cel curent, astfel încât scorul rețelei să fie maxim.

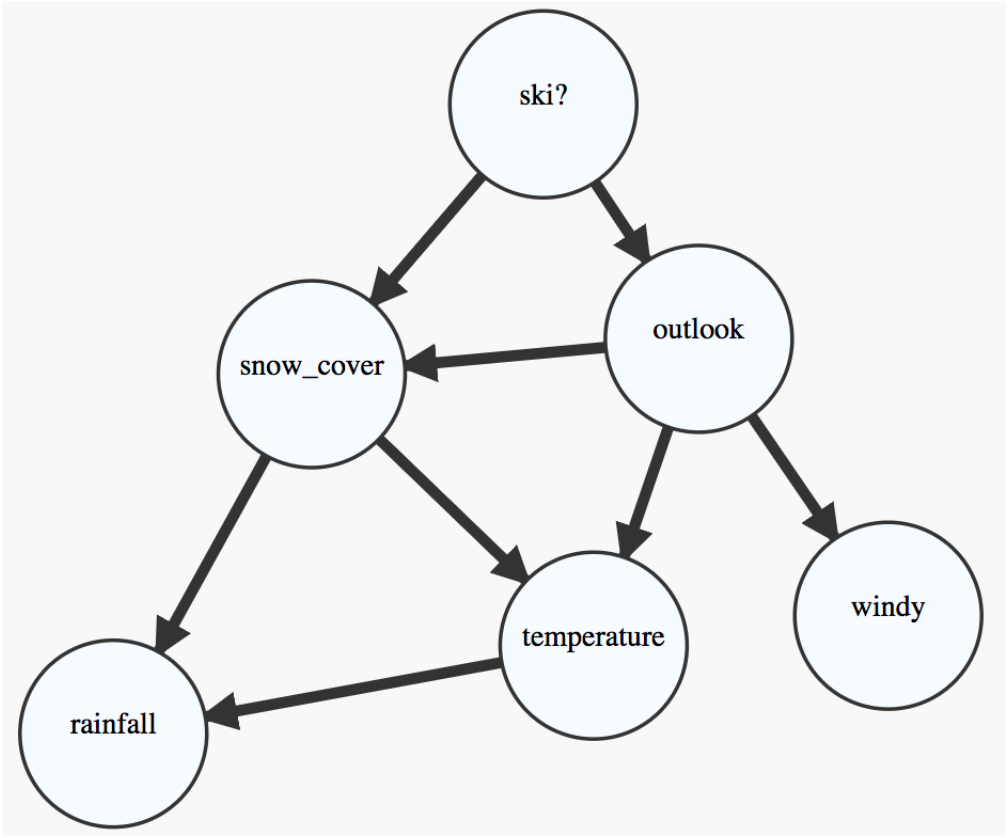


FIGURE 2.1: Strucura unei rețele bayesiene

Numărul maxim de părinți poate fi restricționat pentru un nod (la 2 de exemplu) pentru a evita overfitting-ul. În funcția de scor K2 și în calculul tabelelor probabilităților condiționale se poate folosi estimatorul Laplace. [5]

# Bibliography

[1] Alexandra M Carvalho. "Scoring functions for learning Bayesian networks". In: *Inesc-id Tec. Rep* (2009).

[2] Jiawei Han. *Introduction to Data Mining*. URL: http://hanj.cs.illinois.edu/pdf/ency99.pdf.

[3] Mark A. Hall Ian H. Witten Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann; 3 edition (January 20, 2011), 2013.

[4] Daniel T. Larose. *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining*. John Wiley & Sons, 2014.

[5] Carolina Ruiz. "Illustration of the K2 algorithm for learning Bayes net structures". In: *Department of Computer Science, WPI* (2005).