INM431 Machine Learning

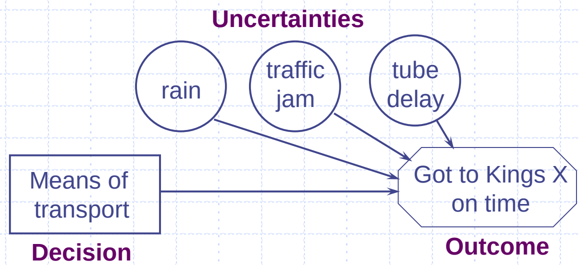
Artur S. d'Avila Garcez
a.garcez@city.ac.uk
http://www.staff.city.ac.uk/~aag/

Content

- Bayesian Networks
- K2 Algorithm

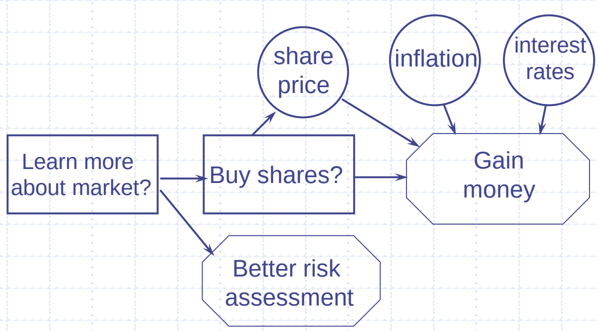
Bayesian Nets

Influence Diagrams:



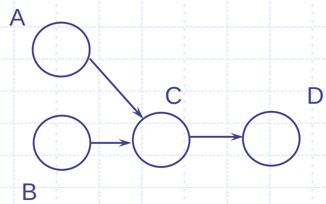
Important: influence diagrams must be **acyclic**

Sequential Decision



The model shows, e.g. that the decision to buy shares can affect the share price…

Uncertainties may influence each other



P(A)
P(B)
P(C|A∧B)
P(C|~A∧B)
P(C|A∧~B)
P(C|~A∧~B)
P(D|C)
P(D|~C)

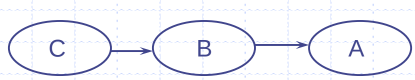
So, denoting "∧" by "," we can say that
P(A,B,C,D) = P(D|C) P(C|A,B) P(A) P(B)

A Simple Example

1. The facts

A is "Mike is not answering his phone"
B is "Mike is at home"
C is "It is raining"

2. The fact dependencies



Note: C affects A indirectly in this case, so they can be treated as independent!

Associated Table of Dependencies

Let:
$P(A|B) = 0.1$, $P(A|\sim B) = 1$
$P(B|C) = 0.8$, $P(B|\sim C) = 0.5$
$P(C) = 0.6$

The probability that it is raining and Mike is at home but not answering the phone:

$P(A,B,C) = P(A|B)\ P(B|C)\ P(C)$

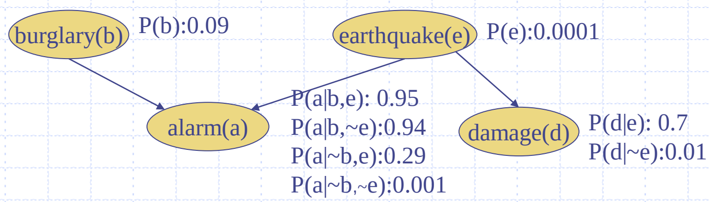
Other possible calculations…

1. Calculate any of the dependent probabilities
e.g. the probability that Mike is not answering the phone: $P(A) = ?$

2. Compute the probability of some event given the evidence
e.g. the probability that it is raining given that Mike isn't answering the phone: $P(C|A) = ?$

3. Compute the most likely set of events, given the evidence
e.g. What is the most likely explanation for Mike not answering the phone?

Bayesian network: example



- P(burglary|alarm)?
- P(earthquake|alarm)?
- P(burglary,earthquake|alarm)?

DGMs: Directed Graphical Models conditional independence (1)

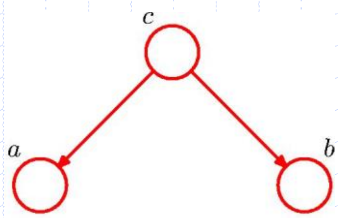
Consider a, b, c such that

$$p(a|b,c) = p(a|c)$$

We say that a is **conditionally independent** of b given c.

Conditional independence properties of the joint distribution can be read directly from the graph without having to perform any analytical manipulations

DGMs: conditional independence (2)



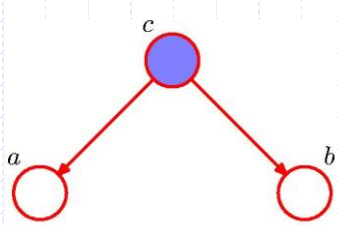
$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c)$$

Note: node *c is said to be tail-to-tail with respect* to the path from *a* to *b*

11

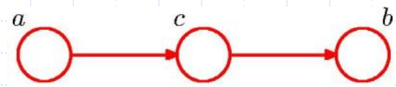
DGMs: conditional independence (3)



$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

12

DGMs: conditional independence (4)



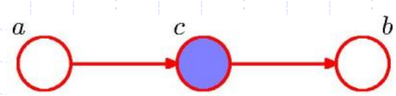
$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

Note: node *c is said to be head-to-tail with respect* to the path from *a* to *b*

13

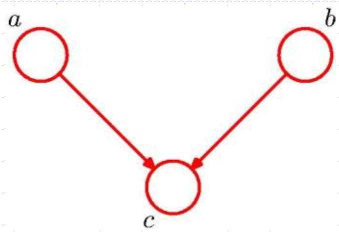
DGMs: conditional independence (5)



$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

14

DGMs: conditional independence (6)



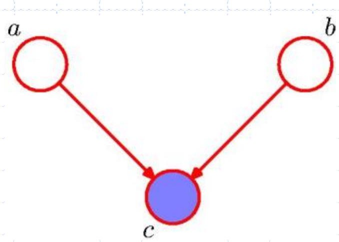
$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

$$p(a,b) = p(a)p(b)$$

Note: node *c is said to be head-to-head with respect* to the path from *a* to *b*

15

DGMs: conditional independence (7)



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

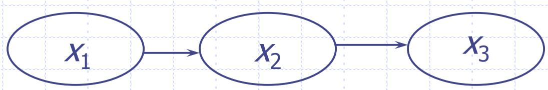
16

Structure Learning

Consider the dataset D:

case	x_1	x_2	x_3
1	1	0	0
2	1	1	1
3	0	0	1
4	1	1	1
5	0	0	0
6	0	1	1
7	1	1	1
8	0	0	0
9	1	1	1
10	0	0	0

Find the Bayesian net that describes the observed data the most!



Score and search…

Start from an initial structure (generated randomly or from domain knowledge) and move to the neighbour with the best score in the structure space until a local maximum of the score function is reached.

This greedy learning process can re-start several times with different initial structures to improve the result.

How many Bayesian nets?

Number of Directed Acyclic Graphs (DAGs) is super-exponential on the number of variables

Number of variables in DAG	Number of the possible DAGs
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1,138,779,265
8	78,370,2329,343
9	1,213,442,454,842,881
10	4,175,098,976,430,598,100

Score function: evaluates how well a given DAG matches the data, e.g. apply maximum likelihood and select the DAG that predicts the data with the highest probability

K2 algorithm

1. Applies to discrete variables; x in {0,1,2,…}
2. Assumes a maximum number n of parents for each node
3. Starts from an initial Bayesian network and moves to add parents incrementally to each node (deterministically or stochastically) until a local maximum is reached (i.e. score and search)

© Artur Garcez

A note on greedy search

- Starts at a specific point (an initial tree, network, etc.) in the hypothesis space;
- Considers all nearest neighbours of the current point, and moves to the neighbour that has the highest score (with a probability in the case of stochastic search);
- If no neighbours have higher score than the current point (i.e., we have reached a local maximum), the algorithm stops.

K2 heuristic

4. Assumes a total order on the set of variables such that, e.g. if n=2 and order is x_1,x_2,x_3,x_4 then:

   x_1 can't have parents,

   x_2 may have x_1 as parent,

   x_3 may have x_1 and x_2 as parents,

   x_4 may have two of x_1,x_2,x_3 as parents.

K2 score function

$$g(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

To compare the networks where node x_i has sets of parents π_i; the highest score wins.

r_i is the size of the list of all possible values of x_i

q_i is the size of list with the Cartesian product of all possible values for the parents of x_i

N_{ijk} is the number of times in D that x_i takes its k^{th} value and the parents of x_i in π_i take the j^{th} instance of the Cartesian product

$N_{ij} = \Sigma_k N_{ijk}$ i.e. number of times the parents take the j^{th} instance

In Summary

A scoring function evaluates how well a given Bayesian network G matches the data D.

Given a scoring function, the best Bayesian network is the one that maximizes this scoring function.

An ad-hoc scoring function is used based on maximum likelihood.

For details:

Gregory F. Cooper and Edward Herskovits, A Bayesian method for the induction of probabilistic networks from data, MLJ, Oct 1992

Other scoring functions (1)

- Criteria for model selection among a finite set of models!
- Seek to maximize likelihood by adding parameters (i.e. increasing the number of edges in the graph).
- Doing so may result in overfitting. Therefore, add a penalty for larger DAGs, e.g.:

$$\max\ (\log P(D|G)) - \frac{k}{2} \log N$$

where, D = data, G = graph, k = number of parameters in the model (e.g. number of coefficients of a regression model, number of entries in the probability tables of a Bayesian net), N = number of examples in D

It is common to use the natural log, i.e. *ln*

Other scoring functions (2)

- In practice, given two or more graphs:
- For each graph, estimate maximum likelihood L = max (ln P(D|G)) from your dataset
- Calculate the Bayesian Information Criterion BIC = k *ln* N – 2 L
- The graph with the lowest BIC is preferred

- Akaike information criterion (AIC): Similar to BIC but uses information loss
- AIC = 2 k – 2 *ln* L
- L is the maximum value of the likelihood function for a model

Other scoring functions (3)

Some scoring functions are based on the concept of Mutual Information I(X,Y): it measures how much information X and Y share, i.e. how much knowing one reduces uncertainty about the other.

See

http://www.cs.technion.ac.il/~dang/journal_papers/friedman1997Bayesian.pdf

http://www.jmlr.org/papers/volume7/decampos06a/decampos06a.pdf