Advanced probabilistic methods

Lecture 2

Pekka Marttinen

Aalto University

January, 2019

Lecture 2 overview[1]

- Bayesian networks (also called 'belief networks')
  - Definition
  - Motivation
- Independence in Bayesian networks
  - d-separation
  - Markov equivalence
- Computation using Bayesian networks
- Ch. 3 in Barber

[1] These slides build upon the book *Bayesian Reasoning and Machine Learning* and the associated teaching materials. The book and the demos can be downloaded from *www.cs.ucl.ac.uk/staff/D.Barber/brml.*

Bayesian networks

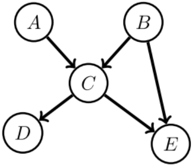
- A Bayesian network is a directed acyclic graph (DAG) in which nodes represent random variables, whose joint distribution can be written as

$$p(x_{1,}\dots,x_D) = \prod_{i=1}^{D} p(x_i|\text{pa}(x_i)),$$

where $\text{pa}(x_i)$ represent the parents of x_i.

- Example:

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



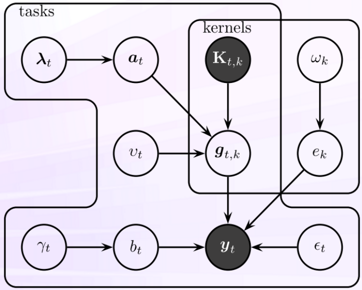
$p(E|B,C)$

Bayesian networks in machine learning

- **An important conceptual tool**
  - BNs are a concise way to represent and communicate the structure and assumptions of a model
- **Computational efficiency**
  - Compact representation of the joint distribution
  - Efficient algorithms exist to compute conditional distributions

Example 1: Prediction of drug sensitivity

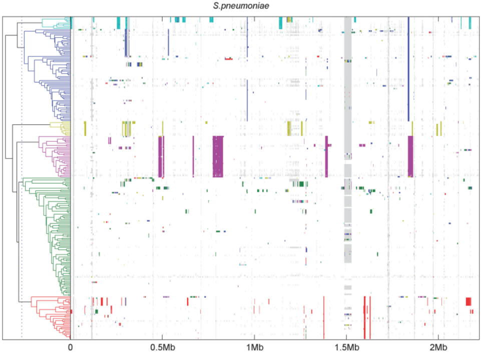
- Solution combines: multiple kernel learning, multiview learning, multitask learning, Bayesian inference
- Team Aalto rank 1/47. (Costello et al. *Nature Biotechnology*, 2014)
- Bayesian wetworks were used to communicate the model structure in a compact way



$$
\begin{aligned}
\lambda_{t,i} &\sim \mathcal{G}(\lambda_{t,i}; \alpha_\lambda, \beta_\lambda) \\
a_{t,i}|\lambda_{t,i} &\sim \mathcal{N}(a_{t,i}; 0, \lambda_{t,i}^{-1}) \\
\upsilon_t &\sim \mathcal{G}(\upsilon_t; \alpha_\upsilon, \beta_\upsilon) \\
\boldsymbol{g}_{t,k}|\boldsymbol{a}_t, \mathbf{K}_{t,k}, \upsilon_t &\sim \mathcal{N}(\boldsymbol{g}_{t,k}; \mathbf{K}_{t,k}\boldsymbol{a}_t, \upsilon_t^{-1}\mathbf{I}) \\
\omega_k &\sim \mathcal{G}(\omega_k; \alpha_\omega, \beta_\omega) \\
e_k|\omega_k &\sim \mathcal{N}(e_k; 0, \omega_k^{-1}) \\
\gamma_t &\sim \mathcal{G}(\gamma_t; \alpha_\gamma, \beta_\gamma) \\
b_t|\gamma_t &\sim \mathcal{N}(b_t; 0, \gamma_t^{-1}) \\
\epsilon_t &\sim \mathcal{G}(\epsilon_t; \alpha_\epsilon, \beta_\epsilon) \\
\boldsymbol{y}_t|b_t, e_{.}, \boldsymbol{g}_{t,.}, \epsilon_t &\sim \mathcal{N}\left(\boldsymbol{y}_t; \sum_{k=1}^{K} e_k \boldsymbol{g}_{t,k} + b_t\mathbf{1}, \epsilon_t^{-1}\mathbf{I}\right)
\end{aligned}
$$

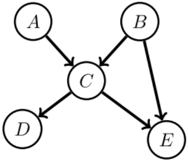
Example 2: Analysis of bacterial genomes

- Dynamic programming algorithms for Hidden Markov Models were used to learn class labels at different locations in the genomes
- By the time of publication, the first method for detecting recombination in large whole genome data sets (Marttinen et al. *Nucleic Acids Research*, 2012)



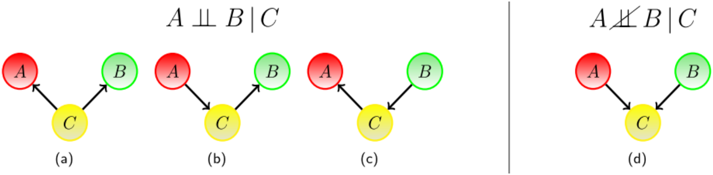
Reading independence statements from the DAG

- Motivating example: do the following independence statements hold:
  - $A \perp\!\!\!\perp B$
  - $A \perp\!\!\!\perp B | E$
  - $D \perp\!\!\!\perp E | C$?



Independence in Bayesian networks (1/2)

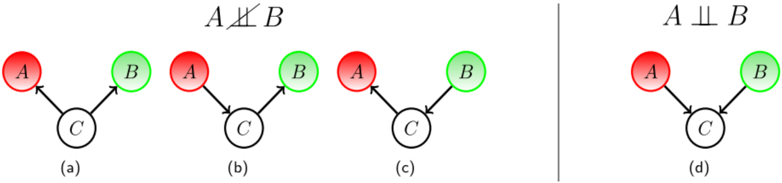
- Possible BNs with three nodes and two links


$A \perp\!\!\!\perp B \mid C$
$A \not\!\perp\!\!\!\perp B \mid C$
A
B
C
A
B
C
A
B
C
A
B
C

(a) (b) (c) (d)

- In (a), (b), and (c), A and B are **conditionally independent** given C.
  - $p(a, b|c) = p(a|c)p(b|c)$
- In (d), A and B are not conditionally independent given C
  - $p(a, b|c) \propto p(a)p(b)p(c|a, b)$

Independence in Bayesian networks (2/2)

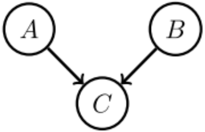


- In (a), (b), and (c), A and B are marginally dependent
- In (d) the variables A and B are **marginally independent**

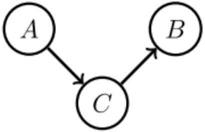
$$p(a, b) = \sum_{c} p(a, b, c) = \sum_{c} p(a)p(b)p(c|a, b) = p(a)p(b)$$

Collider

- A collider (v-structure, head-to-head meeting) has two incoming arrows along a chosen path



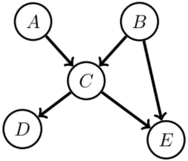
- C a collider
  - $A \perp\!\!\!\perp B$
  - $A \not\!\perp\!\!\!\perp B|C$
  - e.g. $A =$'Talented in sports', $B =$'Talented in maths', $C =$'Admitted to school'



- C a non-collider
  - $A \not\!\perp\!\!\!\perp B$
  - $A \perp\!\!\!\perp B|C$
  - e.g. $A =$'Cumulative sum of $n-1$ dice throws', $C =$'Cumsum of n throws', $B =$'Cumsum of $n+1$ throws'

Blocked paths

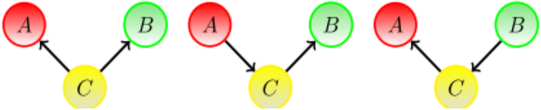
- **A path** between variables A and B **is blocked** by a set of variables $\mathcal{C}$, if
  - there is a collider in the path s.t. neither the collider nor any of its descendants is in the conditioning set $\mathcal{C}$
  - there is a non-collider in the path that is in the conditioning set $\mathcal{C}$.
- Sets of variables $\mathcal{A}$ and $\mathcal{B}$ are **d-separated** by $\mathcal{C}$ if all paths between $\mathcal{A}$ and $\mathcal{B}$ are blocked by $\mathcal{C}$.
  - d-separation implies: $\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{C}$



$A \perp\!\!\!\perp B$?
$A \perp\!\!\!\perp B | E$?
$D \perp\!\!\!\perp E | C$?

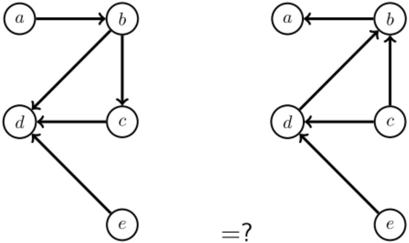
Markov equivalence

- Two graphs are **Markov equivalent**, if they
  - entail the same conditional independencies
  - equivalently: have the same d-separations
- For example:

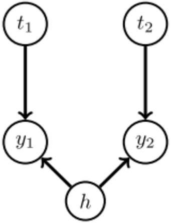


Determining Markov equivalence

- **skeleton:** undirected graph obtained by removing directions
- **immorality**: a collider structure $A \rightarrow C \leftarrow B$, such that there is no direct edge between A and B
- Two graphs are Markov equivalent if and only if they have the same skeleton and the same set of immoralities

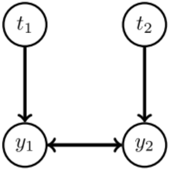


Limitations of expressibility*

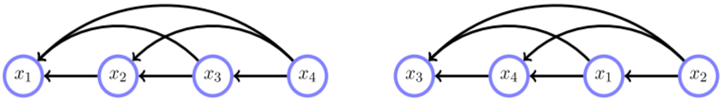


- $t_1 \perp\!\!\!\perp t_2, y_2$ and $t_2 \perp\!\!\!\perp t_1, y_1$
- No Bayesian network for t_1, t_2, y_1, y_2 exists that could capture these independence statements (why?)

- A generalized class of models with bi-directed edges



How to select a DAG to model the system?



- Full graph can represent any distribution:

$$p(x_1, x_2, x_3, x_4) = p(x_1|x_2, x_3, x_4)p(x_2|x_3, x_4)p(x_3|x_4)p(x_4),$$

or equally valid

$$p(x_1, x_2, x_3, x_4) = p(x_3|x_4, x_1, x_2)p(x_4|x_1, x_2)p(x_1|x_2)p(x_2).$$

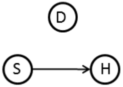
    - Misses all benefits of structure!

- Graph can only be determined up to Markov equivalence class

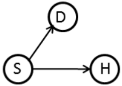
Example on using too simple or too complex a DAG (1/2)

- Simulate data from the 'true model'
- Train a model with training data, try to predict D given S and H in the test data
- S ='Smoking', H ='Hypertension', D ='(Some) Disease'

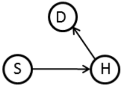
M_1 (too simple)



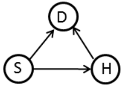
M_2 ('true model')

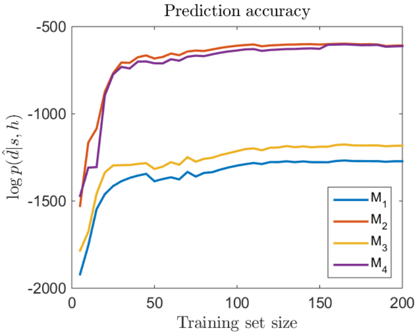


M_3 (wrong)



M_4 (too complex)





Example on using too simple or too complex DAG (2/2)

- Having a bit too simple model is much worse for prediction accuracy than having a bit too complex model
  - Advisable to include all components that may be useful in the prediction model
  - However: way too complex models waste data to learn redundant parameters
- Other interesting points:
  - M_3 seems better than M_1. Why?
  - The negative impact of having too complex a model is most severe when the amount of training data is limited (overfitting).

Possible ways to specify the graph

1. Construct the graph using assumptions about the system
   - add edges based on perceived **direct causalities**
   - Details in the following slides [2]
2. Learn structure from data
   - Ch. 9.5

- Before use, the model should always be checked
  - cross-validation
  - inspection or residuals
  - ...

[2] The slides about causal DAGs follow the derivations of Richard E. Neapolitan (2004) *Learning Bayesian Networks*, Ch. 1.5

Definition of causality (1/2)*

- Let S ='Sprinkler on', G ='Grass wet'
- By observing the values of S and G, we would surely find them dependent, so $p(s, g) \neq p(s)p(g)$
- Non-symmetric:
  - Turning the sprinkler on makes grass wet
  - Watering the grass (by some means other than the sprinkler) does not turn the sprinkler on
  - Interpretation: S is a cause of G

www.123rf.com

Definition of causality (2/2)*

- The causality can be defined by intervetions (manipulations of variables)
  - Set the value of a putative cause to a certain value.
  - Investigate if the distribution of the putative effect changes
- For example ($S =$'Sprinkler on', $G =$'Grass wet'):

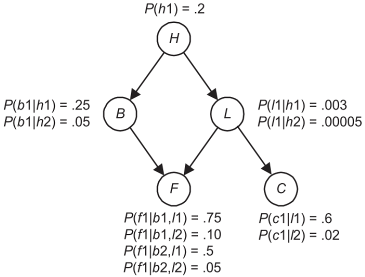
$$p(G = 1|do(S = 1)) \neq p(G = 1|do(S = 0))$$
$$p(S = 1|do(G = 1)) = p(S = 1|do(G = 0))$$

  Therefore: S is a cause of G

- Interventions form the basis of *randomized controlled experiments* that are used in clinical trials, for example to assess the effectiveness of a drug.

Causal DAG

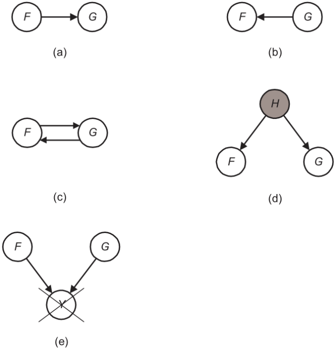
- Suppose our understanding of the causal influences is the following:
  - A history of smoking (H) causes bronchitis (B) and lung cancer (L).
  - Bronchitis (B) and lung cancer (L) can cause fatigue (F).
  - Lung cancer can cause a positive chest X-ray (C)
- A few additional assumptions aside (see next slide), the following DAG with empirically determined conditional distributions could be used to represent our knowledge



Neapolitan (2004, Fig 10)

Underlying assumptions in causal DAGs

- Correlation between F and G could be explained by the following causal structures:
- When constructing DAGs using causal edges, we must assume
  - No **feedback loops** (c)
  - No **hidden common causes** (d)
  - No **selection bias** (e)

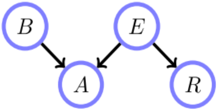


Neapolitan (2014, Fig 1.12)

Inference

- Inference corresponds to using the distribution to answer a question about the environment.
- Examples
  - What is the probability $p(x = 4 | y = 1, z = 2)$?
  - What is the most likely joint state of the distribution $p(x, y)$?
  - What is the probability the stock market will do down tomorrow?
- Computational Efficiency
  - For singly-connected graps (e.g. trees), there exist efficient algorithms based on the concept of message passing.
  - In general, the case of multiply-connected models is computationally inefficient.
  - Ch 5 & 6

Computation - example (1/3)



- A: 'Alarm is on', B: 'There's a burglar in the house', E:'There's an earthquake', R:'Radio reports of an earthquake'
- Compute $p(B = 1|A = 1)$,the probability that there's a burglar, given the alarm is on.
- Conditional probabilities:

$p(A = 1\|B, E)$	B	E
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$p(R = 1\|E)$	E
1	1
0	0

$$p(E = 1) = 0.000001$$

$$p(B = 1) = 0.01$$

Computation - example (2/3)

$$
\begin{aligned}
p(B=1|A=1) &\overset{1}{=} \frac{p(B=1,A=1)}{p(A=1)} \\
&\overset{2}{=} \frac{\sum_e \sum_r p(B=1,A=1,E=e,R=r)}{\sum_b \sum_e \sum_r p(B=b,A=1,E=e,R=r)} \\
&\overset{3}{=} \frac{\sum_e \sum_r p(A=1|B=1,E=e)p(B=1)p(E=e)p(R=r|E=e)}{\sum_b \sum_e \sum_r p(A=1|B=b,E=e)p(B=b)p(E=e)p(R=r|E=e)}
\end{aligned}
$$

1: definition of conditional probability, 2: marginalization, 3: factorization of the joint distribution according to the BN

Computation - example (3/3)

By reordering to simplify computations, we get further that

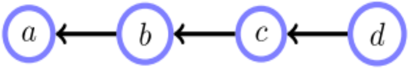
$$\ldots = \frac{\sum_e p(A=1|B=1,E=e)p(B=1)p(E=e)\sum_r p(R=r|E=e)}{\sum_b \sum_e p(A=1|B=b,E=e)p(B=b)p(E=e)\sum_r p(R=r|E=e)},$$

and, because $\sum_r p(R=r|E=e)=1$, we finally get

$$\ldots = \frac{\sum_e p(A=1|B=1,E=e)p(B=1)p(E=e)}{\sum_b \sum_e p(A=1|B=b,E=e)p(B=b)p(E=e)} \approx 0.99.$$

- Note: even further re-ordering would have been possible.
- Note 2: People are not in general very good at estimating this kind of probabilities. Could you have come up with the result approximately from the given probabilities, without actually doing the computations?

Message passing - a simple example (1/2)*



- Compute marginal $p(a = 0)$ in the given graph

$$p(a=0) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} \sum_{d\in\{0,1\}} p(a=0,b,c,d)$$

$$= \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} \sum_{d\in\{0,1\}} p(a=0|b)p(b|c)p(c|d)p(d)$$

Naive computation: summation of $2^{T-1} = 8$ terms

Message passing - a simple example (2/2)*

- A more efficient approach is to eliminate one variable at a time:

$$p(a=0) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a=0|b)p(b|c) \underbrace{\sum_{d\in\{0,1\}} p(c|d)p(d)}_{\gamma_d(c)}$$

$$= \sum_{b\in\{0,1\}} p(a=0|b) \underbrace{\sum_{c\in\{0,1\}} p(b|c)\gamma_d(c)}_{\gamma_c(b)}$$

$$= \sum_{b\in\{0,1\}} p(a=0|b)\gamma_c(b)$$

Computational cost: $2 \times (T-1) = 6$ summations

- *Variable elimination*: eliminate variables starting from the end of the chain (or a leaf of a tree)
- Pass a message (information) from the eliminated variable to its neighbor in the chain.

Main points

- The definition of Bayesian networks
- Reading (conditional) independendies using d-separation
- Understanding why it is important to select the model (network) structure appropriately
- Computation of marginal and conditional distributions using a BN