Advanced Probabilistic Machine Learning
SSY316

Graphical Models (1)

Alexandre Graell i Amat

alexandre.graell@chalmers.se

https://sites.google.com/site/agraellamat

November 21, 2023

CHALMERS

Graphical models

A powerful framework to represent and learn structured probabilistic models

Capture way in which a joint distribution can be factorized into product of factors, each depending only on a subset of variables

Useful for

- Visualize the structure of a probabilistic model
- Encode structural information (dependencies) about involved RVs
- Structure computations: provide graph-based algorithms for computation, inference, and forecasting

Graphical models

- Image Processing
- Speech Processing
- Natural Language Processing
- Document Processing
- Pattern Recognition
- Bioinformatics
- Computer Vision
- Economics
- Physics
- Social Sciences
- ...

Graphical models

Three types of graphical models:

- Bayesian networks (directed acyclic graphs): Represent a set of RVs and their conditional dependence structure
- Markov random fields (undirected graphs): Represent a set of RVs and their Markov structure
- Factor graphs: More convenient for the purposes of inference and learning

Bayesian networks (Bayes nets)

Bayesian networks: encode a factorization of a joint distribution.

Two types of elements:

- Nodes: represent RVs
  - Empty nodes: unobserved RVs
  - Shaded nodes: observed RVs
- Edges: represent relationships between RVs

Bayesian networks

Bayesian networks: encode a factorization of a joint distribution.

K RVs $\{\mathsf{x}_1, \ldots, \mathsf{x}_K\}$ with joint probability distribution $p(x_1, \ldots, x_K)$

Chain rule:

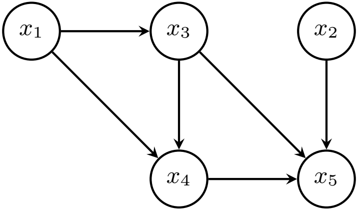
$$\begin{aligned} p(x_1, \ldots, x_K) &= p(x_1)p(x_2|x_1) \cdots p(x_K|x_1, \ldots, x_{K-1}) \\ &= \prod_{k=1}^{K} p(x_k|x_1, \ldots, x_{k-1}) \end{aligned}$$

To build BN:

1. Introduce a node for each x; associate node with $p(x|\cdot)$
2. For each $p(x|\cdot)$ draw a directed edge to node x from nodes corresponding to RVs on which the distribution is conditioned
3. Edge from node x to node $\tilde{x}$: x parent of $\tilde{x}$, $\tilde{x}$ child of x

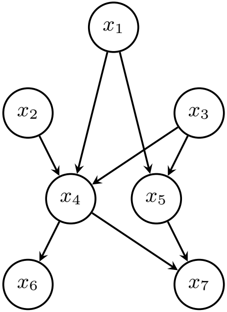
Bayesian networks
(from joint distribution to BN)

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_3)p(x_5|x_2, x_3, x_4)$$



1. Introduce a node for each x; associate node with $p(x|\cdot)$
2. For each $p(x|\cdot)$ draw a directed edge to node x from nodes corresponding to RVs on which distribution is conditioned

Bayesian networks (from BN to joint distribution)



$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Factorization into conditional distributions given by structure of BN.

Bayesian networks

Bayesian network: A directed acyclic graph (DAG) whose nodes represent RVs $\{\mathsf{x}_1, \ldots, \mathsf{x}_K\}$ with an associated joint distribution that factorizes as

$$p(x_1, \ldots, x_K) = \prod_{k=1}^{K} p(x_k | x_{\mathcal{P}(x_k)})$$

$\mathcal{P}(x_k)$: set of indices of parents of node x_k in the DAG.

- $\mathsf{x}_{\mathcal{P}(\mathsf{x}_k)}$: Account for statistical dependence of x_k with all the preceding variables $\mathsf{x}_1, \ldots, \mathsf{x}_{k-1}$ according to selected order
- BN encodes

$$\mathsf{x}_k \perp \{\mathsf{x}_1, \ldots, \mathsf{x}_{k-1}\} \backslash \mathcal{P}(\mathsf{x}_k)$$

Example: Bayesian polynomial regression

Goal: Make predictions for target variable t given some new value x.

- Training data set $\mathcal{D} = (x_{\mathcal{D}}, t_D) = \{(x_1, t_1), \ldots, (x_N, t_N)\}$

Model:

For $\boldsymbol{\phi}(x) = (1, x^2, \ldots, x^M)^\mathsf{T}$,

$$\begin{aligned} t &= \boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \beta^{-1}) \\ \boldsymbol{w} &\sim p(\boldsymbol{w}) \end{aligned}$$

Equivalently,

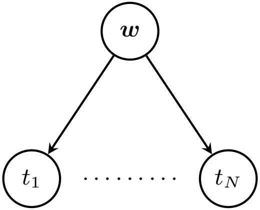
$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|\mu(x, \boldsymbol{w}), \beta^{-1})$$

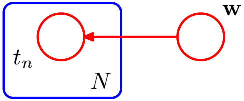
RVs: $\boldsymbol{w}$ and $t_{\mathcal{D}}$
Parameters: $(x_1, \ldots, x_N)$, σ^2, α

Example: Polynomial regression

Joint distribution $p(t_{\mathcal{D}}, \boldsymbol{w})$:

$$p(\boldsymbol{t}_{\mathcal{D}}, \boldsymbol{w}) = p(\boldsymbol{t}_{\mathcal{D}}|\boldsymbol{w})p(\boldsymbol{w}) = p(\boldsymbol{w}) \prod_{i=1}^{N} p(t_i|\boldsymbol{w})$$

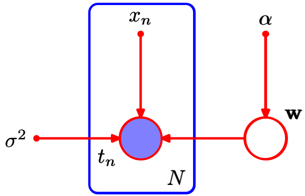




Example: Polynomial regression

- Joint distribution conditioned on input data and model parameters,

$$p(t_{\mathcal{D}}, \boldsymbol{w} | x_{\mathcal{D}}, \alpha, \sigma^2) = p(\boldsymbol{w}|\alpha) \prod_{i=1}^{N} p(t_i | \boldsymbol{w}, x_i, \sigma^2)$$



- Deterministic parameters represented by smaller solid circles
- Some RVs are observed $\longrightarrow$ shaded nodes
- Unobserved RVs: latent or hidden RVs (e.g., $\boldsymbol{w}$)

Bayesian networks

When are they useful?

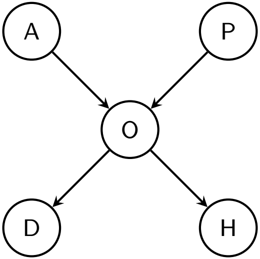
- When can identify causality relationships among RVs $\longrightarrow$ a natural order on variables; RVs that appear later caused by subset of preceding variables

Causing RVs for RV x_k included in $\mathcal{P}(x_k)$ $\longrightarrow$ when conditioning on $\mathsf{x}_{\mathcal{P}(\mathsf{x}_k)}$, x_k independent on all other preceding RVs.

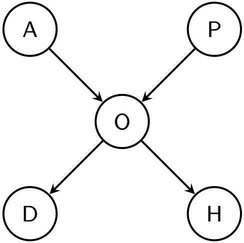
Bayesian networks: Example

- Lack of physical activity (A) and bad dietary patterns (P) cause obesity (O)
- Obesity causes diabetes (D) and heart disease (H)

Bayesian network:



Bayesian networks: Example



Joint distribution:

$$p(\mathsf{A},\mathsf{P},\mathsf{O},\mathsf{D},\mathsf{H}) = p(\mathsf{A})p(\mathsf{P})p(\mathsf{O}|\mathsf{A},\mathsf{P})p(\mathsf{D}|\mathsf{O})p(\mathsf{H}|\mathsf{O})$$

Marginal distribution:

$$p(\mathsf{H}) = \int p(\mathsf{A},\mathsf{P},\mathsf{O},\mathsf{D},\mathsf{H})\mathsf{d}a\,\mathsf{d}p\,\mathsf{d}o\,\mathsf{d}d$$

Ancestral sampling

Problem: Obtaining marginals is not easy

Idea: Draw samples from a given probability distribution.

- Easy to draw samples from a BN! $\longrightarrow$ Ancestral sampling

Ancestral sampling

Assume:

$$p(x_1, \ldots, x_K) = \prod_{k=1}^{K} p(x_k | x_{\mathcal{P}(x_k)})$$

(ordered variables $\{\mathrm{x}_1, \ldots, \mathrm{x}_K\}$, with no arrow from any node to any lower numbered node)

Goal: Draw samples from $p(x_1, \ldots, x_K)$

Ancestral sampling:

1. Draw sample for $\mathrm{x}_1 \sim p(x_1)$
2. Draw sample for $\mathrm{x}_2 \sim p(x_2 | x_{\mathcal{P}(2)})$
3. ...
4. Draw sample for $\mathrm{x}_K \sim p(x_K | x_{\mathcal{P}(K)})$

We have obtained a sample from the joint distribution.

Ancestral sampling

Sample from a marginal distribution:

- Take sampled values for required nodes and ignore those for remaining ones
- $p(x_2, x_4)$: sample from $p(x_1, \ldots, x_K)$, retain $\hat{x}_2, \hat{x}_4$ and discard $\{\hat{x}_{j \neq 2,4}\}$

Conditional independence

- Two RVs a and b are conditionally independent given c if

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

We write $a \perp b|c$

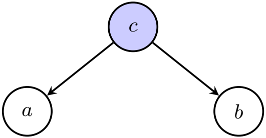
- Can also be written as

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

Graphical models: Conditional independence properties can be read directly from graph (d-separation).

Conditional independence

Tail-to-tail nodes:

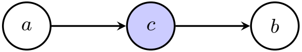


$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c) \implies \mathsf{a} \perp \mathsf{b}|\mathsf{c}$$

a and b independent if node in between is observed

Conditional independence

Head-to-tail nodes:

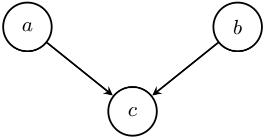


$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c) \implies \mathsf{a} \perp \mathsf{b}|\mathsf{c}$$

a and b independent if node in between is observed

Conditional independence

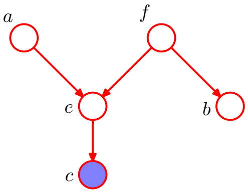
Head-to-head nodes:



$$p(a,b) = \int p(a,b,c)dc = \int p(c|a,b)p(a)p(b)dc = p(a)p(b) \quad \Longrightarrow \quad \mathsf{a} \perp \mathsf{b}|\emptyset$$

a and b independent if node in between and any of its descendants are not observed

d-separation



- Are a and b conditionally independent given c?
- More in general: Is a given subset of variables $\mathcal{A}$ independent of another set $\mathcal{B}$ conditioned on a third subset $\mathcal{C}$? ($\mathsf{x}_{\mathcal{A}} \perp \mathsf{x}_{\mathcal{B}} | \mathsf{x}_{\mathcal{C}}$)

Goal: Determine independencies directly from the directed acyclic graph.

d-separation: If all paths from from any node in $\mathcal{A}$ to any node in $\mathcal{B}$ given the nodes in $\mathcal{C}$ are blocked, then $\mathcal{A} \perp \mathcal{B} | \mathcal{C}$

d-separation

Previous examples giving blocked paths:

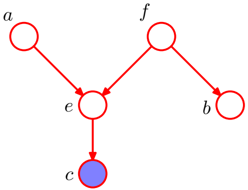
- observed tail-to-tail nodes
- observed head-to-tail nodes
- unobserved head-to-head nodes (with unobserved descendants)

d-separation: Let G be a directed graph and $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ disjoint sets of nodes (RVs). Then, if all paths from any node in $\mathcal{A}$ to any node in $\mathcal{B}$ given the nodes in $\mathcal{C}$ are blocked, $\mathcal{A}$ and $\mathcal{B}$ are said to be d-separated by $\mathcal{C}$ and $\mathcal{A} \perp \mathcal{B} | \mathcal{C}$.

A path between $\mathcal{A}$ and $\mathcal{B}$ is blocked if the path includes either:

- A head-to-tail or tail-to-tail node which is in $\mathcal{C}$; or
- A head-to-head node, and neither the node nor any of its descendants in $\mathcal{C}$

d-separation

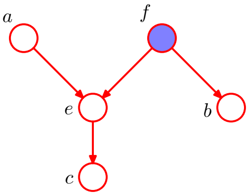


Are a and b conditionally independent given c?

- Path from a to b not blocked by f, since a tail-to-tail node and f unobserved
- Path not blocked by e, as a head-to-head node with an observed descendant

Hence, $a \perp b | c$ does not follow from the DAG

d-separation

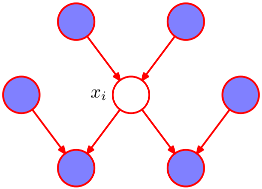


Are a and b conditionally independent given f?

- Path from a to b blocked by f, since a tail-to-tail node and observed
- Path blocked by e, as a head-to-head node and neither it its descendants are observed

Hence, $a \perp b|c$ follows from the DAG

Markov blanket



In a directed graphical model: A node is conditionally independent of all other nodes given its parents, children, and co-parents.

Markov blanket of node x_i: The minimal set of nodes that isolates x_i from the rest of the graph.

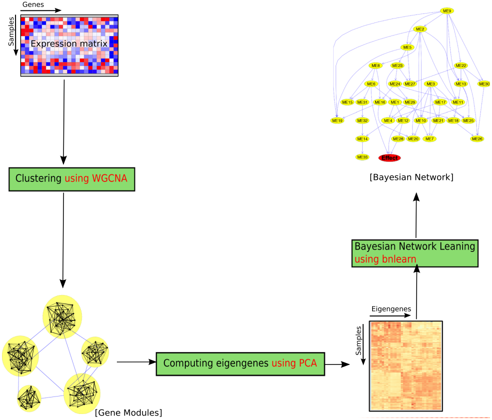
Structure Learning

What if the Bayesian network is not known?

Bayesian networks can be learned from data without a pre-specified structure:

- Different algorithms can be employed to learn network structure by analyzing the data and inferring most likely graph structure that best fits observed dependencies.
- Once structure and parameters are learned, Bayesian networks can be used for prediction

Predicting blood disease from gene expression profile



R. Agrahari *et al.*, "Applications of Bayesian network models in predicting types of hematological malignancies," Scientific Reports, 2018

Bayesian networks: Examples

M. Berkan Sesen *et al.*, “Bayesian Networks for Clinical Decision Support in Lung Cancer Care,” PLoS ONE, 2013.

A. Greppi, M. De Giuli, C. Tarantola, ‘Bayesian networks for stock picking,” 2013.

F. B. Hatipoglu, U. Uyar, “Examining the dynamics of macroeconomic indicators and banking stock returns with bayesian networks, *Business and Economics Research Journal*, 2019.

Y. Zuo, E. Kita, “Stock price forecast using Bayesian network,” Elsevier, 2012.

Reading

"Pattern recognition and machine learning,"

Chapter 8 (Intro, 8.1 (until 8.1.2), 8.2)