Lecture 7

Probabilistic Graphical Models

Instructor: Shibo Li

shiboli@cs.fsu.edu

Department of Computer Science
Florida State University

Overview

- A marriage between the graph theory and probability theory: it uses graphs to represent probabilistic models and facilitate inference
- The graphical structures reflect the conditional independency of the model (intuitive, convenient and expressive for modeling)
- The inference relies on the graphical structures (easy to implement, apply, analyze and improve)
- Neural networks are instances of graphical models

Outline

- Bayesian networks
  - Graphical representation
  - Conditional independence
  - D-separation, Bayes ball algorithm
  - Markov blanket
- Markov random field
  - Conditional independence
  - Relation to directed graphs
- Inference
  - Factor-graphs
  - Sum-product algorithm
  - Max-product, max-sum algorithms

Outline

- Bayesian networks
- Markov random fields
- Inference

- Bayes' Rule (theorem) revisited

$$p(\mathbf{x}_2|\mathbf{x}_1) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_1)}$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)\ldots p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$$

Why?

The decomposition of the joint probability defines a sampling procedure. We sequentially sample each variable given the previously sampled ones

- Consider a probabilistic model over 3 random variables: *a,b,c*

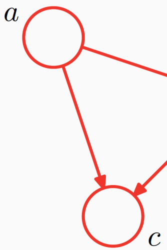
$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

- Question: can we use a graph to represent their joint probability?

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

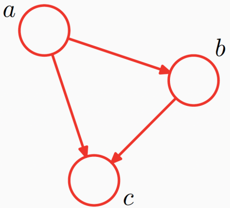
- Question: can we use a graph to represent their joint probability?

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



- Given the joint probability,
  - Use a node to represent each random variable (RV)
  - For each conditional distribution in the joint probability, $p(a|b_1,\ldots, b_m)$, add an edge from each b_i to a ($1 \leq i \leq m$). The RVs in the condition parts are represented as the parents
  - If no condition parts, the node has no parents

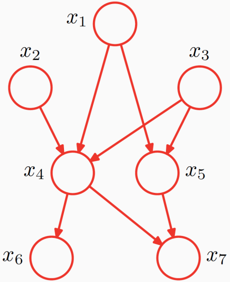
$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$



Bayesian Networks – representation

- Another example

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5).$$



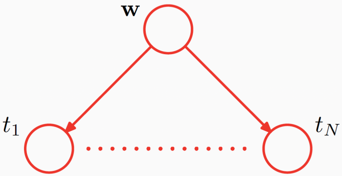
- We name this representation as a Bayesian network
- Bayesian networks must be a Directed Acyclic Graphs (DAG)! Why?

- We name this representation as a Bayesian network
- Bayesian networks must be a Directed Acyclic Graphs (DAG)! Why?

A cycle means each random variable (RV) can be sampled only if *all* the other RVs in the cycle have been sampled. That means, the RVs in the cycle cannot be sequentially sampled. This violates Bayes' Rule, since Bayes' Rule guarantees all the random variables can be sequentially sampled via the joint probability decomposition.

- Polynomial regression

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



Bayesian Networks

- How to be more specific and succinct?

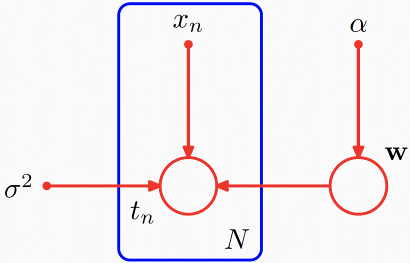
$$\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbf{I}) \qquad \mathcal{N}(t_n|\sum_{j=0}^{d-1} w_j x_n^j, \sigma^2)$$

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{\boxed{N}} p(t_n|\mathbf{w}, x_n, \sigma^2)$$

parameters

observations

Bayesian Networks

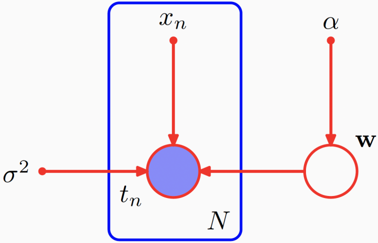


Small solid nodes: deterministic parameters, uninterested observations

Big empty nodes: latent random variables

Plate with label *N*: *N* replicates

- In the training data, the outputs have been observed



Shaded nodes: observed latent variables

- The network structure is determined by the factorization of the joint probability; different factorization leads to different structures

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

What are the networks?

So, equivalent models may have different structures

Bayesian Networks - Notes

- How to design the factorization of the joint probability is the key of the probabilistic modeling.
- Using the full Bayes formula will lead to a fully connected network, which represents the most general modelling (without any assumptions). But this is not what we want.
- For probabilistic modeling, we nearly always use domain knowledge to simplify the joint probability, which can be reflected by the network structure. The simplification is called conditional independence.

- Linear Gaussian model

- For multivariate Gaussian variables $x_1, \ldots, x_N$

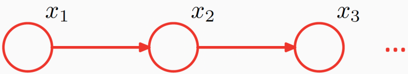
$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i, v_i\right)$$

Question1: what is the network structure if we do not make any assumption? Fully connected

Question2: How many parameters do we need to estimate? $O(N^2)$

Bayesian Networks

- Linear Gaussian model: Let us choose a chain structure



$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i, v_i\right)$$

Question2: How many parameters do we need to estimate? O(N)

- In general, the simplification of the Bayes‘ Rule reflects our ideas, tricks and knowledge in probabilistic modeling

- How is the simplification reflected?

Conditional independence!

Conditional Independence

- Consider a probabilistic model over 3 random variables: *a,b,c*

a is conditional independent of b given c if

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

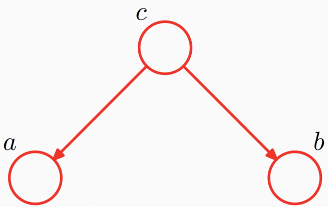
Why?

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence

- What is the Bayesian network?

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

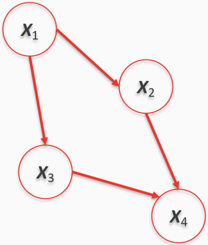


The network structure is simplified as well

Conditional Independence

- Practically , how do we design a Bayesian network?

Consider a sampling (generative) process



We usually do not explicitly consider all possible conditional independences!

...

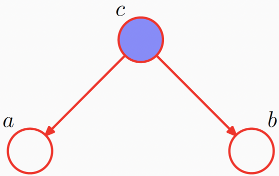
Conditional Independence

- Question: For a (complex) Bayesian network, given arbitrary nonintersecting sets of nodes *A, B, C*, how do we test the conditional independency?

$$A \perp\!\!\!\perp B \mid C$$

- This is important to analyze our model

- Basic case I: *tail-to-tail*



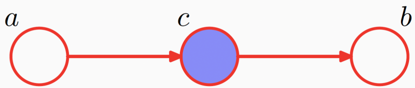
$$a \not\perp\!\!\!\perp b \mid \emptyset$$

$$a \perp\!\!\!\perp b \mid c$$

Why?

D-separation

- Basic case II: *head-to-tail*



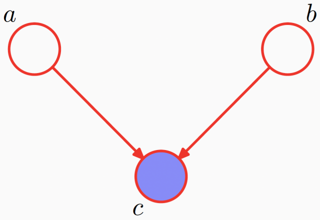
$$a \not\perp\!\!\!\perp b \mid \emptyset$$

$$a \perp\!\!\!\perp b \mid c$$

Why?

D-separation

- Basic case III (a little odd): ***head-to-head***



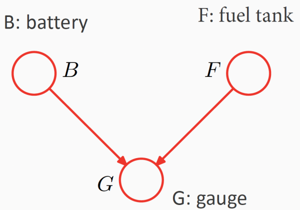
$$a \perp\!\!\!\perp b \mid \emptyset$$

$$a \not\!\perp\!\!\!\perp b \mid c$$

Why?

D-separation

- *head-to-head:* explain away effect

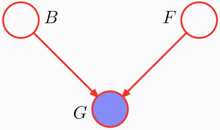


$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9.
\end{aligned}
$$

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

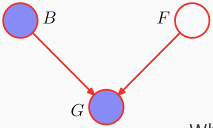
D-separation

- *head-to-head:* explain away effect



$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \simeq 0.257$$

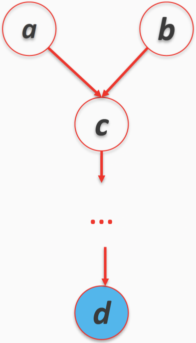
>



$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \simeq 0.111$$

Why? Batter being dead partly takes away the effect of zero Gauge

- *head-to-head: more general case*



$a \not\perp\!\!\!\perp b \mid$ $\boxed{c \text{ or any descendent}(c)}$

D-separation

- In general, for a (complex) Bayesian network, given arbitrary nonintersecting sets of nodes *A, B, C,* how to test the conditional independency?

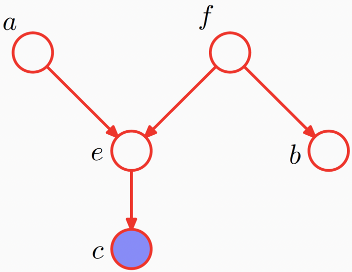
$$A \perp\!\!\!\perp B \mid C$$

D-separation (Bayes ball algo.)

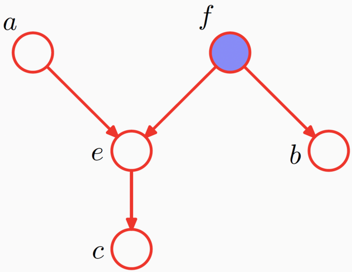
- Step 1: Shade all the nodes in *C*
- Step 2: For every path from any node in A to any node in B
  - If the path contains a node, such that
    - the arrows on the path meet ***head-to-tail*** or ***tail-to-tail*** at a node in C, the path is blocked and continue, OR
    - the arrows on the path meet head-to-head at a node, and neither the node or any of its descendent is in C,

    the path is blocked and continue
  - Otherwise, return $A \perp\!\!\!\perp B \mid C$ does not hold
- Step 3: if every path is blocked, return $A \perp\!\!\!\perp B \mid C$ holds

D-separation examples



A = {a}, B = {b}, C = {c}



A = {a}, B = {b}, C = {f}

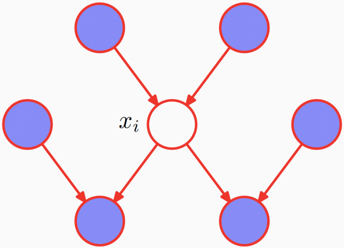
Markov-blanket

- Consider a Bayesian network with D nodes, $\boldsymbol{x}_1$, …, $\boldsymbol{x}_D$
- For a particular node $\boldsymbol{x}_i$, conditioned on what set of variables, $\boldsymbol{x}_i$ are independent to the remaining variables?

$$
\begin{aligned}
p(\mathbf{x}_i | \mathbf{x}_{\{j \neq i\}}) &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_D)}{\int p(\mathbf{x}_1, \ldots, \mathbf{x}_D) \, \mathrm{d}\mathbf{x}_i} \\
&= \frac{\prod_k p(\mathbf{x}_k | \mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k | \mathrm{pa}_k) \, \mathrm{d}\mathbf{x}_i}
\end{aligned}
$$

Markov-blanket

- Answer: $\boldsymbol{x}_i$'s parents, $\boldsymbol{x}_i$'s children and the children's co-parents
- These variables are called the Markov-blanket of $\boldsymbol{x}_i$

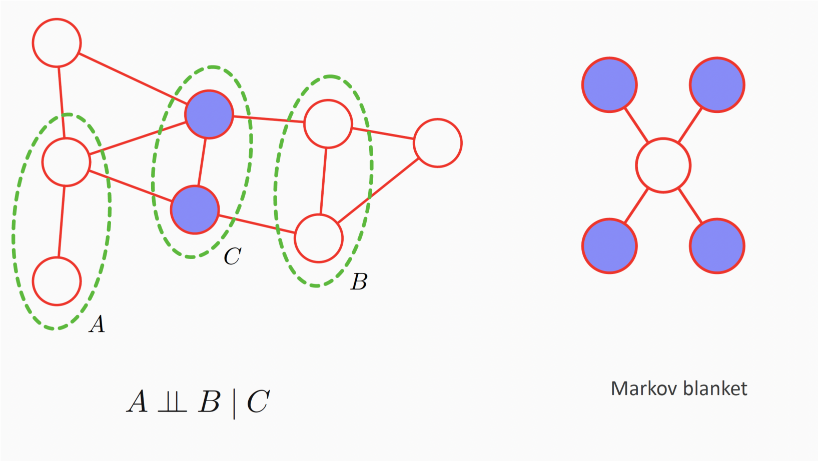


Some thoughts

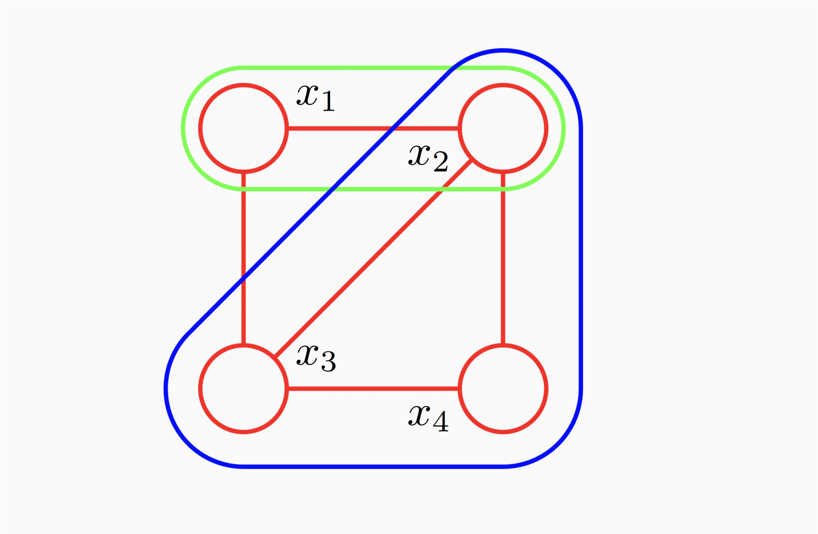
- D-separation is a bit subtle to test the conditional independency

- Can we have easier graphical representations that allow more natural tests? e.g., only based on paths without considering arrow directions?

Markov Random Fields



Cliques and Maximum Cliques



Joint Distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(\mathbf{x}_C)$$

Where $\psi_C(\mathbf{x}_C) \geqslant 0$ is the *potential function* over maximum clique C

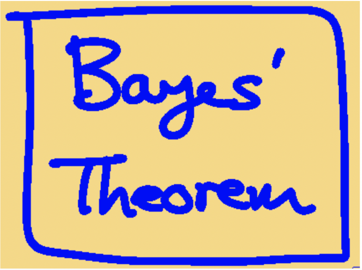
$$Z = \sum_{\mathbf{x}} \prod_{C} \psi_C(\mathbf{x}_C)$$

is the normalization constant, also called *partition* function

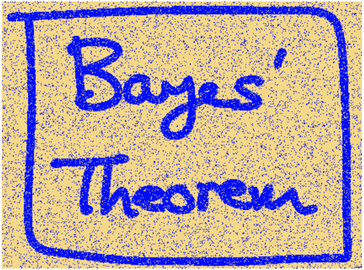
Energy and the Boltzmann distribution

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

Illustraion: Image Denoise

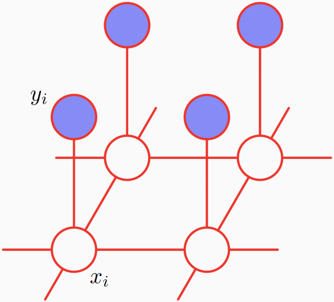


Ground-truth



noisy observation

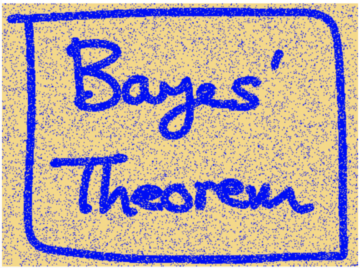
Illustraion: Image Denoise

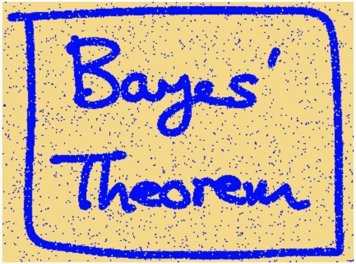


$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

Illustraion: Image Denoise

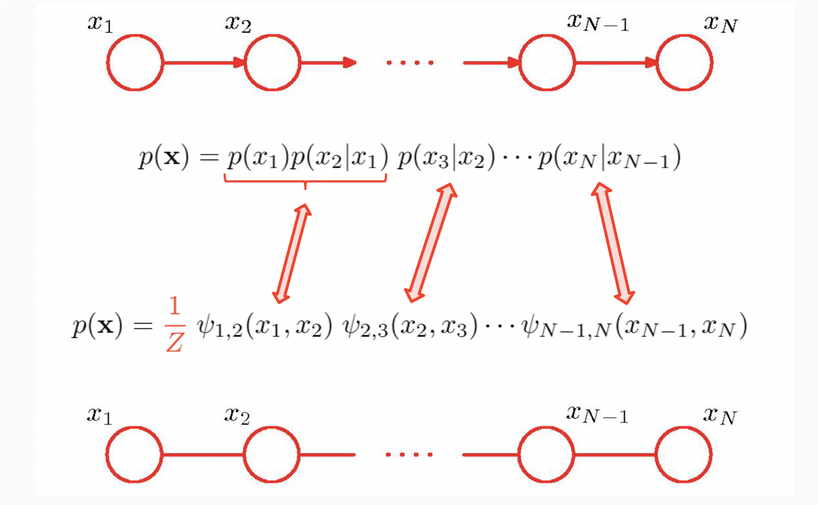




noisy observation

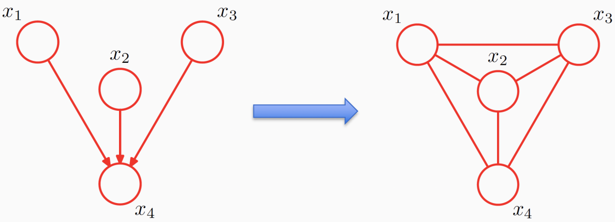
restored version (ICM)

Convert Directed to Undirected Graphs



Convert Directed to Undirected Graphs

Add additional links: “marrying parents”, i.e., moralization



$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) = \psi(x_1, x_2, x_3, x_4)$$

What you need to know

- How to construct Bayes networks and Markov random field
- How to convert a BN to MRF (moralization)
- BN is an acyclic directed graph, why? (Bayes' Rule)
- Conditional independence
- Head-to-tail, tail-to-tail and head-to-head
- Explain away effect
- D-separation (Bayes ball algorithm)
- BNs are NOT equivalent to MRFs!