CS 8501 Advanced Topics in Machine Learning

Lecture 04: Directed Graphical Models

Yangfeng Ji

Information and Language Processing Lab

Department of Computer Science

University of Virginia

https://yangfengji.net/

Introduction

Central Questions

- How can we compactly represent the joint distribution $p(x \mid \theta)$?
- How can we use this distribution to infer one set of variables given another in a reasonable amount of computation time?
- How can we learn the parameters of this distribution with a reasonable amount of data?

Number of Parameters

Assume each x_i is a random variable with T possible values, how many parameters that we need to represent the following distribution?

$$p(x_{1:V})$$

Example

Consider a joint distribution on (x_1, x_2, x_3), where each $x_i \in \{1, \dots, T\}$

- Without any assumption, we need each θ_{ijk} represent a specific probability of

$$p(x_1 = i, x_2 = j, x_3 = k) = \theta_{ijk}$$

- In total, we need $T^3 - 1$ parameters $\theta = \{\theta_{ijk}\}$
  - Because $\sum_i \sum_j \sum_k \theta_{ijk} = 1$

Factorization

What if there is no independence assumption, and just factorize the distribution as

$$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)?$$

- $p(x_1)$: $T - 1$ parameters
- $p(x_2 \mid x_1)$: $T(T-1)$ parameters
- $p(x_3 \mid x_1, x_2)$: $T^2(T-1)$ parameters
- In total, $T^3 - T^2 + T^2 - T + T - 1 = T^3 - 1$

Independence

If all three random variable are independent with each other, then we can factorize the joint distribution as

$$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2) \cdot p(x_3)$$

- Each $p(x_i)$ need $T - 1$ parameters
- In total, we need $3(T - 1)$ parameters

Efficient Representation

- The essence of efficient representation is *independence*
- In many cases, we need to exploit the (conditional) independence of random variables for efficient representation

Conditional Independence

X and Y are conditionally independent given Z, denoted

$$X \perp\!\!\!\perp Y | Z$$

if and only if the joint probability can be written as

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$$

Markov Chains

Consider the distribution $p(x_1, x_2, x_3)$ with $x_1 \perp\!\!\!\perp x_3 | x_2$, we have

$$p(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1)$$

with $p(x_3 \mid x_2, x_1) = p(x_3 \mid x_2)$ we have

$$p(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)$$

which is a *first-order* Markov chain

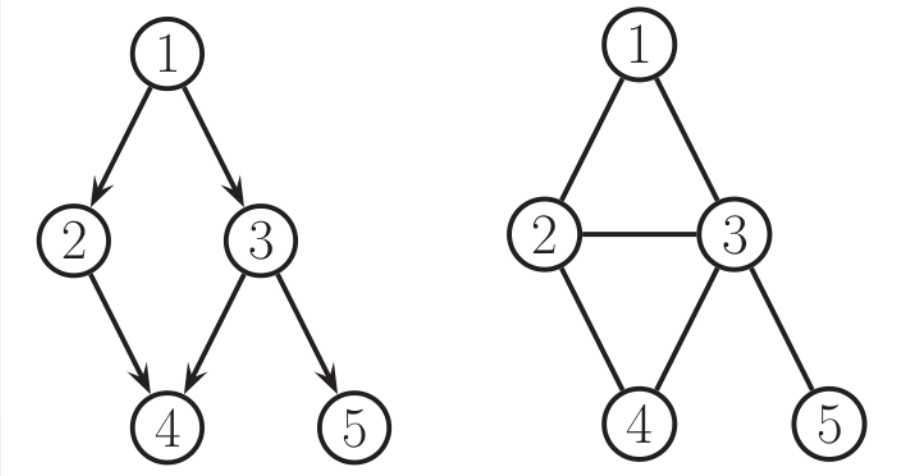
A first-order Markov chain with discrete values can be fully described by

- the initial state $p(x_1 = i)$, and
- the state transition matrix $p(x_t = j \mid x_{t-1} = i)$

Graphical Models

A graphical model is a way to represent a joint distribution with its conditional independence

- Based on the graphical properties, there are two kinds graphical models
  - Directed graphs: Bayes nets (this lecture)
  - Undirected graphs: Markov random fields (next lecture)



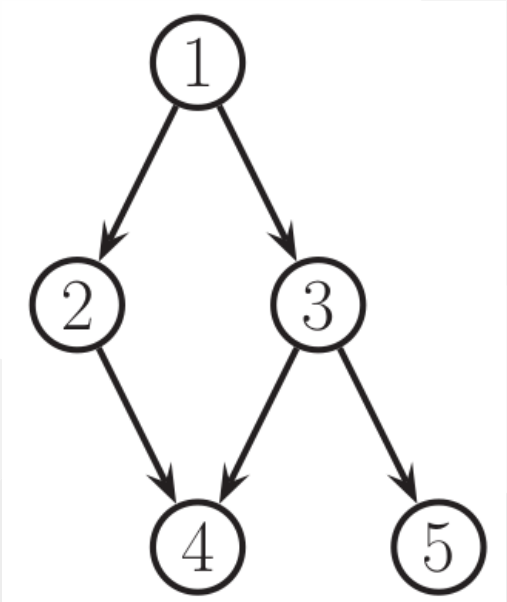
Graph Terminology

- Graph
- Nodes: parent nodes, children nodes, etc
- Edges
- Adjacency matrix
- Directed vs. undirected
- Cycle (or loop)
- Directed acyclic graph (DAG)

More in section 10.1.4

Topological Ordering

A topological ordering is a numbering of the nodes such that parents have lower numbers than their children.



Directed Graphical Models

Different names refer to the same thing

- Directed graphical models: *the most descriptive name*
- Bayesian networks (Bayes nets): *not related to Bayes' rule*
- Belief networks: *probability represents subjective belief*
- Causal networks: *directed arrows are sometimes interpreted as representing causal relations*

Ordered Markov Property

In a DAG, a node only depends on its immediate parents, not on all ancestors

$$x \perp\!\!\!\perp \boldsymbol{x}_{\mathrm{anc}(x)\setminus\mathrm{pa}(x)} \big| \boldsymbol{x}_{\mathrm{pa}(x)}$$

Consider the following Markov chain

$$\cdots \rightarrow x_{t-2} \rightarrow x_{t-1} \rightarrow x_t \rightarrow \cdots$$

In general, we have

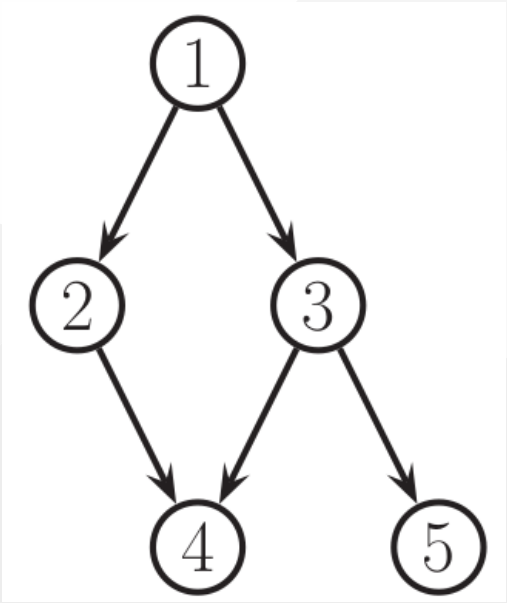
$$p_G(\boldsymbol{x}_{1:V}) = \prod_{t=1}^{V} p(x_t \mid \boldsymbol{x}_{\mathrm{pa}(t)})$$

Factorization

Recall that the conditional independence can help us simplify the factorization.

For the following running example, we have

$$p(\boldsymbol{x}_{1:5}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3)$$



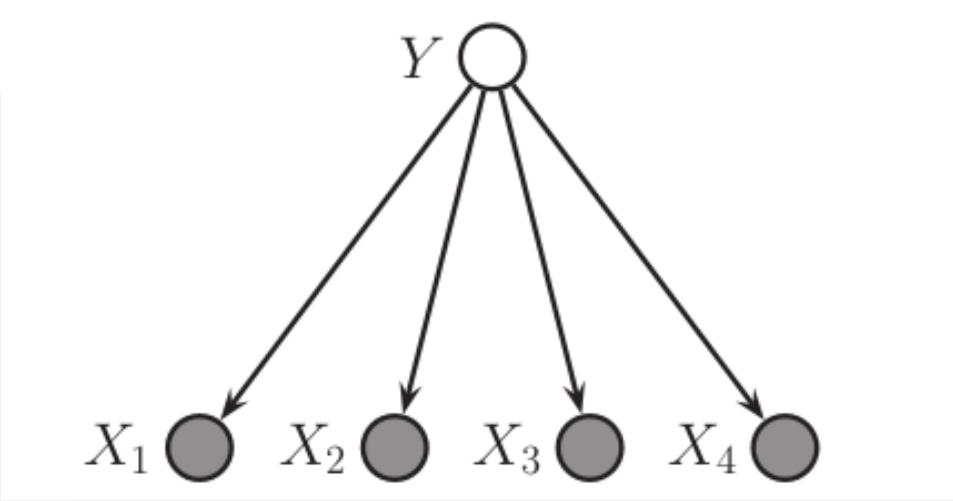
Examples

- This section focuses on graphical representations
- Inference will be discussed in the next section

Naive Bayes Classifiers

The graphical representation of $p(\boldsymbol{X}_{1:4}, Y)$

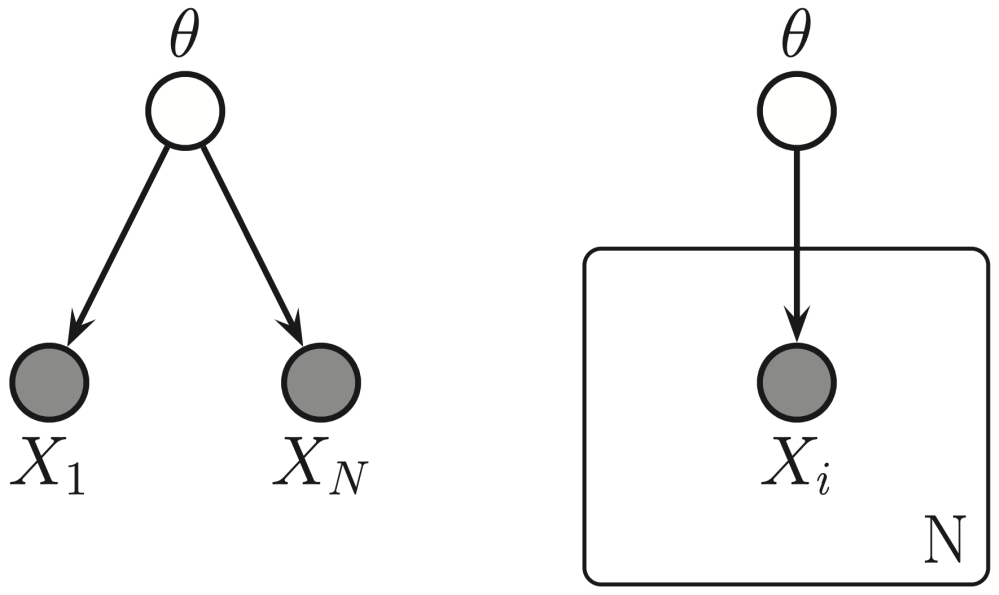
$$p(\boldsymbol{X}_{1:4}, Y) = p(Y) \prod_{t=1}^{4} p(X_t \mid Y)$$



Shaded notes are observed

Plate Notation

Plate notation is a useful graphical representation for conditionally IID examples



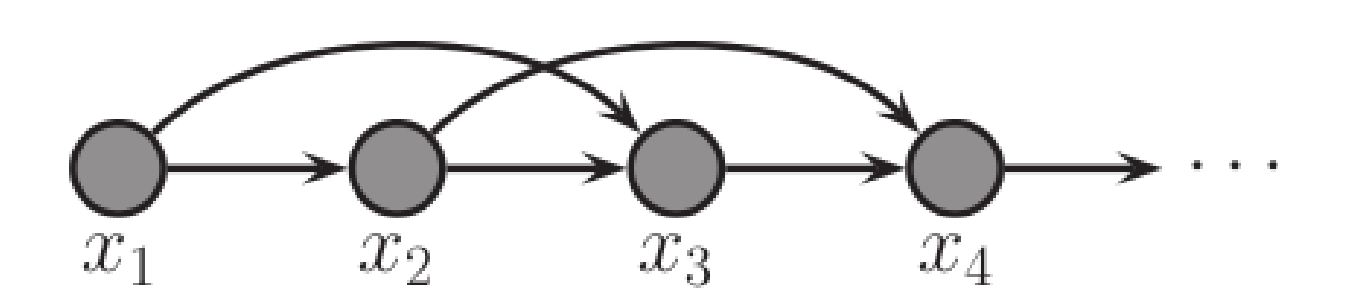
$$p(\theta, \boldsymbol{X}) = p(\theta)\{\prod_{i=1}^{N} p(X_i \mid \theta)\}$$

Markov Chains

First and second order Markov chains

Transition probability: Each component in the factorization, such as

- $p(X_t \mid X_{t-1})$ in the first order case
- $p(X_t \mid X_{t-1}, X_{t-2})$ in the second order case

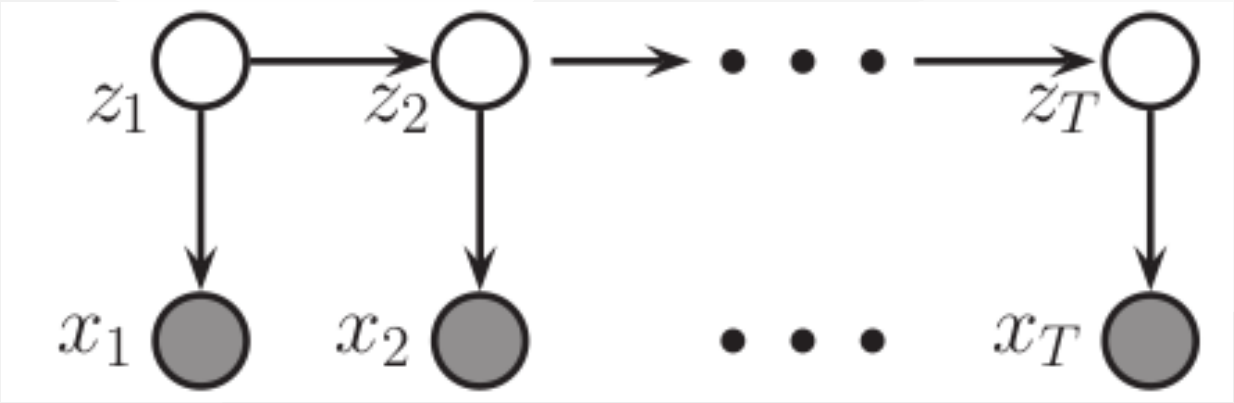


Hidden Markov Models

A first-order hidden Markov models

Two building blocks

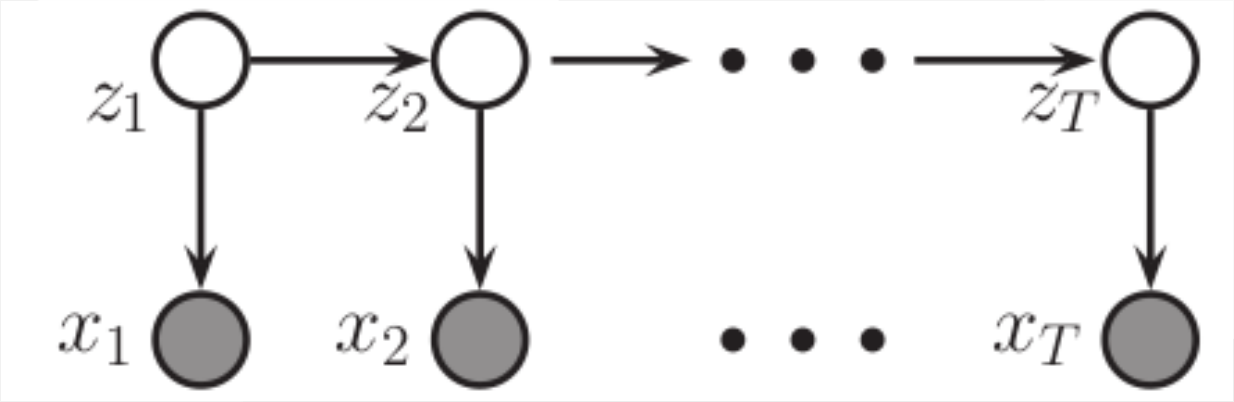
- Transition probability (or transition model): $p(z_t = j \mid z_{t-1} = i) = a_{ij}$
- Emission probability (or observation model): $p(x_t = k \mid z_t = j) = b_{jk}$



Hidden Markov Models (II)

With continuous observations

- Transition probability (or transition model): $p(z_t = j \mid z_{t-1} = i) = a_{ij}$
- Emission probability (or observation model): $p(x_t = k \mid z_t = j) = \mathcal{N}(\mu_j, \sigma_j^2)$

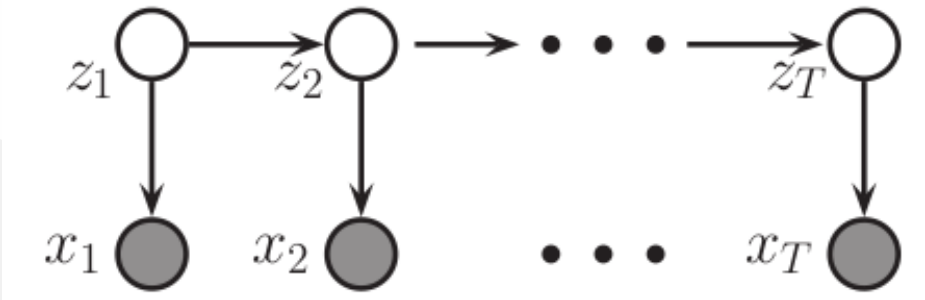


More content about hidden Markov models: [Murphy, 2012; Chapter 17]

State Space Models

Continuous variables on both hidden states and observations

- Transition model: $z_t = g(u_t, z_{t-1}, \varepsilon_t)$
- Observation model: $x_t = h(z_t, u_t, \delta_t)$

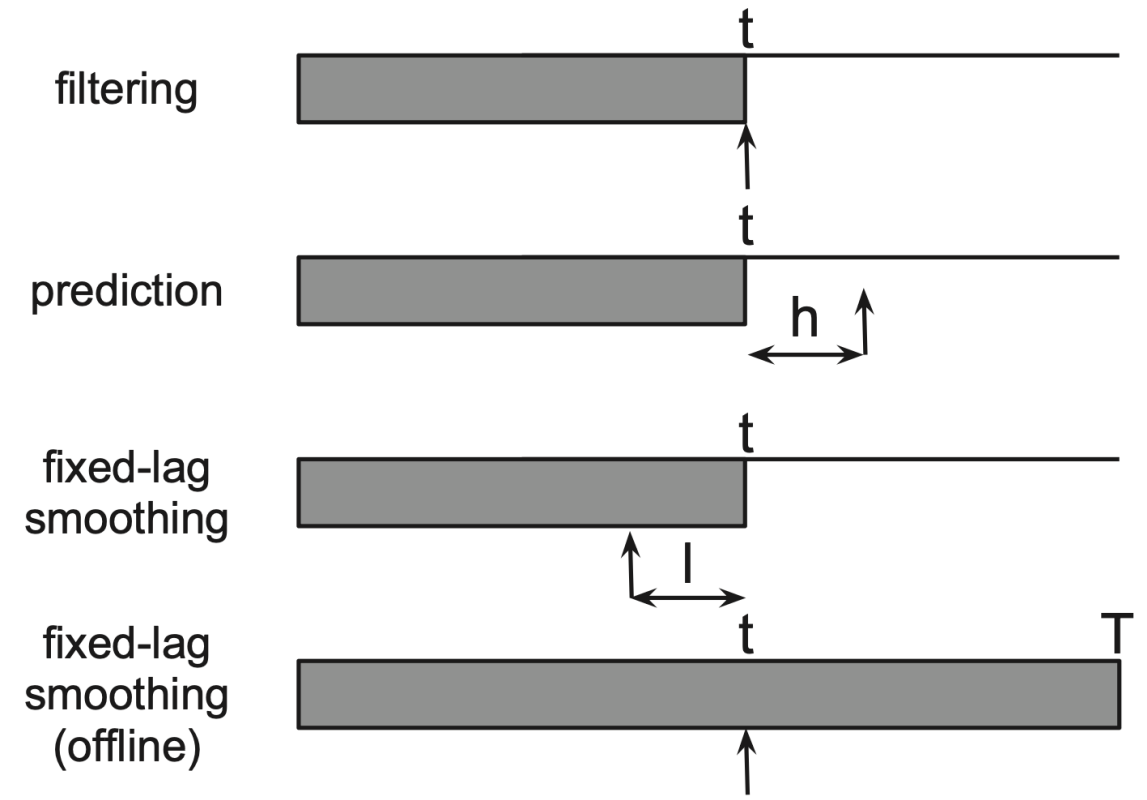


Linear dynamic systems:

- $z_t = A_t z_{t-1} + B_t u_t + \varepsilon_t$
- $x_t = C_t z_t + D_t u_t + \delta_t$

State Space Models: Predictions

Three different types of predictions in State Space Models



More content about state space models: [Murphy, 2012; Chapter 18]

Dynamic Bayesian Networks

- Discrete-state DBN
  - HMMs
  - Factorial HMMs
  - Hierarchical HMMs
  - ...
- Continuous-state DBN
  - KFM
  - Switching KFM
  - ...

More information: [Murphy 2002, PhD Dissertation]

Inference

Inference

A typical task of inference is to estimate the conditional probability of hidden variables x_h given visible variable x_v

$$p(x_h \mid x_v, \theta) = \frac{p(x_h, x_v \mid \theta)}{p(x_v \mid \theta)} = \frac{p(x_h, x_v \mid \theta)}{\sum_{x'_h} p(x'_h, x_v \mid \theta)}$$

Computing $\sum_{x'_h} p(x'_h, x_v \mid \theta)$ is non-trivial

Discrete Random Variables

Consider the previous example, if the goal is to estimate $p(x_4 \mid x_1, x_2, x_3, x_5)$, then

$$p(x_1, x_2, x_3 \mid x_4, x_5) = \frac{p(x_1, x_2, x_3, x_4, x_5)}{p(x_4, x_5)}$$

A straightforward way of computation

$$p(x_4, x_5) = \sum_{x_1, x_2, x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3)$$

The number of summation operations: K^3 where K is the number of values for each random variable x_j

An Alternative Way

Another way of computation

$$p(x_4, x_5) = \sum_{x_2, x_3} \{ \sum_{x_1} p(x_1) p(x_2|x_1) p(x_3|x_1) \} p(x_4|x_2, x_3) p(x_5|x_3)$$

- The number of summation $K + K^2$
- The benefit will be more significant, if the probability distribution has large K, many random variables, and sparse dependency

Sum-product Algorithm

Essentially, the idea in the following formula is **switching the sum and product operations** based on dependency. Therefore, it is also called the **sum-product algorithm**.

$$p(x_4, x_5) = \sum_{x_2, x_3} \{ \sum_{x_1} p(x_1) p(x_2|x_1) p(x_3|x_1) \} p(x_4|x_2, x_3) p(x_5|x_3)$$

- A general version of this algorithm can be used to compute conditional probability directly
- Another name of this algorithm: belief propagation

Last Comments

Variational inference and sampling methods offer two different ways to handle this problem

Learning

Foreword

There is no direct relation between Bayesian networks and Bayesian statistics.

Difference between Learning and Inference

In general, three categories of variables on graph

- x_v: visible variables
- x_h: hidden variables
- θ: model parameter as in $p(x_v, x_h \mid \theta)$

The difference between inference and learning

- Inference: estimate the probability of x_h given x_v
- Learning: estimate θ, usually a point estimate

MAP

A typical way of learning in graphical models is MAP

$$\hat{\theta} = \mathrm{argmax}_\theta \log p(\theta) + \sum_{i=1}^{N} \log p(x_v^{(i)} \mid \theta)$$

where

- i is the index of training examples
- $\log p(x_v^{(i)} \mid \theta)$ is the likelihood of *visible* variables only

$$\log p(x_v^{(i)} \mid \theta) = \log \sum_{x_h} p(x_v^{(i)}, x_h \mid \theta)$$

Marginal Likelihood

Computing the marginal likelihood

$$p(x_v;\theta) = \sum_{v_h} p(x_v, x_h;\theta)$$

is the challenge not only for **inference** but also **learning**.

Learning

- Learning from complete data: With all variable observed, we have likelihood

$$\log p(x;\theta) = \log \prod p(x_t \mid x_{\mathrm{pa}(x_t)};\theta) = \sum \log p(x_t \mid x_{\mathrm{pa}(x_t)};\theta)$$

- Learning with hidden variables: will be discussed in the future lectures

Conditional Independence

- Based on section 10.5

Three Basic Directed Graph Structures

Markov Chain

$$X \to Y \to Z$$

Conditional independence

- $X \not\perp\!\!\!\perp Z$
- $X \perp\!\!\!\perp Z | Y$

Three Basic Directed Graph Structures (II)

Common Cause

$$X \leftarrow Y \rightarrow Z$$

Conditional independence

- $X \not\perp\!\!\!\perp Z$
- $X \perp\!\!\!\perp Z | Y$
- The Beer and Diapers story

Three Basic Directed Graph Structures (III)

Explaining Away (also called the v-structure in the textbook)

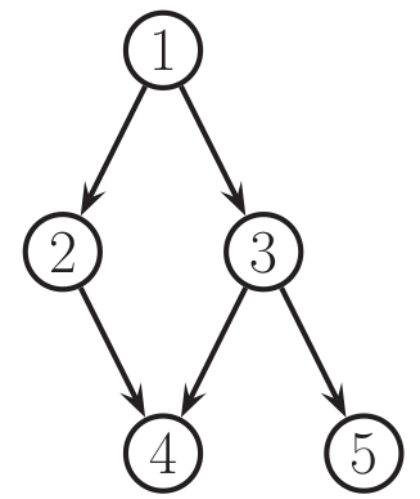
$$X \rightarrow Y \leftarrow Z$$

Conditional independence

- $X \perp\!\!\!\perp Z$
- $X \not\perp\!\!\!\perp Z | Y$
- One event can be caused by two reasons, the identification of one reason will reduce the probability about another reason happened.

Example

Reading independence from graph



- $X_2 \perp\!\!\!\perp X_5$?
- $X_2 \perp\!\!\!\perp X_5 | X_1$?
- $X_2 \perp\!\!\!\perp X_5 | X_1, X_4$?

Thank You!