1 - Directed probabilistic graphical model (DPGM)

Introduction

Directed probabilistic graphical models (DPGMs), also called **Bayesian networks**, provide a compact way to represent high-dimensional joint distributions by exploiting **conditional independence** structure.

- They offer a general and effective way to **parameterize** a joint distribution using **local conditional distributions**.
- The graph makes modeling assumptions (conditional independencies) explicit and often leads to **more efficient inference and learning**.
- Directed graphs can naturally encode **causal** hypotheses (when appropriate). Undirected models (MRFs) are often more natural when we only want to represent **symmetric dependencies** without committing to a direction of influence.

Recall the chain rule: any joint distribution can be written as

$$p\left(x_{1:n}\right) = \prod_{i=1}^{n} p\left(x_i \mid x_{1:i-1}\right).$$

A Bayesian network replaces the potentially huge conditioning set $x_{1:i-1}$ with a smaller set of "relevant" variables x_{A_i} :

$$p\left(x_i \mid x_{1:i-1}\right) = p\left(x_i \mid x_{A_i}\right).$$

Graphical representation

Let $G = (V, E)$ be a **directed acyclic graph (DAG)** whose nodes correspond to random variables $X_1, \ldots, X_N$. Because the graph is acyclic, it admits a topological ordering:

Definition

A **topological ordering** of the vertices of the graph is an ordering $1, \ldots, N_G$ such that $i \in \text{pa}(j)$ implies that $i < j$. Acyclicity ensures that a topological ordering always exists.

Definition

Consider each node represent a random variable. we say that they satisfies **ordered Markov property** if

$$x_i \perp \boldsymbol{x}_{\text{pred}(i) \setminus \text{pa}(i)} | \boldsymbol{x}_{\text{pa}(i)}$$

where

- $\text{pa}(i)$ are the parents of node i, and
- $\text{pred}(i)$ are the predecessors of node i in the ordering.

If the ordered Markov property holds, then the joint distribution factorizes as

$$p\left(\boldsymbol{x}_{1:N_G}\right) = p\left(x_1\right) p\left(x_2 \mid x_1\right) p\left(x_3 \mid x_1, x_2\right) \ldots p\left(x_{N_G} \mid x_1, \ldots, x_{N_G-1}\right) = \prod_{i=1}^{N_G} p\left(x_i \mid \boldsymbol{x}_{\text{pa}(i)}\right)$$

This factorization typically requires **far fewer parameters** than an unconstrained joint distribution, especially when each node has only a few parents.

Ancestral sampling: To sample from a Bayesian network prior, visit nodes in topological order and sample each node from its conditional given its parents:

$$x_i \sim p\left(x_i \mid x_{\text{pa}(i)}\right).$$

This yields i.i.d. samples $(x_1, \ldots, x_N) \sim p(x)$.

Gaussian Bayes nets

Consider a DPGM where all variables are real-valued and each conditional is linear-Gaussian:

$$p(x_i | \boldsymbol{x}_{\text{pa}(i)}) = \mathcal{N}(x_i | \mu_i + \boldsymbol{w}_i^\top \boldsymbol{x}_{\text{pa}(i)}, \sigma_i^2)$$

where $\boldsymbol{w}_i = (w_{i,j})_{j \in \text{pa}(i)}$ is the weight vector.

Now we want to show that the joint distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$:

Equivalently, we can write **structural equations**

$$x_i = \mu_i + \sum_{j \in \text{pa}(i)} w_{i,j}(x_i - \mu_i) + \sigma_i z_i, \qquad z_i \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

If we consider the vector form, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{N_G})$, then we have

$$(\boldsymbol{x} - \boldsymbol{\mu}) = \mathbf{W}(\boldsymbol{x} - \boldsymbol{\mu}) + \mathbf{S}\boldsymbol{z}$$

where

$$\mathbf{W} = \begin{pmatrix} 0 & & & & \\ w_{2,1} & 0 & & & \\ w_{3,1} & w_{3,2} & 0 & & \\ \vdots & & & \ddots & \\ w_{N_G,1} & w_{N_G,2} & \ldots & w_{N_G,N_G-1} & 0 \end{pmatrix}$$

(because of the topological ordering, so $\mathbf{W}$ is strictly lower-triangular) Now denote $\boldsymbol{e} := \mathbf{S}\boldsymbol{z}$ to be a vector of noise terms, then we can write

$$\boldsymbol{e} = (\mathbf{I} - \mathbf{W})(\boldsymbol{x} - \boldsymbol{\mu})$$

note that $\mathbf{I} - \mathbf{W}$ is invertible, so we denote $\mathbf{U} = (\mathbf{I} - \mathbf{W})^{-1}$ and can write

$$\boldsymbol{x} - \boldsymbol{\mu} = (\mathbf{I} - \mathbf{W})^{-1}\boldsymbol{e} = \mathbf{U}\boldsymbol{e} = \mathbf{U}\mathbf{S}\boldsymbol{z}$$

Hence the covariance is given by

$$\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{x}] = \mathrm{Cov}[\boldsymbol{x} - \boldsymbol{\mu}] = \mathrm{Cov}[\mathbf{U}\mathbf{S}\boldsymbol{z}] = \mathbf{U}\mathbf{S}\,\mathrm{Cov}[\boldsymbol{z}]\mathbf{S}\mathbf{U}^\top = \mathbf{U}\mathbf{S}^2\mathbf{U}^\top$$

Conditional independence (CI) and graph semantics

A conditional independence statement is written as

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B | \boldsymbol{x}_C$$

where $A, B, C \subseteq V$ are sets of nodes.

Definition

Let's denote

1. $I(G)$ to be the set of all CI statements encoded by the graph, i.e. A, B and C are sets of vertices.
2. $I(p)$ to be the set of all CI statements that hold true in some distribution p.

If p factorizes over G, then $I(G) \subseteq I(p)$. In this case, we say that G is an **I-map (independence map)** for p.

Even if G is an I-map, the inclusion can be strict: the distribution p may have additional independencies not reflected in G.

Markov properties and d-separation

Recall the definition of **global Markov property** for undirected graph:

Definition

- **(Separation)** Let $A, B, C \subseteq V$. We say that A and B are separated by C in G if every path from any $a \in A$ to any $b \in B$ contains at least one vertex in C.
- **(global Markov property)** We say that p satisfies the global Markov property for G if whenever a separation is present in G there is a corresponding conditional independence in p.

However the definition of global Markov property is different for directed graphs, consider the following **v-structure**:

$$s \to m \leftarrow t$$

since G satisfies ordered Markov property, so we can factorize:

$$p(s,t,m) = p(s)p(t)p(m|s,t)$$

divide both side by $p(m)$ we get:

$$p(s,t|m) = \frac{p(s)p(t)p(m|s,t)}{p(m)}$$

and in general we don't have $p(s,t|m) = p(s|m)p(t|m)$, so s and t are *not* conditionally independent given m. This is called explaining away, or Berkson's paradox.

Definition

For a directed acyclic graph G, we say a set of nodes A is **d-seperated** from a different set of nodes B given a third observed set C iff for each undirected path P from every

node $a \in A$ to every node $b \in B$, at least one of the following conditions hold:

1. P contains pipe, $s \to m \to t$ or $s \leftarrow m \leftarrow t$, where $m \in C$,
2. P contains a fork, $s \leftarrow m \to t$, where $m \in C$,
3. P contains a v-structure, $s \to m \leftarrow t$, where m is not in C and neither is any descendant of m.

Definition

We say p satisfies **global Markov property** for DAG G if for every sets of variables A, B, C:

$$\mathbf{X}_A \perp_G \mathbf{X}_B | \mathbf{X}_C \iff A \text{ is d-seperated from } B \text{ given } C$$

Definition

We say p satisfies **local Markov property** for G if

$$x_i \perp \boldsymbol{x}_{\mathrm{nd}(i) \setminus \mathrm{pa}(i)} | \boldsymbol{x}_{\mathrm{pa}(i)}$$

where the **non-descendants** of a node $\mathrm{nd}(i)$ are all the nodes except for its descendants.

Propsition

For DAGs, the **ordered**, **local**, and **global** Markov properties are equivalent (under standard Bayesian network semantics).

Markov blanket

Definition

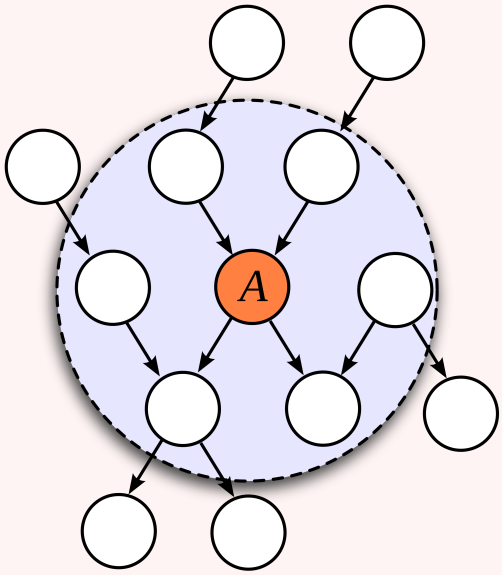
Markov blanket of a node i is the smallest set of nodes that renders a node i conditionally independent of all other nodes in the graph.

Propsition

The Markov blanket of i is given by

$$\mathrm{mb}(i) = \mathrm{ch}(i) \cup \mathrm{pa}(i) \cup \mathrm{copa}(i)$$

where $\mathrm{ch}(i)$ is the children of i and $\mathrm{copa}(i)$ is the set of co-parents of i.



Representational power and Markov equivalence

- It is always possible to build a *trivial I-map*: a fully connected DAG has $I(G) = \varnothing$.
- A *minimal I-map* can be found (conceptually) by starting from a dense graph and removing edges while preserving the I-map property.
- A *perfect map* (where $I(G) = I(p)$) does not always exist for a given p.

Two DAGs can encode the same CI structure.

Definition

Two DAGs are **Markov equivalent** (a.k.a. I-equivalent) if $I(G_1) = I(G_2)$.

When are two DPGMs I-equivalent?

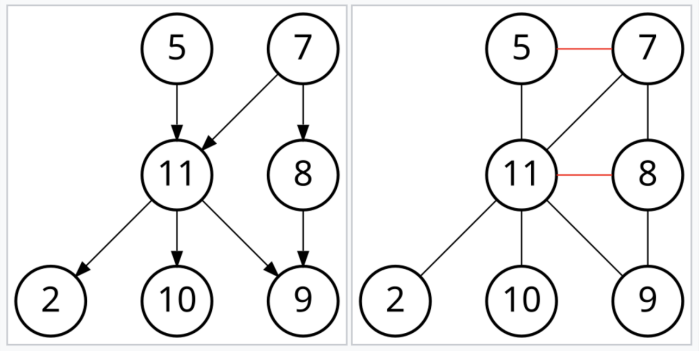
- Two graphs have the same **skeleton** if we drop the directionality of the arrows, we obtain the same undirected graph in each case.

- If G, G' have the same skeleton and the same v-structures, then $I(G) = I(G')$.

Optional: Moral graph

Definition

Let G be a directed acyclic graph; the **moral graph** G^m is formed from G by joining any non-adjacent parents and dropping the direction of edges.



Definition

We say that p satisfies the **global Markov property** with respect to G if whenever A and B are separated by C in $\left(G_{\text{an}(A \cup B \cup C)}\right)^m$ we have $\boldsymbol{x}_A \perp \boldsymbol{x}_B \mid \boldsymbol{x}_C$.