Independence Properties of Directed Probabilistic Graphical Models

November 11, 2020

Table of contents

Motivation

Motivation

Consider, for example, a Bayesian Hidden Markov Model.

$$\underset{\textit{joint}}{p(\pi, \theta, X, Y)} = \underbrace{p(\pi)\, p(\theta)}_{\text{prior}} \quad \underbrace{p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)}_{\text{(complete data) likelihood}}$$

Motivation

Consider, for example, a Bayesian Hidden Markov Model.

$$\underset{\textit{joint}}{p(\pi, \theta, X, Y)} = \underbrace{p(\pi)\, p(\theta)}_{\text{prior}} \quad \underbrace{p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)}_{\text{(complete data) likelihood}}$$

You may see statements like: *The future is independent of the past given the current hidden state.*

Motivation

Consider, for example, a Bayesian Hidden Markov Model.

$$\underbrace{p(\pi, \theta, X, Y)}_{\textit{joint}} = \underbrace{p(\pi)\, p(\theta)}_{\text{prior}} \quad \underbrace{p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)}_{\text{(complete data) likelihood}}$$

You may see statements like: *The future is independent of the past given the current hidden state.*

- How can we know this?

Motivation

Consider, for example, a Bayesian Hidden Markov Model.

$$\underset{joint}{p(\pi, \theta, X, Y)} = \underbrace{p(\pi)\, p(\theta)}_{\text{prior}} \quad \underbrace{p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)}_{\text{(complete data) likelihood}}$$

You may see statements like: *The future is independent of the past given the current hidden state.*

- How can we know this?
- More generally: How can we easily answer queries about (conditional *or* marginal) independence ?

Motivation

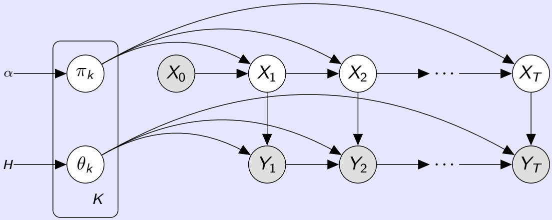
Consider, for example, a Bayesian Hidden Markov Model.

$$\underbrace{p(\pi, \theta, X, Y)}_{\textit{joint}} = \underbrace{p(\pi)\, p(\theta)}_{\text{prior}} \quad \underbrace{p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)}_{\text{(complete data) likelihood}}$$

You may see statements like: *The future is independent of the past given the current hidden state.*

- How can we know this?
- More generally: How can we easily answer queries about (conditional *or* marginal) independence ?

Representation as a ***probabilistic graphical model***



Directed Probabilistic Graphical Models

Joint distributions

The starting point for a directed probabilistic graphical model is a particular factorization of a joint density:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid \pi_i) \tag{2.1}$$

where the conditioning set π_i is referred to as the parents of variable i.

(2.1) simplifies the factorizations which are *always* true, by the chain rule of probability:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, .., X_{i-1})$$

Joint distributions

The starting point for a directed probabilistic graphical model is a particular factorization of a joint density:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid \pi_i) \tag{2.1}$$

where the conditioning set π_i is referred to as the parents of variable i.

(2.1) simplifies the factorizations which are *always* true, by the chain rule of probability:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, .., X_{i-1})$$

In other words, (2.1) restricts our consideration to a certain subset of joint probability distributions.

Joint distributions

The starting point for a directed probabilistic graphical model is a particular factorization of a joint density:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid \pi_i) \tag{2.1}$$

where the conditioning set π_i is referred to as the parents of variable i.

(2.1) simplifies the factorizations which are *always* true, by the chain rule of probability:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, .., X_{i-1})$$

In other words, (2.1) restricts our consideration to a certain subset of joint probability distributions.

Consider, e.g., the structure imposed in Bayesian models by independent priors or conditionally i.i.d likelihoods.

Directed probabilistic graphical models

Once we have specified our desired factorization via (2.1), we can identify it with a directed acyclic graph (DAG) $\mathcal{G} = (E, V)$ by:

- identifying each random variable with a node

Directed probabilistic graphical models

Once we have specified our desired factorization via (2.1), we can identify it with a directed acyclic graph (DAG) $\mathcal{G} = (E, V)$ by:

- identifying each random variable with a node
- drawing a directed arc from A to B if A is a parent of B

Directed probabilistic graphical models

Once we have specified our desired factorization via (2.1), we can identify it with a directed acyclic graph (DAG) $\mathcal{G} = (E, V)$ by:

- identifying each random variable with a node
- drawing a directed arc from A to B if A is a parent of B

We call this representation a directed probabilistic graphical model (or a Bayesian network) .

Directed probabilistic graphical models

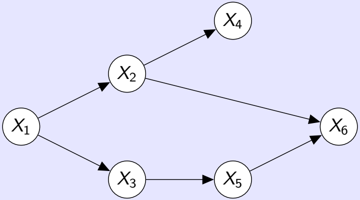
Once we have specified our desired factorization via (2.1), we can identify it with a directed acyclic graph (DAG) $\mathcal{G} = (E, V)$ by:

- identifying each random variable with a node
- drawing a directed arc from A to B if A is a parent of B

We call this representation a directed probabilistic graphical model (or a Bayesian network) .

Example

For example, the directed acyclic graph (DAG) below



corresponds to the factorization (Any guesses?)

Directed probabilistic graphical models

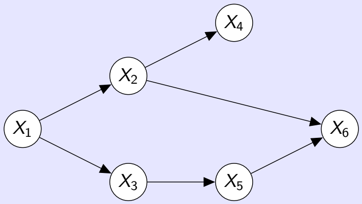
Once we have specified our desired factorization via (2.1), we can identify it with a directed acyclic graph (DAG) $\mathcal{G} = (E, V)$ by:

- identifying each random variable with a node
- drawing a directed arc from A to B if A is a parent of B

We call this representation a directed probabilistic graphical model (or a Bayesian network) .

Example

For example, the directed acyclic graph (DAG) below



corresponds to the factorization (Any guesses?)

$$p(X) = p(X_1)\, p(X_2 \mid X_1)\, p(X_3 \mid X_1)\, p(X_4 \mid X_2)\, p(X_5 \mid X_3)\, p(X_6 \mid X_5, X_2)$$

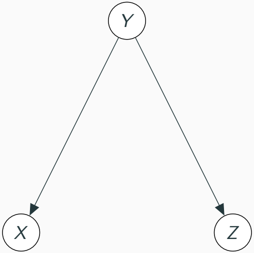
Independence in Canonical Graphs

Three canonical graphs

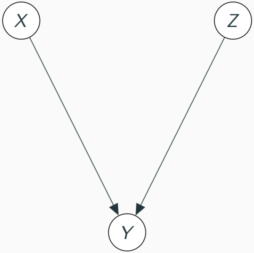
Three canonical graphs

Cascade

Common parent



v-structure

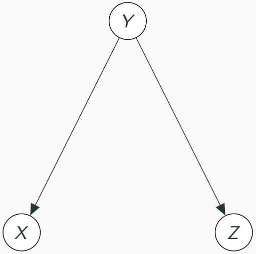


Three canonical graphs : Marginal Independence

Cascade

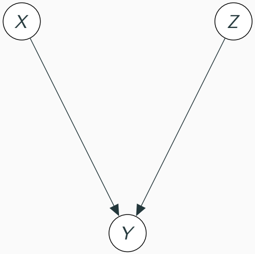
$Z \not\perp\!\!\!\perp X$

Common parent



$Z \not\perp\!\!\!\perp X$

v-structure

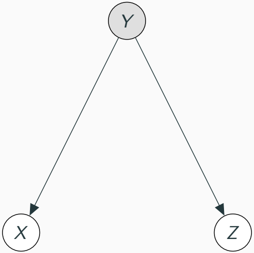


$Z \perp\!\!\!\perp X$

Cascade

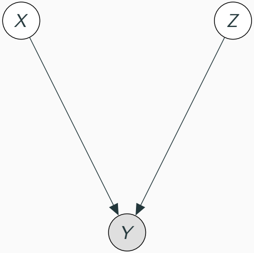
$Z \perp\!\!\!\perp X \mid Y$

Common parent



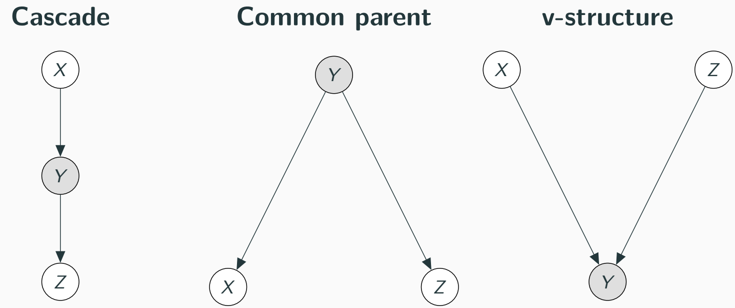
$Z \perp\!\!\!\perp X \mid Y$

v-structure



$Z \not\perp\!\!\!\perp X \mid Y$

Three canonical graphs : Take Home

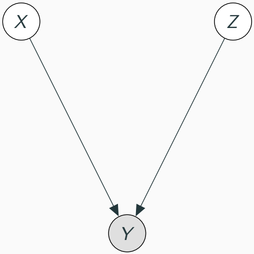


Knowing Y decouples X and Z

Knowing Y couples X and Z

Competing explanations

v-structure



The independence properties of the v-structure is commonly understood through a competing explanations paradigm.

Suppose your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.

Let

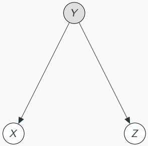
$$
\begin{aligned}
X &= \{\text{your house got robbed}\} \\
Z &= \{\text{an earthquake occurred nearby}\} \\
Y &= \{\text{your burglar alarm goes off}\}
\end{aligned}
$$

Then it is (perhaps) intuitive that

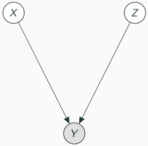
$$
\begin{aligned}
&Z \perp\!\!\!\perp X \\
&Z \not\!\perp\!\!\!\perp X \mid Y
\end{aligned}
$$

Relevance to real models

Common parent



v-structure



In real models ...

- the common parent structure shows up with conditional i.i.d data models. (So imagine Y is a parameter and X and Z are two observations.) The observations are conditionally independent, but integrating out the random parameter induces dependencies in the observations. Note in particular that the observations are, in general, *dependent* in the predictive posterior.
- the v-structure shows up with independent priors. (So imagine X and Z are model parameters given independent priors and Y is an observation.) Then the parameters are independent when generating data (i.e. in the prior), but they become dependent when doing inference (i.e. in the posterior).

Independence in Directed PGM's

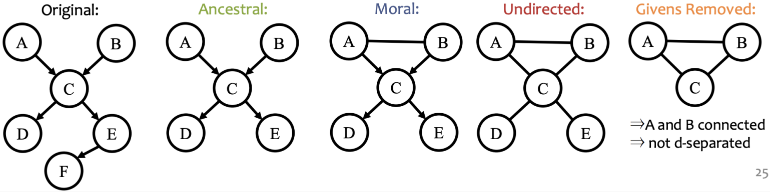
d-separation

If variables X and Z are **d-separated** given a **set** of variables E
Then X and Z are **conditionally independent** given the **set** E

Definition #2:
Variables X and Z are d-separated given a **set** of evidence variables E iff there does **not** exist a path in the **undirected ancestral moral** graph **with E removed**.

1. **Ancestral graph**: keep only X, Z, E and their ancestors
2. **Moral graph**: add undirected edge between all pairs of each node's parents
3. **Undirected graph**: convert all directed edges to undirected
4. Givens Removed: delete any nodes in E

Example Query: $A \perp\!\!\!\perp B \mid \{D, E\}$



25

Image Credit: Matt Gormley (CMU).

Worksheet for practice

Fundamental property of Bayes networks

An oft-stated fact is:

A node is independent of its non-descendants given its parents.

This can easily be proven via d-separation.

- The first step ("ancestral graph") will remove all of X's children.
- The fourth step ("remove givens") will remove X's parents.
- Thus, X will be disconnected from the rest of the graph.

Markov blankets

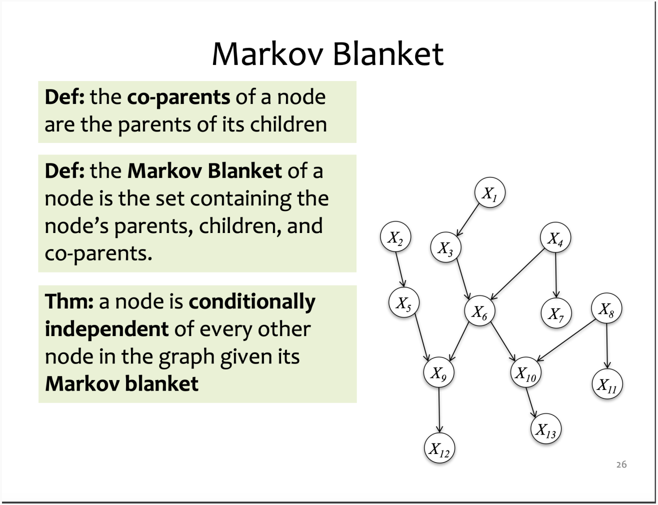


Image Credit: Matt Gormley (CMU).

Markov blankets

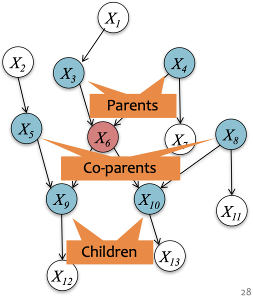
Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Thm: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



28

Image Credit: Matt Gormley (CMU).

Markov Blankets: Why *co*-parents?

Why is it not sufficient for the Markov Blanket to only include the parents and children of X_i?

Markov Blankets: Why *co*-parents?

Why is it not sufficient for the Markov Blanket to only include the parents and children of X_i?

The phenomenon of explaining away means that the observations of child nodes will not block paths to the co-parents.

This is why step 2 of the d-separation algorithm ("moralization") connects parents.

In the previous graph, the transformed graph would still have paths from X_6 to, for example, X_8 (and to X_{11}).

Proof of Markov Blanket statement

Let us consider the conditional distribution of some variable X_i given the factorization in (2.1):

$$p(X_i \mid X_{-i}) = \frac{p(X_1, \ldots, X_n)}{\int p(X_1, \ldots, X_n) d\, X_i}$$

$$= \frac{\prod_{k=1}^{n} p(X_k \mid \pi_k)}{\int \prod_{k=1}^{n} p(X_k \mid \pi_k)\ d\, X_i}$$

All terms will cancel in the numerator and denominator except for terms of the form

1. $p(X_i \mid \pi_i)$, i.e. terms where i is the node itself
2. $\{p(X_k \mid \pi_k) : i \in \pi_k\}$, i.e. terms where i is one of the parents.

Terms of type (1) will depend on X_i's parents, and terms of type (2) will depend on X_i's children and co-parents.