

Introduction

Probabilistic Graphical Models

Jerónimo Hernández-González

Revisiting concepts

- ▶ What is a Probabilistic Graphical Model?
- ▶ Which types of PGMs do you know?
- ▶ Which are the differences between them?
- ▶ Why do we use PGMs?

The Artificial Intelligence problem simplified

Given a problem,

Flu diagnosis

we use a model to represent it

train a classifier

so that we can ask questions
to the model

predict flu diagnosis for a new
patient

The logical approach to AI

Given a problem,

1. Representation:

- a) Determine which are the facts and rules that are relevant to the problem.
- b) Represent them by means of logical theory.

2. Inference:

- a) Rewrite your question as a logic formula such that the answer (true/false) to that formula is your answer.
- b) Determine whether the formula is satisfied or not using logical inference (modus ponens, modus tollens, ...).

The logical approach to AI

Given a problem,

1. Representation:

- a) Determine which are the facts and rules that are relevant to the problem.
- b) Represent them by means of logical theory.

2. Inference:

- a) Rewrite your question as a logic formula such that the answer (true/false) to that formula is your answer.
- b) Determine whether the formula is satisfied or not using logical inference (modus ponens, modus tollens, ...).

Most of the time, things are not just true or false...

The probabilistic approach to AI

Given a problem which involves **uncertainty**,

1. Representation.
 - a) Identify the relevant variables (known and unknown) related to your question
 - b) Define a probability distribution over all the variables identified.
2. Inference.
 - a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
 - b) Obtain the probability distribution by using probabilistic inference.

Example I: Medical diagnosis

In a health center, a patient comes in coughing.

A physician wants to determine how likely it is that she has flu.

1. Representation.

a) We identify the relevant variables:

1.1 whether the patient coughs $C = \{T, F\}$

1.2 whether the patient has fever $F = \{T, F\}$

1.3 whether the patient has a flu $I = \{T, F\}$

b) We define the distribution $P(C, F, I)$ with the help of the physician, who tells the probability of every combination of (C, F, I)

2. Inference.

a) Our question, rewritten as a probabilistic query, is:

Which is the probability distribution, $P(I|C = T)$?

b) We obtain the probability distribution over I using inference.

Example II: Medical diagnosis

In a health center, a second patient comes in. She is coughing and suffers hemoptysis.

A physician wants to know whether she has flu, lung cancer or none.

1. Representation. Need to improve the model

a) We identify the relevant variables:

1.1 C and F as in the previous example

1.2 whether the patient has hemoptysis $M = \{T, F\}$

1.3 patient's diagnosis $D = \{\text{Cancer}, \text{Flu}, \text{None}\}$

b) We define the dist. $P(C, F, M, D)$ with the help of the physician, who tells the probability of every combination of (C, F, M, D)

2. Inference. A different type of query

a) Our question, rewritten as a probabilistic query, is:

Which is the diagnosis d that $\arg \max_d P(D = d | C = T, M = T)$?

b) We use inference to obtain d .

Types of probabilistic queries

Given a probability distribution $P(\mathbf{X}) = P(X_1, \dots, X_n)$,
a partition of the set of variables $\mathbf{X} = (\mathbf{Y}, \mathbf{H}, \mathbf{E})$,
and an value-assignment \mathbf{e} to the subset of variables \mathbf{E} ,
we identify **two different types** of probabilistic queries:

1. **Conditional probability queries.** Find the *marginal* distribution:

$$P(\mathbf{Y} | \mathbf{E} = \mathbf{e})$$

2. **MAP queries.** Find the value-assignment \mathbf{y} to the subset \mathbf{Y} that maximizes $P(\mathbf{Y} = \mathbf{y} | \mathbf{E} = \mathbf{e})$:

$$\arg \max_{\mathbf{y}} P(\mathbf{Y} = \mathbf{y} | \mathbf{E} = \mathbf{e})$$

The probabilistic approach to AI without experts!

Given a problem which involves uncertainty and data,

1. Representation.

- a) Identify the relevant variables (known and unknown) related to your question
- b) Define the set of possible probability distributions over all the variables identified.

2. Learning.

Based on the available data, select the single probability distribution in the set that is more likely.

3. Inference.

- a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
- b) Obtain the probability distribution by using probabilistic inference.

The probabilistic approach to AI without experts!

Given a problem which involves **uncertainty**, **data** and **initial beliefs**,

1. Representation.

- a) Identify the relevant variables (known and unknown) related to your question
- b) Define the set of possible prob. distributions over all the variables identified and weigh them w.r.t. your initial belief.

2. Learning.

With the available data, refine your initial beliefs weighing more the probability distributions in the set that are more likely

3. Inference.

- a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
- b) Obtain the probability distribution by using probabilistic inference (a weighted combination of all prob. distributions).

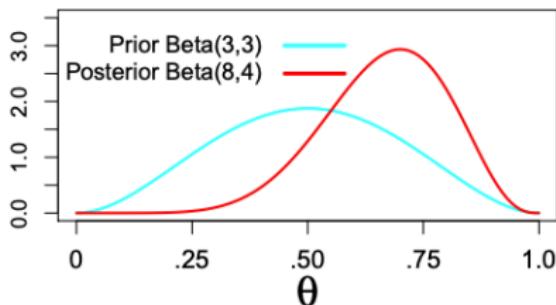
The probabilistic approach to AI without experts!

Examples first case: A novel physician learns from the records of a retired doctor.

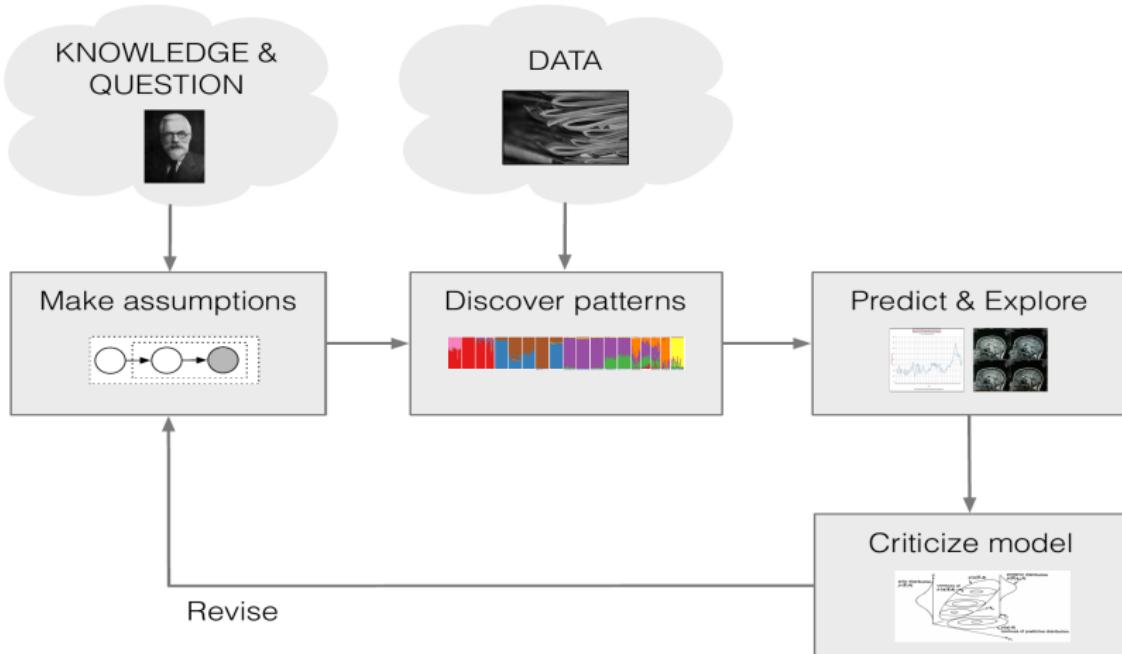
Frequentist coin example (HTTTTT): $\theta = 5/6$

Examples second case: A novel physician updates her academic beliefs with the records of a retired doctor.

Bayesian coin example (HTTTTT): $p(\theta)$



The probabilistic approach to data science



[Box, 1980; Rubin, 1984; Gelman+ 1996; Blei, 2014]

The probabilistic approach to AI: drawbacks

Example: a genetic engineer

She wants to deal with nucleotide sequences.

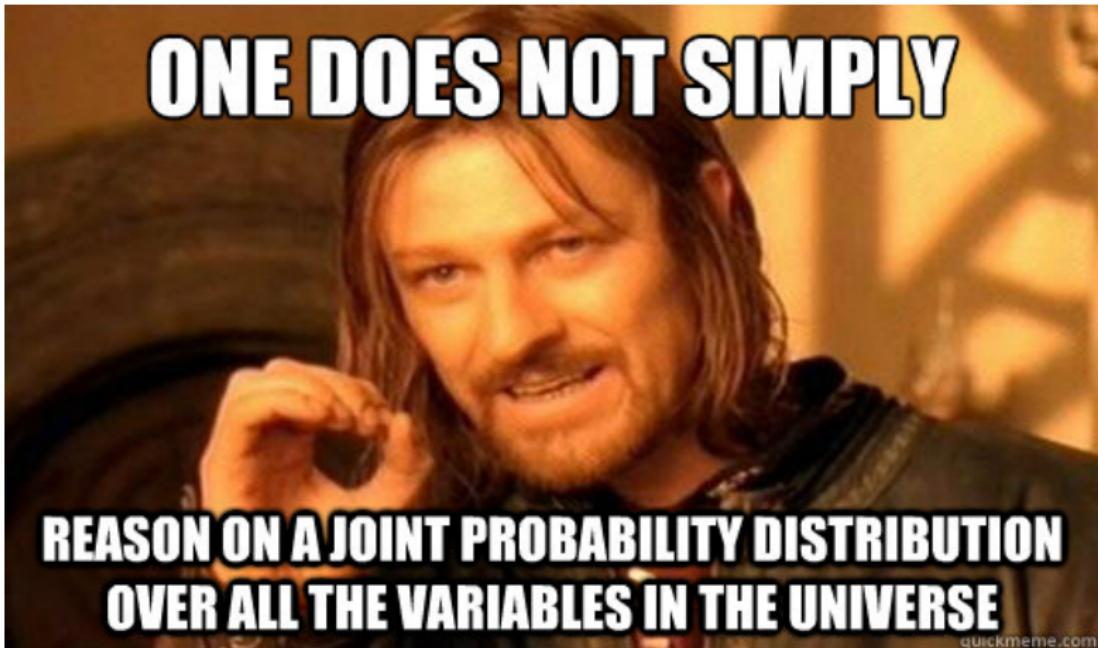
We identify n relevant variables (one per nucleotide) with 4 possible values each $\mathbf{X} = (X_1, \dots, X_n)$:

- ▶ How much memory does the joint distribution $P(X)$ need?

$$4^n$$

- ▶ What is the cost of finding the most probable assignment of values to variables?

The probabilistic approach to AI: drawbacks



Introduction

Probabilistic Graphical Models

Jerónimo Hernández-González

Probabilistic Graphical Models

We need a way to:

- ▶ Encode probability distributions over large number of variables in a compact way
- ▶ Efficiently answer queries
- ▶ Learn from the available data

Probabilistic Graphical Models

We need a way to:

- ▶ Encode probability distributions over large number of variables in a compact way
- ▶ Efficiently answer queries
- ▶ Learn from the available data

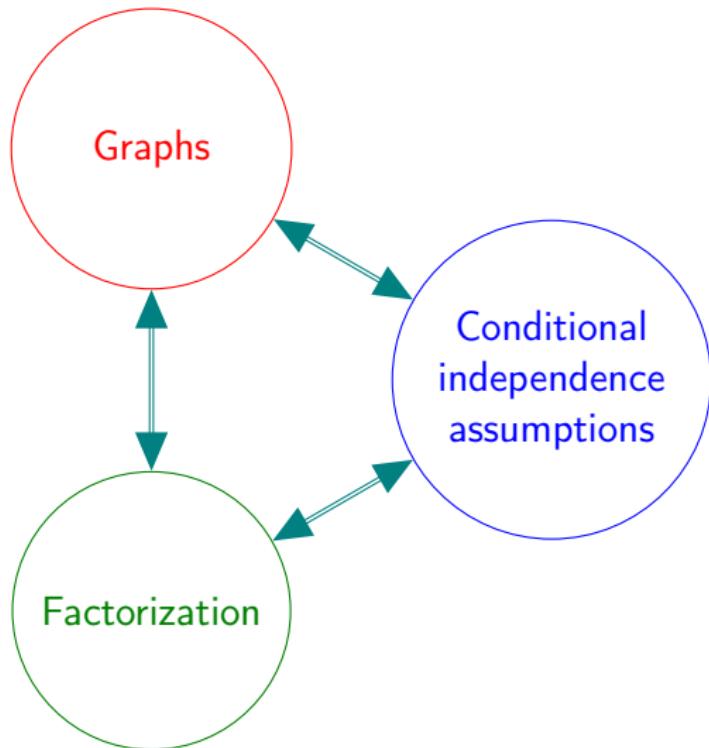
PGMs do the job!

- ▶ Use a graph:
Each variable is represented by a vertex in the graph. *Possible dependencies* between variables are encoded by adding edges to the graph
- ▶ Why graphs?
 - ▶ Easy to visualize and understand.
 - ▶ Mathematically adequate.

Probabilistic graphical models

- ▶ Given by:
 - ▶ Structure: a graph
 - ▶ Parameters: a set of conditional probabilities
- ▶ Three main types:
 - ▶ Directed acyclic graphs: Bayesian networks (a.k.a. belief networks or causal networks).
 - ▶ Undirected graphs: Markov networks (a.k.a. Markov random fields)
 - ▶ Bipartite graphs: Factor graphs.
- ▶ They represent:
 - ▶ Factorizations of probability distributions: Chain rule with a reduced set of conditioning variables
 - ▶ Dependency models: A set of conditional independences

Three views of probabilistic modeling



Probabilistic independence

2 random variables are **independent** if, for all values that Z and Y can take,

$$P(Y, Z) = P(Y) \cdot P(Z)$$

An equivalent condition is:

$$P(Y|Z) = P(Y) \quad \text{or} \quad P(Z|Y) = P(Z)$$

We note it as $Y \perp\!\!\!\perp Z$

** Same definition when \mathbf{Y} and \mathbf{Z} are disjoint sets of variables.

Probabilistic independence

Examples

The probability that there is electricity in this room today is...

- ▶ **independent** of whether Borussia Dortmund wins European Champions League this year.
- ▶ **not independent** of whether there was electricity yesterday.
- ▶ **not independent** of whether there is electricity in the room next door.
- ▶ **independent** of whether I took a shower this morning.

Probabilistic **conditional** independence

Given 3 random variables, X, Y, Z we say that X is **conditionally independent** of Y given Z if,

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

We note it as $X \perp\!\!\!\perp Y|Z$.

** Same definition when \mathbf{X} , \mathbf{Y} and \mathbf{Z} are disjoint sets of variables.

When modeling a problem, some conditional independence relations are usually clear.

Probabilistic conditional independence

Examples

- ▶ Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \dots, 10\}$.
- ▶ Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
Let n_a be the number Alice heard, and n_b the one Bob heard.
- ▶ Are n_a and n_b (marginally) independent?

Probabilistic conditional independence

Examples

- ▶ Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \dots, 10\}$.
- ▶ Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
Let n_a be the number Alice heard, and n_b the one Bob heard.
- ▶ Are n_a and n_b (marginally) independent?
No! We'd expect $P(n_a = 1 | n_b = 1) \neq P(n_a = 1)$
** n_b gives some evidence of what n_k might be

Probabilistic conditional independence

Examples

- ▶ Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \dots, 10\}$.
- ▶ Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.

Let n_a be the number Alice heard, and n_b the one Bob heard.

- ▶ Are n_a and n_b (marginally) independent?
No! We'd expect $P(n_a = 1 | n_b = 1) \neq P(n_a = 1)$
** n_b gives some evidence of what n_k might be
- ▶ Are n_a and n_b conditionally independent given n_k ?
Yes! If we know what Kevin actually said, n_a and n_b are no longer related.

$$P(n_a = 1 | n_b = 1, n_k = 2) = P(n_a = 1 | n_k = 2)$$

** n_k fully explains any possible relationships between n_a and n_b

Probabilistic conditional independence

Examples

- ▶ Two random variables that are marginally independent **can** also be conditionally independent given a third variable:

The probability that Borussia Dortmund wins next year Champions League is independent on whether there is electricity in this room given that there is electricity in the room next door.

- ▶ However that's not always the case:
Given two coins C_1 and C_2 , these are marginally independent

$$P(C_1|C_2 = H) = P(C_1)$$

Consider now a third variable: S is *true* if both coins show the same face

$$P(C_1|C_2 = H, S = \text{true}) \neq P(C_1|S = \text{true})$$

Independence: modeling advantages

Independence is good for simplicity!

If we have to represent a probability distribution $P(\mathbf{X})$, and we know that we can split $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{Z},$$

then we can represent $P(\mathbf{X}) = P(\mathbf{Y}) \cdot P(\mathbf{Z})$, that is, we need to represent just two smaller pieces, $P(\mathbf{Y})$ and $P(\mathbf{Z})$.

Example: Say \mathbf{X} has 10 binary variables and \mathbf{Y} and \mathbf{Z} have 5 binary variables each. How much memory are we saving?

Independence: modeling advantages

Independence is good for simplicity!

If we have to represent a probability distribution $P(\mathbf{X})$, and we know that we can split $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{Z},$$

then we can represent $P(\mathbf{X}) = P(\mathbf{Y}) \cdot P(\mathbf{Z})$, that is, we need to represent just two smaller pieces, $P(\mathbf{Y})$ and $P(\mathbf{Z})$.

Example: Say \mathbf{X} has 10 binary variables and \mathbf{Y} and \mathbf{Z} have 5 binary variables each. How much memory are we saving?

From 1024 to 64 parameters!

Structural statistical model

A set of (conditional) independence statements

Simple explanation with two variables, X and Y

There is a single independence statement that $P(X, Y)$ can satisfy:

$$X \perp\!\!\!\perp Y$$

There are two possible structural models:

1. $\mathcal{M}_1 = \emptyset$
2. $\mathcal{M}_2 = \{X \perp\!\!\!\perp Y\}$.

A distribution respects model \mathcal{M} if it satisfies all the independencies in \mathcal{M}

All the possible prob. distributions $p(X, Y)$ can be classified into:

- a) “Distributions that respect \mathcal{M}_1 ”
- b) “Distributions that respect \mathcal{M}_2 ”.

Structural statistical model

A set of (conditional) independence statements

Generalized explanation, with n variables

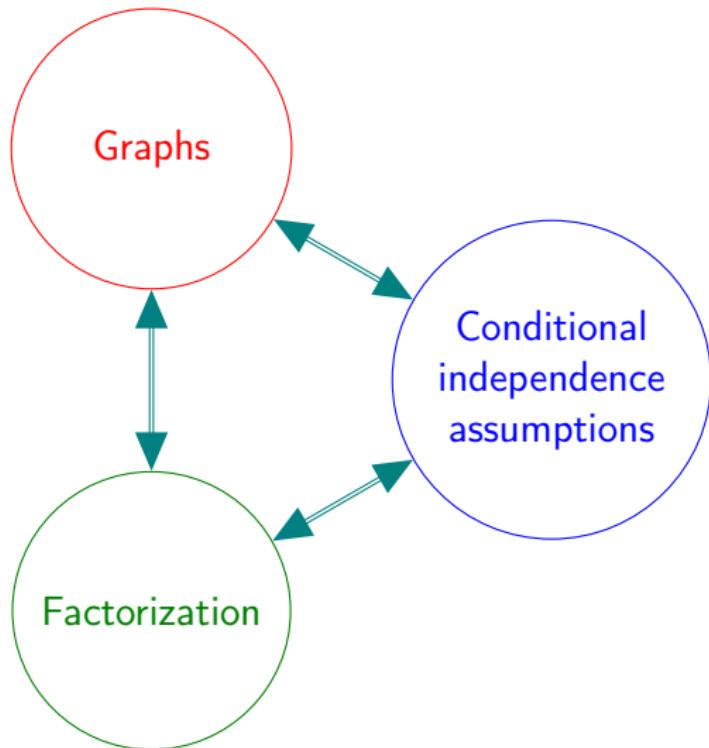
A structural statistical model over a set of variables \mathbf{V} is described by a set of conditional independence assumptions like:

$$\begin{aligned}\mathcal{M} = \{ & \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}, \\ & \mathbf{X} \perp\!\!\!\perp \mathbf{Z} | \mathbf{Y}, \\ & \mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \}\end{aligned}$$

where $\mathbf{V} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$.

A distribution respects model \mathcal{M} if it satisfies all the independencies in \mathcal{M}

Three views of probabilistic modeling



Factorization

Example

Let us consider the following complex function:

$$f(x, y, z, t) = 6yzt + y^4z - 6\sqrt{x}zt - \sqrt{x}y^3z + 6yt \log t + \\ y^4t \log t - 6\sqrt{xt} \log t - \sqrt{xy^3} \log t$$

Factorization

Example

Let us consider the following complex function:

$$f(x, y, z, t) = 6yzt + y^4z - 6\sqrt{x}zt - \sqrt{x}y^3z + 6yt \log t + y^4t \log t - 6\sqrt{xt} \log t - \sqrt{xy}^3 \log t$$

It can be re-expressed as:

$$f(x, y, z, t) = (z + \log t) \cdot (y - \sqrt{x}) \cdot (6t + y^3)$$

Factorization

Example

Let us consider the following complex function:

$$f(x, y, z, t) = 6yzt + y^4z - 6\sqrt{x}zt - \sqrt{x}y^3z + 6yt \log t + y^4t \log t - 6\sqrt{xt} \log t - \sqrt{xy}^3 \log t$$

It can be re-expressed as:

$$f(x, y, z, t) = (z + \log t) \cdot (y - \sqrt{x}) \cdot (6t + y^3)$$

We say that f factorizes as a product of three factors:

$$f(x, y, z, t) = f_1(z, t) \cdot f_2(x, y) \cdot f_3(y, t)$$

Factorization

Example

Note that the following factorization:

$$f(x, y, z, t) = f_1(z, t) \cdot f_2(x, y) \cdot f_3(y, t)$$

involves:

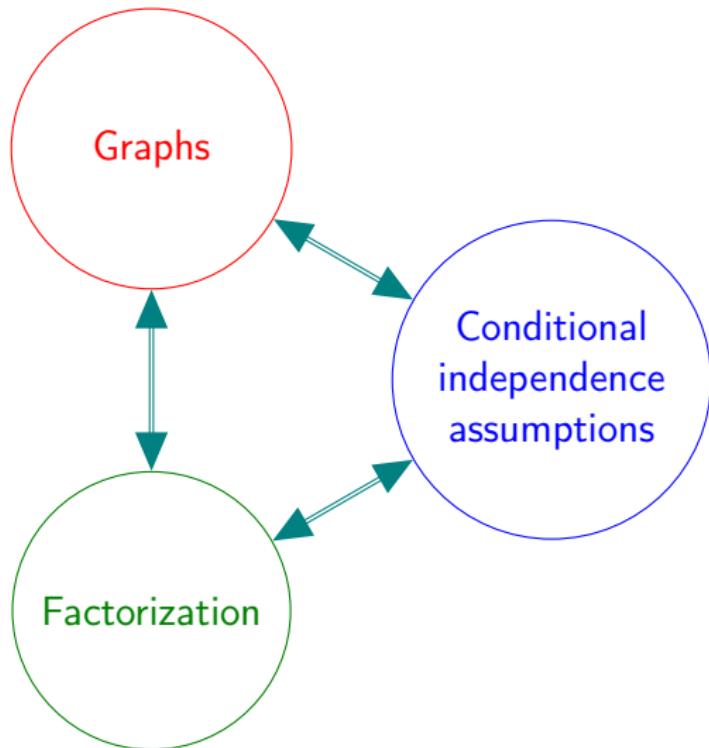
- ▶ $\text{Scope}(f) = \{x, y, z, t\}$
- ▶ $\text{Scope}(f_1) = \{z, t\}$
- ▶ $\text{Scope}(f_2) = \{x, y\}$
- ▶ $\text{Scope}(f_3) = \{z, t\}$

When a function (our prob. distribution) factorizes into small parts, we can represent it in a smaller space.

We are interested in probability distributions that factorize!!

$$\text{Ex.: } P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$$

Three views of probabilistic modeling



Probabilistic graphical models

- ▶ Given by:
 - ▶ Structure: a graph
 - ▶ Parameters: a set of conditional probabilities
- ▶ Three main types:
 - ▶ Directed acyclic graphs: Bayesian networks (a.k.a. belief networks or causal networks).
 - ▶ Undirected graphs: Markov networks (a.k.a. Markov random fields)
 - ▶ Bipartite graphs: Factor graphs.
- ▶ They represent:
 - ▶ Factorizations of probability distributions: Chain rule with a reduced set of conditioning variables
 - ▶ Dependency models: A set of conditional independences

Distributions and factors

Approaches:

- ▶ Distributions
 - ▶ Joint distribution
 - ▶ Conditioning
 - ▶ Marginalization
- ▶ Factors
 - ▶ How are they different from distributions?
 - ▶ Product
 - ▶ Marginalization
 - ▶ Reduction

Inference

PGMs allow for probability based operations

- ▶ which can be done more efficiently (exact/approximate)
 - ▶ Exact: marginalize, conditioning, belief propagation,...
 - ▶ Approx.: random sampling
- ▶ in order to answer two types of queries

Learning

- ▶ From **data** (and expert **knowledge**)
- ▶ Learning algorithms for...
 - ▶ **Parametric learning**
 - ▶ **Structural learning** (**NP-complete**)
 - ▶ **Quantitative**: Scoring functions (e.g. likelihood)
 - ▶ **Qualitative**: Conditional independence tests
- ▶ To obtain **robust** models, we need to control the trade-off between **model complexity** and amount of **available data**

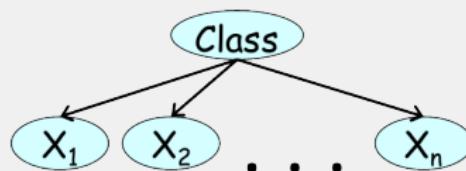
Relation with supervised learning

A generative approach

1. Learn the underlying statistical model
2. Classify unseen instances into the most probable class
 - ** Bayes classifier: lower bound of the classification error

Structures biased towards supervised learning

Few parameters with high discriminative information
Ex.: (Augmented) naive Bayes family



What is all this about?

By the end of this course you should know:

- ▶ Representation: what is a probabilistic graphical model
- ▶ Inference: which queries can we ask to them
- ▶ Learning: how to obtain PGMs from data

What is all this about?

By the end of this course you should know:

- ▶ Representation: what is a probabilistic graphical model
 - ▶ Directed and Undirected
 - ▶ Temporal and plate models
- ▶ Inference: which queries can we ask to them
 - ▶ when (and how) these questions can be answered exactly in polynomial time (exact inference)
 - ▶ what to do when they cannot (approximate inference)
- ▶ Learning: how to obtain PGMs from data
 - ▶ Parameters and structure
 - ▶ With and without complete data

What is all this about?

By the end of this course you should know:

- ▶ Representation: what is a probabilistic graphical model
 - ▶ Directed and Undirected
 - ▶ Temporal and plate models
- ▶ Inference: which queries can we ask to them
 - ▶ when (and how) these questions can be answered exactly in polynomial time (exact inference)
 - ▶ what to do when they cannot (approximate inference)
- ▶ Learning: how to obtain PGMs from data
 - ▶ Parameters and structure
 - ▶ With and without complete data

Furthermore, you should be able to:

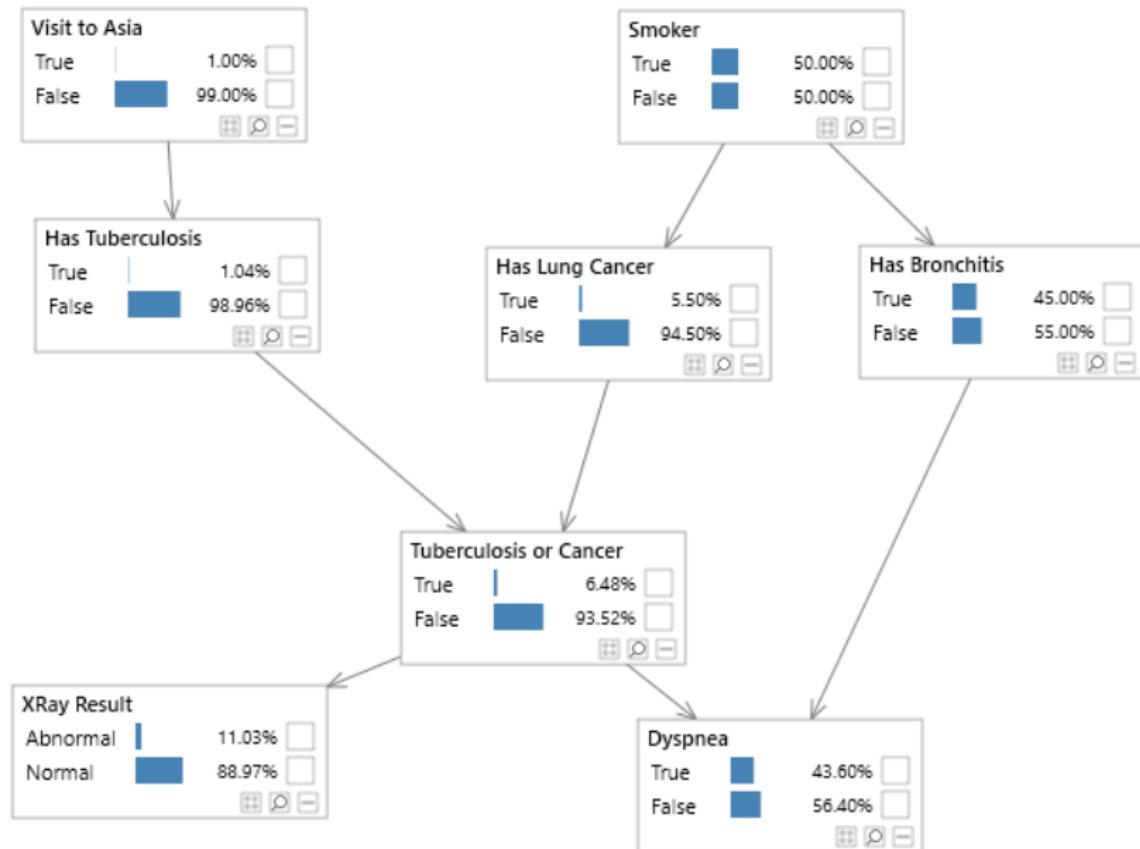
- ▶ apply PGM algorithms to problems of your interest
- ▶ translate PGMs and algorithms into code

Introduction

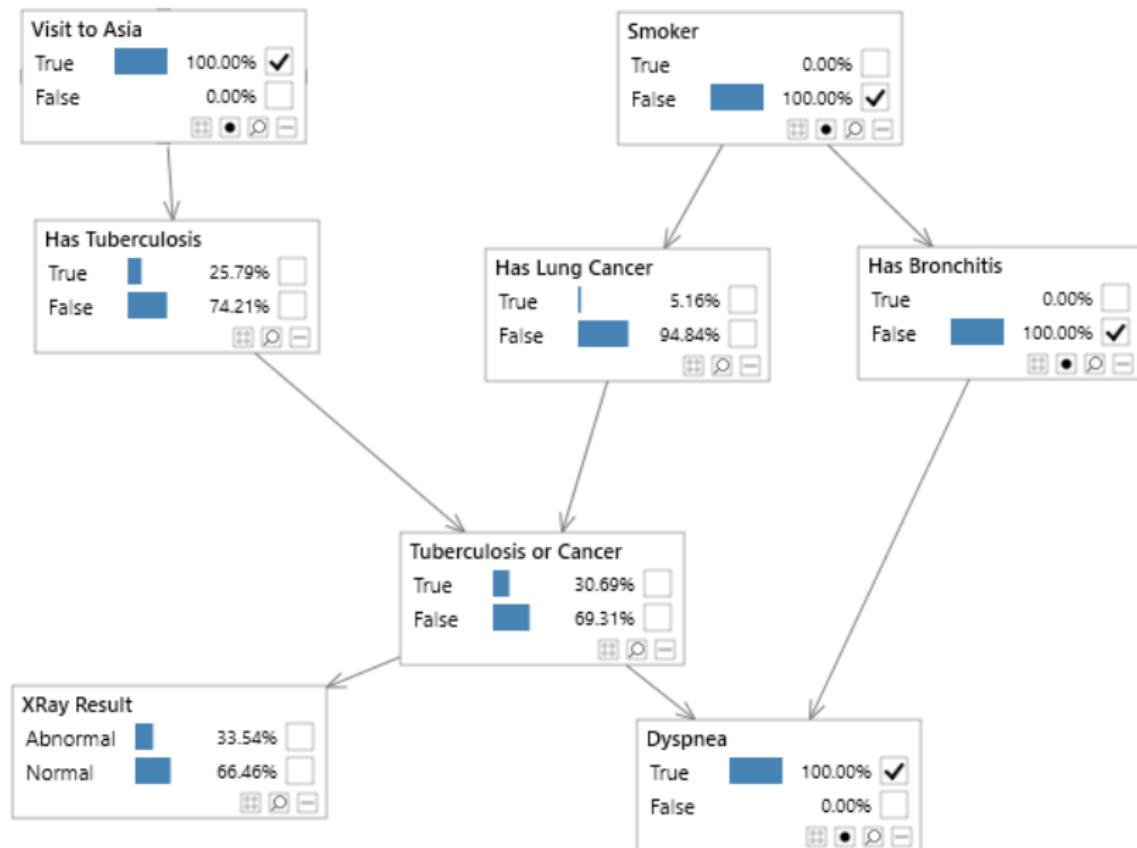
Probabilistic Graphical Models

Jerónimo Hernández-González

Applications (Asia pgm)



Applications (Asia pgm)



Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

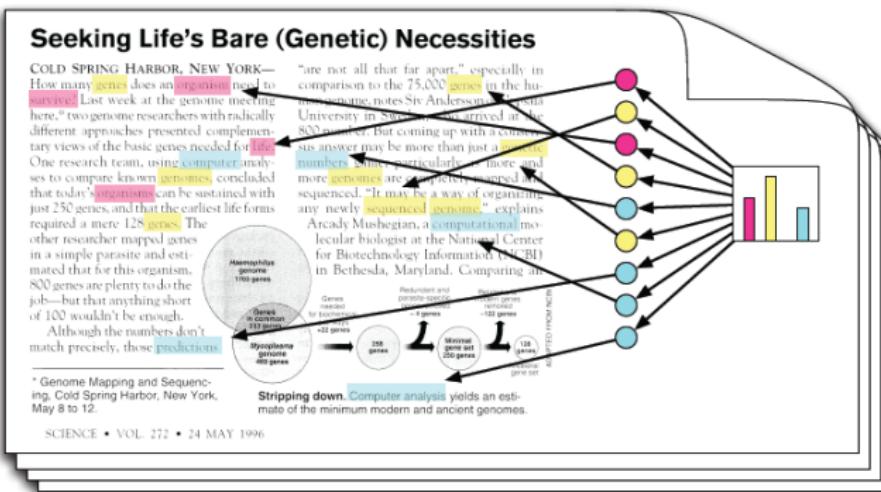
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a geneticist at the University of Sweden who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game. "The particular... more and more organisms are completely mapped and sequenced. 'It may be a way of organizing any newly sequenced genome,'" explains

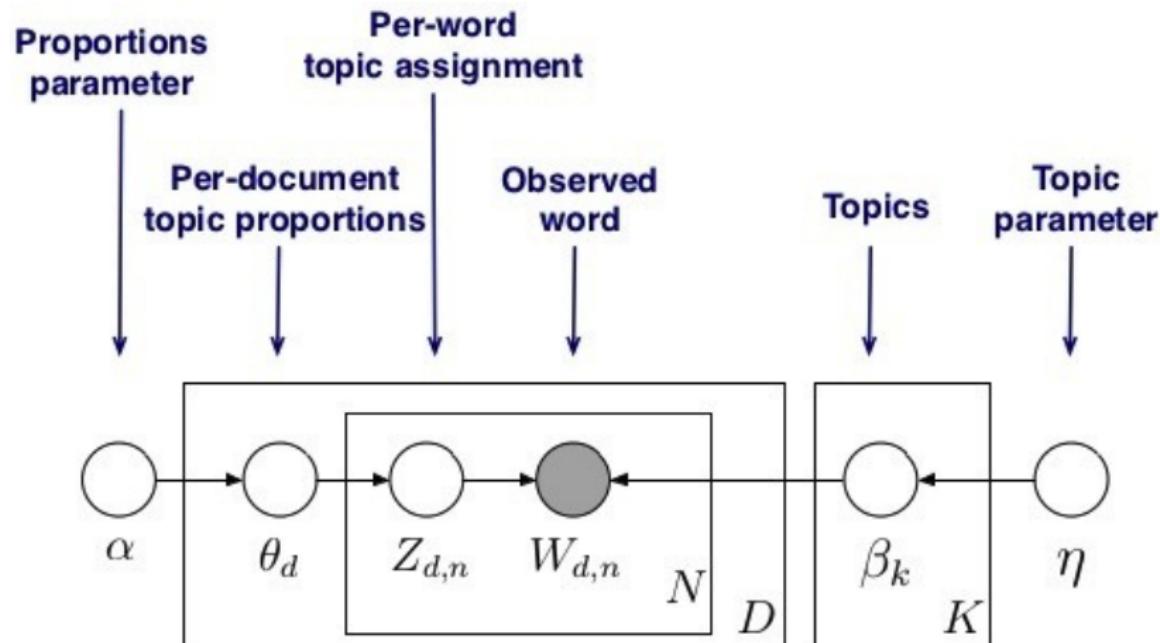
Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

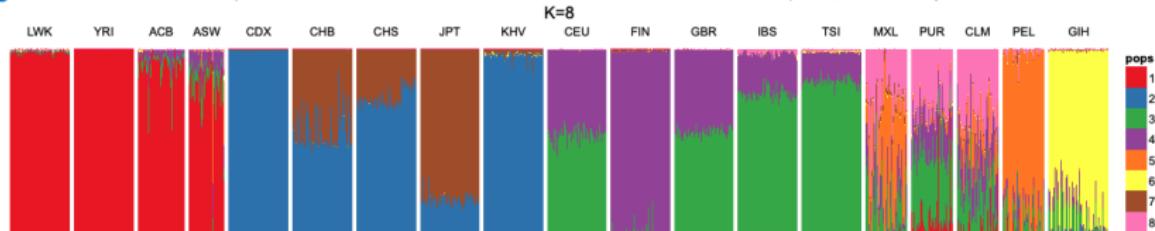
Topic proportions and assignments



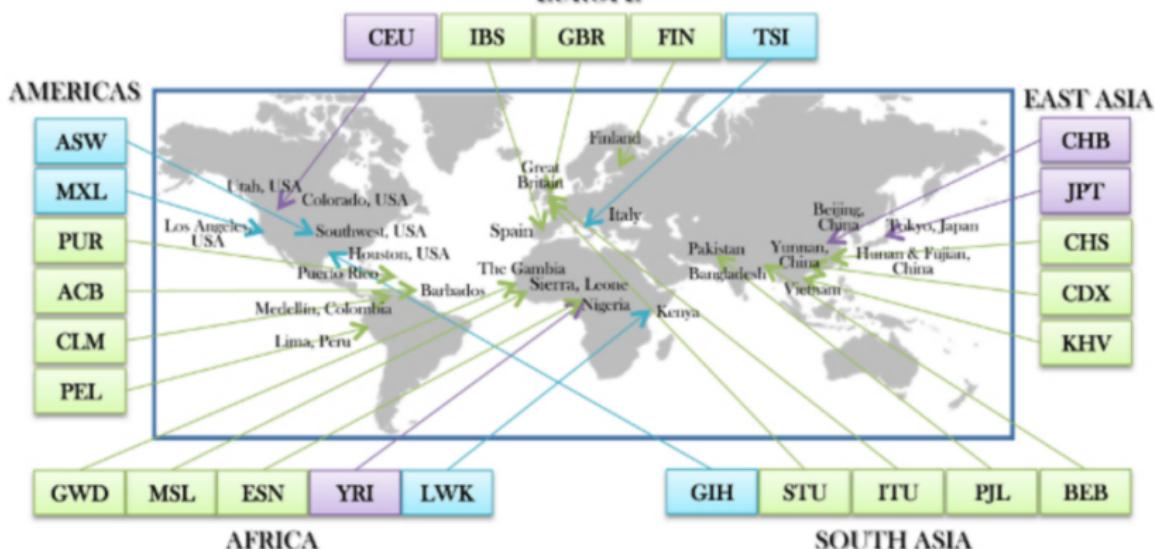


Latent Dirichlet Allocation (LDA)

Applications (Gopalan et al. 2016, doi: 10.1038/ng.3710)



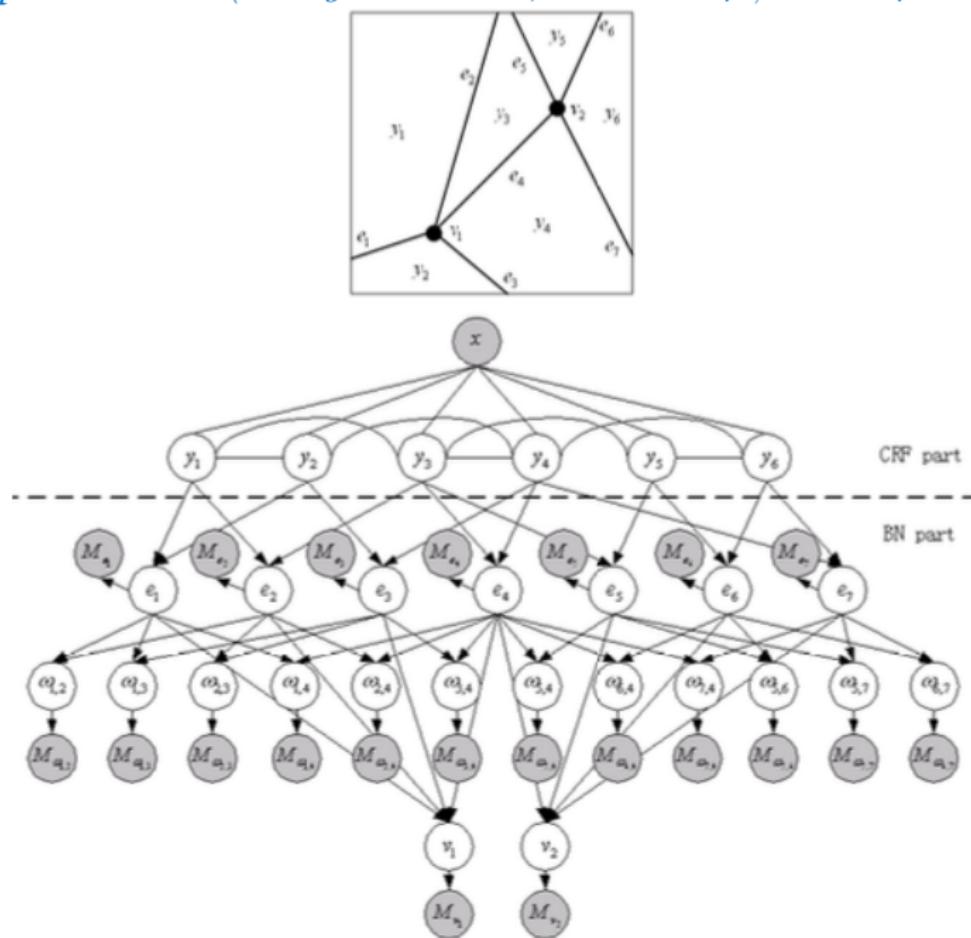
EUROPE



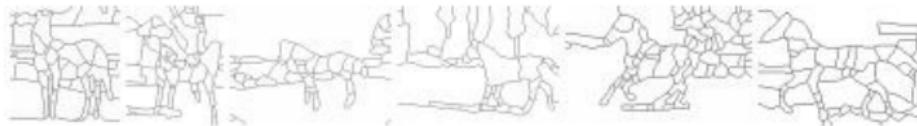
International HapMap Population

HapMap 3 Population

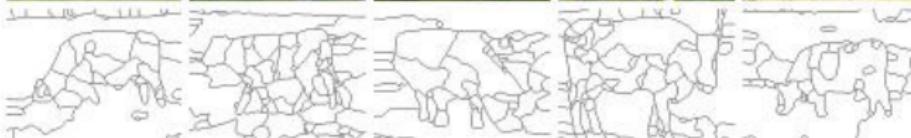
New 1000 Genomes Population



Applications (Zhang et al. 2009, doi: 10.1142/9789814273398_0006)



Applications (Zhang et al. 2009, doi: 10.1142/9789814273398_0006)



Introduction

Probabilistic Graphical Models

Jerónimo Hernández-González

Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

Probability Overview

- ▶ Events
Discrete random variables, continuous random variables, compound events
- ▶ Axioms of probability
What defines a reasonable theory of uncertainty
- ▶ Conditional probabilities
- ▶ Chain rule
- ▶ Bayes rule
- ▶ Joint probability distribution
- ▶ $P(Y|X)$: Facing practical problems
 - ▶ Conditional independencies for model simplification
 - ▶ Principles of parameter estimation

Random Variables

- ▶ Informally, A is a random variable if
 - ▶ A denotes something about which we are uncertain
 - ▶ perhaps the outcome of a randomized experiment
- ▶ Examples
 - ▶ A : True if a randomly drawn person from our class is female
 - ▶ A : The hometown of a randomly drawn person from our class
 - ▶ A : True if two randomly drawn persons from our class have same birthday
- ▶ Define $P(A)$ as “the fraction of possible worlds in which A is true” or “the fraction of times A holds, in repeated runs of the random experiment”
 - ▶ The set of possible worlds is called the sample space, S
 - ▶ A random variable A is a function defined over S

$$A : S \rightarrow \{0, 1\}$$

A bit of formalism

More formally, we have:

- ▶ a sample space S , a.k.a. the set of possible worlds
E.g., set of students in our class
- ▶ a random variable is a function defined over the sample space
Gender: $S \rightarrow \{m, f\}$
Height: $S \rightarrow \mathbb{R}$
- ▶ an event is a subset of S
E.g., the subset of S for which Gender= f
E.g., the subset of S for which (Gender= m) AND (eyeColor=blue)
- ▶ we are often interested in probabilities of specific events
- ▶ and of specific events conditioned on other specific events

A bit of formalism

- ▶ X : random variable
(UPPERCASE)
- ▶ $\mathbf{X} = (X_1, \dots, X_n)$: random vector
(BOLD UPPERCASE)
- ▶ Ω_X : possible values of random variable X .
- ▶ $x \in \Omega_X$: a possible value of random variable X : $X = x$
(lowercase)
- ▶ $\mathbf{x} = (x_1, \dots, x_n)$: a possible (tuple of) value of a random vector \mathbf{X} : $(X_1 = x_1, \dots, X_n = x_n), \forall X_i : x_i \in \Omega_{X_i}$
(bold lowercase)

A bit of formalism

- ▶ X : random variable
(UPPERCASE)
- ▶ $\mathbf{X} = (X_1, \dots, X_n)$: random vector
(BOLD UPPERCASE)
- ▶ Ω_X : possible values of random variable X .
- ▶ $x \in \Omega_X$: a possible value of random variable X : $X = x$
(lowercase)
- ▶ $\mathbf{x} = (x_1, \dots, x_n)$: a possible (tuple of) value of a random vector \mathbf{X} : $(X_1 = x_1, \dots, X_n = x_n), \forall X_i : x_i \in \Omega_{X_i}$
(bold lowercase)

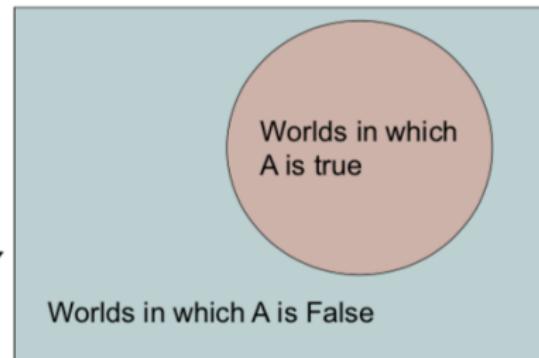
* **Instantiation**

Visualizing A

Sample space
of all possible
worlds

Its area is 1

$P(A) = \text{Area of}$
reddish oval



A bit of formalism

The Axioms of Probability:

- ▶ $0 \leq P(A) \leq 1$
- ▶ $P(\text{True}) = 1$
- ▶ $P(\text{False}) = 0$
- ▶ $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

[di Finetti 1931]:

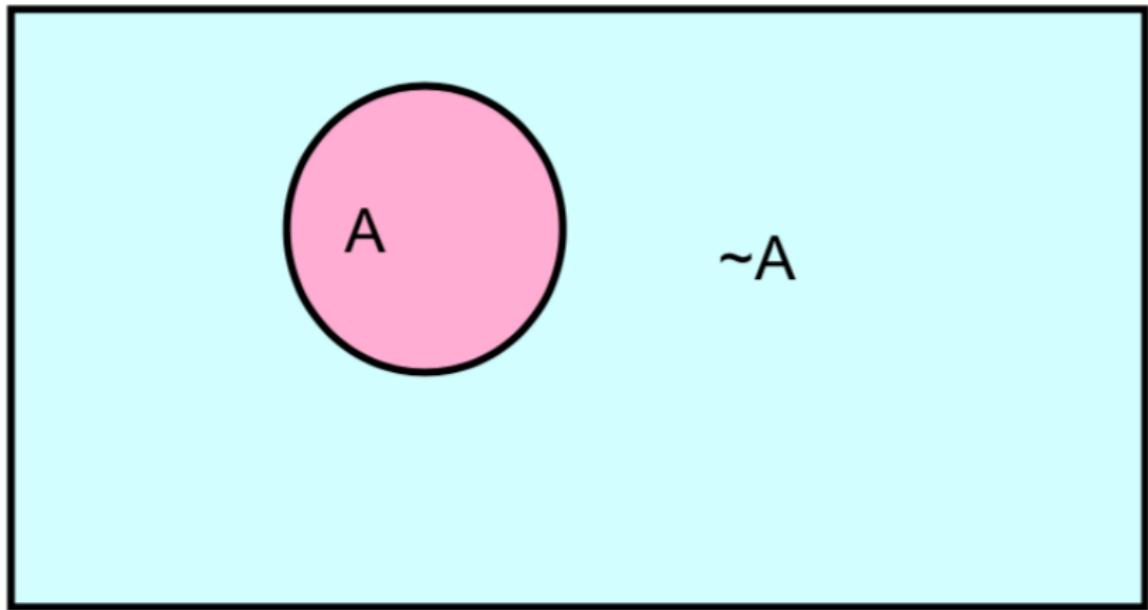
when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

Elementary probability in pictures

$$P(\neg A) + P(A) = 1$$



A useful theorem

$$0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

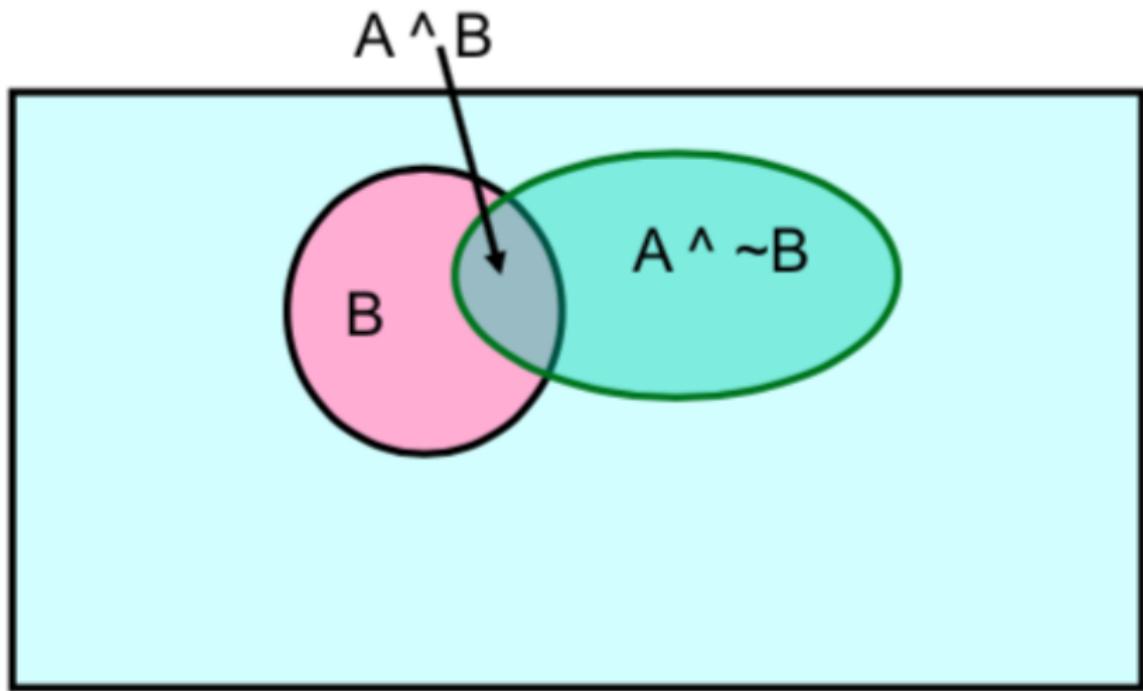
$$A = [A \text{ and } (B \text{ or } \neg B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \neg B)]$$

$$\begin{aligned} P(A) &= P(A \text{ and } B) + P(A \text{ and } \neg B) - P((A \text{ and } B) \text{ and } (A \text{ and } \neg B)) \\ &= P(A \text{ and } B) + P(A \text{ and } \neg B) \text{ - } P(A \text{ and } B \text{ and } A \text{ and } \neg B)} \end{aligned}$$

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B)$$

Elementary probability in pictures

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B) \quad [\text{and} \equiv \wedge]$$

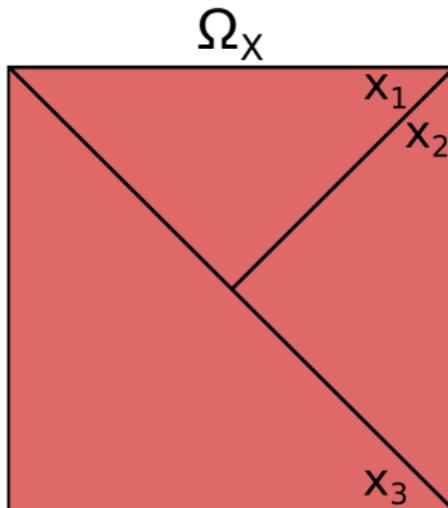


Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x) \right) = 1$$



$$p(X = x_1) = 1/4$$

$$p(X = x_2) = 1/4$$

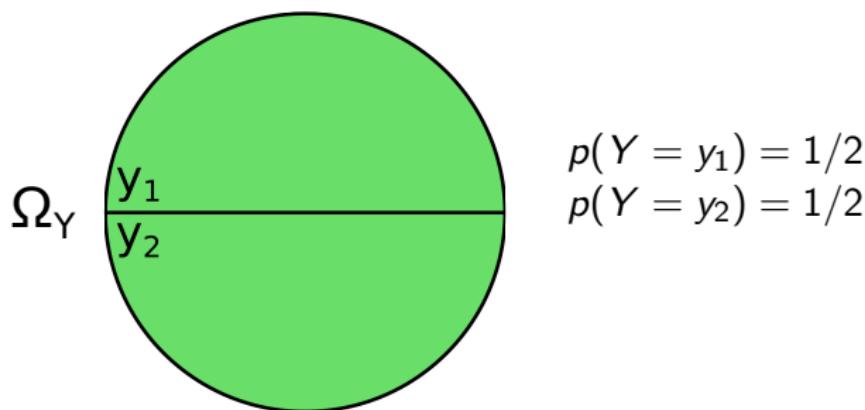
$$p(X = x_3) = 1/2$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x) \right) = 1$$



Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x) \right) = 1$$

$p(\mathbf{X})$: joint probabilities

$$0 \leq p(\mathbf{X} = \mathbf{x}) \leq 1, \forall \mathbf{x} = (x_1, \dots, x_n) \in \Omega_{X_1} \times \dots \times \Omega_{X_n}$$

$$\left(\sum_{\mathbf{x} \in \Omega_{X_1} \times \dots \times \Omega_{X_n}} p(\mathbf{X} = \mathbf{x}) \right) = 1$$

$$p(\mathbf{X}) = p(X_1, \dots, X_n) = p(X_1 \cap \dots \cap X_n)$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x) \right) = 1$$

$p(\mathbf{X})$: joint probabilities

$$0 \leq p(\mathbf{X} = \mathbf{x}) \leq 1, \forall \mathbf{x} = (x_1, \dots, x_n) \in \Omega_{X_1} \times \dots \times \Omega_{X_n}$$

$$\left(\sum_{\mathbf{x} \in \Omega_{X_1} \times \dots \times \Omega_{X_n}} p(\mathbf{X} = \mathbf{x}) \right) = 1$$

$$p(\mathbf{X}) = p(X_1, \dots, X_n) = p(X_1 \cap \dots \cap X_n)$$

Marginalization

$$p(Y) = \sum_{x \in \Omega_X} p(Y|X = x) \cdot p(X = x)$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x) \right) = 1$$

$p(\mathbf{X})$: joint probabilities

$$0 \leq p(\mathbf{X} = \mathbf{x}) \leq 1, \forall \mathbf{x} = (x_1, \dots, x_n) \in \Omega_{X_1} \times \dots \times \Omega_{X_n}$$

$$\left(\sum_{\mathbf{x} \in \Omega_{X_1} \times \dots \times \Omega_{X_n}} p(\mathbf{X} = \mathbf{x}) \right) = 1$$

$$p(\mathbf{X}) = p(X_1, \dots, X_n) = p(X_1 \cap \dots \cap X_n)$$

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y | X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y | X = x) \right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X, Y)}{p(X)} \text{ with } p(X) \neq 0$$

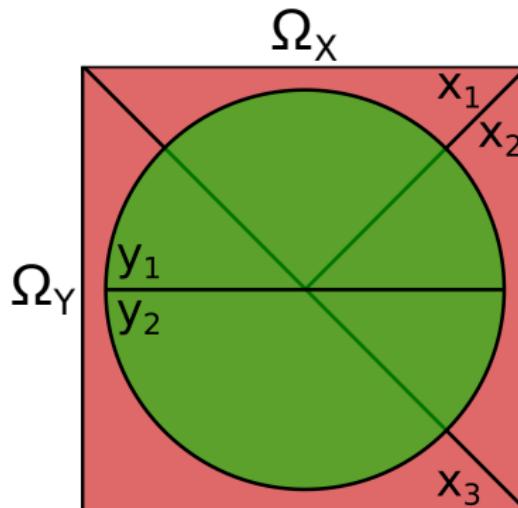
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x) \right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$$\begin{aligned} p(Y = y_1|X = x_1) &= \\ p(Y = y_2|X = x_1) &= \\ p(Y = y_1|X = x_2) &= \\ p(Y = y_2|X = x_2) &= \\ p(Y = y_1|X = x_3) &= \\ p(Y = y_2|X = x_3) &= \end{aligned}$$

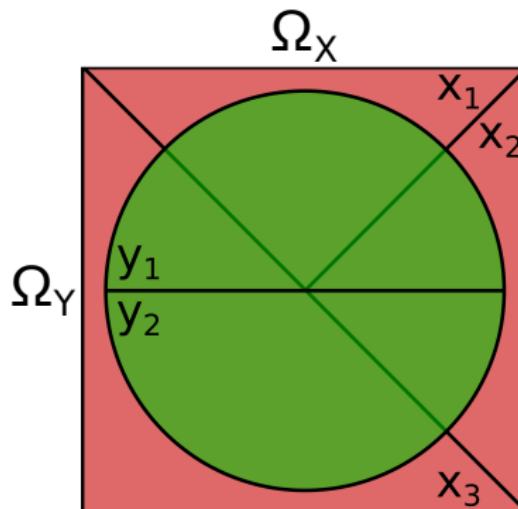
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x) \right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$$p(Y = y_1|X = x_1) = 1$$

$$p(Y = y_2|X = x_1) = 0$$

$$p(Y = y_1|X = x_2) =$$

$$p(Y = y_2|X = x_2) =$$

$$p(Y = y_1|X = x_3) =$$

$$p(Y = y_2|X = x_3) =$$

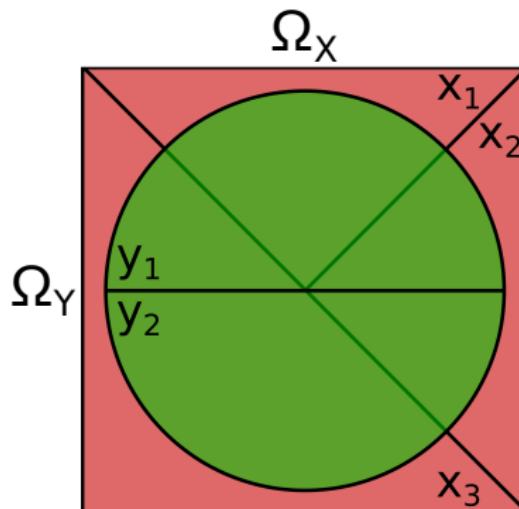
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x) \right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$$p(Y = y_1|X = x_1) = 1$$

$$p(Y = y_2|X = x_1) = 0$$

$$p(Y = y_1|X = x_2) = 1/2$$

$$p(Y = y_2|X = x_2) = 1/2$$

$$p(Y = y_1|X = x_3) =$$

$$p(Y = y_2|X = x_3) =$$

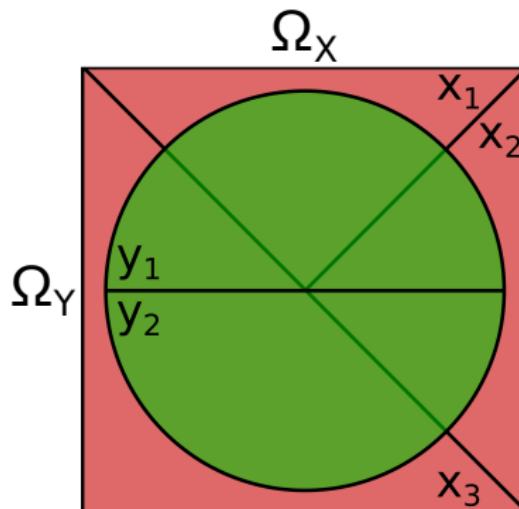
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x) \right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$$p(Y = y_1|X = x_1) = 1$$

$$p(Y = y_2|X = x_1) = 0$$

$$p(Y = y_1|X = x_2) = 1/2$$

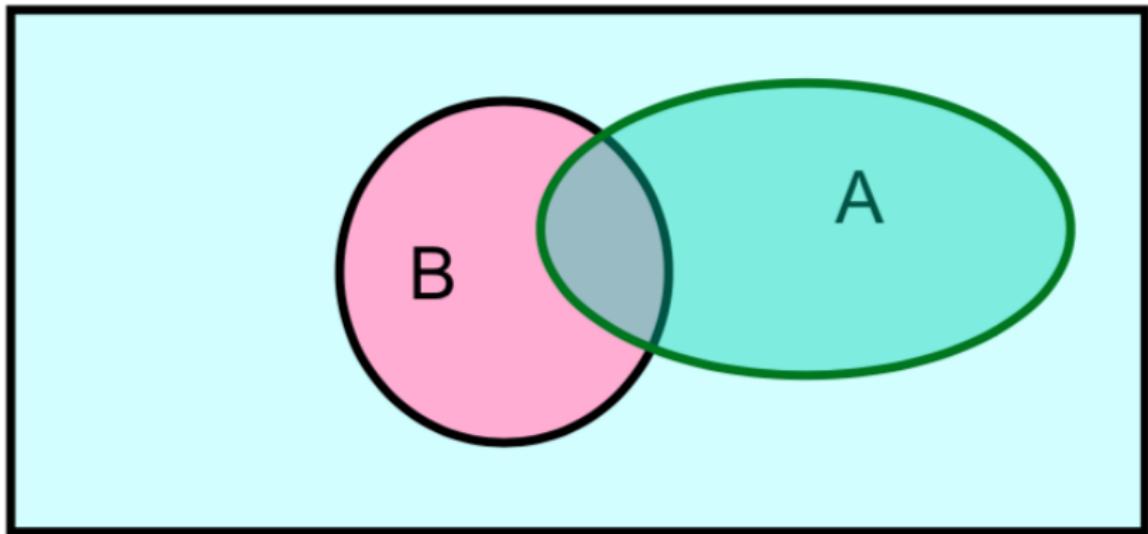
$$p(Y = y_2|X = x_2) = 1/2$$

$$p(Y = y_1|X = x_3) = 1/4$$

$$p(Y = y_2|X = x_3) = 3/4$$

Definition of Conditional Probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$



Definition of Conditional Probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

The Chain Rule

$$P(A, B) = P(A|B) \cdot P(B)$$

Chain rule

Chain rule

For all \mathbf{x} we have that

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1})$$

** it holds for any ordering of X_1, \dots, X_n

- ▶ Joint distribution as a product of conditional probabilities
- ▶ Example:

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

Exercise

Chain rule

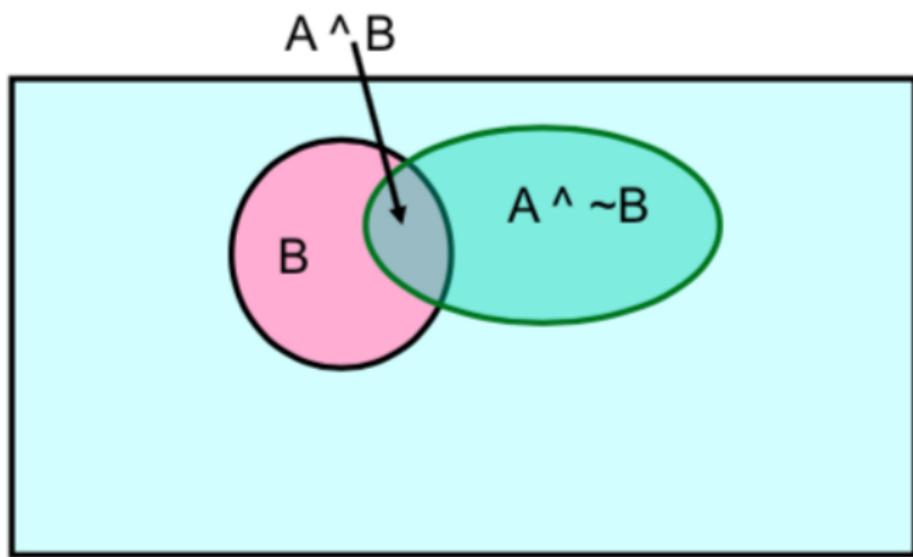
Let be $\mathbf{X} = (X_1, X_2, X_3, X_4)$, prove that for all \mathbf{x} we have that

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

** Remember that $p(x_i|x_1, \dots, x_{i-1}) = \frac{p(x_1, \dots, x_i)}{p(x_1, \dots, x_{i-1})}$

Bayes rule

2 expressions for $P(A, B)$



Bayes rule

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

We call $P(A)$ the “prior”

and $P(A|B)$ the “posterior”



Bayes, Thomas (1763) An essay
towards solving a problem in the doctrine
of chances. *Philosophical Transactions of
the Royal Society of London*, 53:370-418

This is also the Bayes rule

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

$$P(A|B, C) = \frac{P(B|A, C) \cdot P(A, C)}{P(B, C)}$$

Exercise

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

A : you have a flu, B : you just coughed

Assume:

- $P(A) = 0,05$
- $P(B|A) = 0,80$
- $P(B|\neg A) = 0,2$

what is you have a flu given that you just coughed?

Independent events

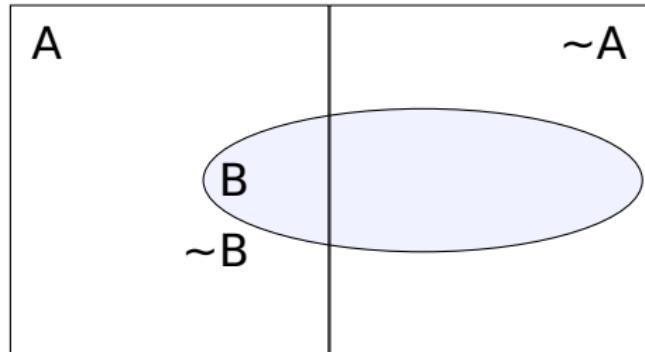
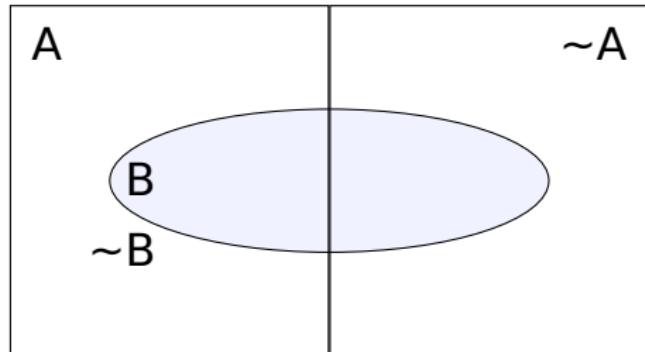
Definition

Two events A and B are independent if

$$P(A, B) = P(A) \cdot P(B)$$

** Intuition **

Knowing A tells us nothing about the value of B (and vice versa)



Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

Joint distribution

What does all this have to do with
function approximation?

instead of $F : X \rightarrow Y$,

learn $P(Y|X)$

Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values
(M variables $\rightarrow 2^M$ combinations)

| A | B | C |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |

Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values (M variables $\rightarrow 2^M$ combinations)
2. Say how probable each combination is

| A | B | C | Prob |
|---|---|---|------|
| 0 | 0 | 0 | 0.30 |
| 0 | 0 | 1 | 0.05 |
| 0 | 1 | 0 | 0.10 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.10 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.10 |

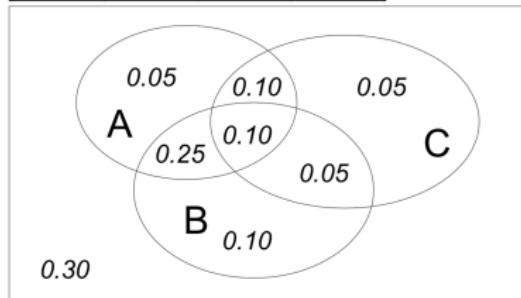
Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values (M variables $\rightarrow 2^M$ combinations)
2. Say how probable each combination is

Subscribed to the axioms of probability if sum to 1

| A | B | C | Prob |
|---|---|---|------|
| 0 | 0 | 0 | 0.30 |
| 0 | 0 | 1 | 0.05 |
| 0 | 1 | 0 | 0.10 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.10 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.10 |



Exercise

Using the joint distribution

You can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{r: \text{ rows matching } E} P(r)$$

| gender | hours_worked | wealth | |
|--------|--------------|--------|-----------|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Exercise

Using the joint distribution

$$P(E) = \sum_{r: \text{rows matching } E} P(r)$$

$$P(\text{Poor} \wedge \text{Male}) = ?$$

| gender | hours_worked | wealth | |
|--------|--------------|--------|-----------|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Exercise

Using the joint distribution

$$P(E) = \sum_{r: \text{rows matching } E} P(r)$$

$$P(\text{Poor}) = ?$$

| gender | hours_worked | wealth | |
|--------|--------------|--------|-----------|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Inference with the joint distribution

You can ask for the probability
of any logical expression
involving a subset of attributes
given another expression
involving other attributes

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$
$$= \frac{\sum_{r: \text{ rows matching } E_1 \text{ & } E_2} P(r)}{\sum_{o: \text{ rows matching } E_2} P(o)}$$

| gender | hours_worked | wealth | |
|--------|--------------|--------|-----------|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Exercise

Inference with the joint distribution

You can ask for the probability
of any logical expression
involving a subset of attributes
given another expression
involving other attributes

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{r: \text{ rows matching } E_1 \text{ & } E_2} P(r)}{\sum_{o: \text{ rows matching } E_2} P(o)}$$

| gender | hours_worked | wealth | |
|--------|--------------|--------|-----------|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

$$P(\text{Male}|\text{Poor}) = ?$$

Learning and the joint distribution

Suppose we want to learn the function $f : \langle G, H \rangle \rightarrow W$

Equivalently, $P(W|G, H)$

Solution:

- ▶ Learn joint distribution from train data
- ▶ Calculate $P(W|G, H)$ for test data

E.g., $\arg \max_{w \in \{rich, poor\}} P(W = w | G = female, H = 40, 5 -)$

Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

Solution?

$P(Y|X)$ sounds like a nice alternative solution to learning
 $F : X \rightarrow Y$

Are we done?

Solution?

$P(Y|X)$ sounds like a nice alternative solution to learning
 $F : X \rightarrow Y$

Are we done?

Main problem

Learning $P(Y|X)$ may require more data than we have

Solution?

$P(Y|X)$ sounds like a nice alternative solution to learning
 $F : X \rightarrow Y$

Are we done?

Main problem

Learning $P(Y|X)$ may require more data than we have

E.g., consider learning the joint distribution for 100 binary variables

- ▶ # of rows in this table?
- ▶ # of data samples to learn faithfully?
- ▶ # of rows never observed?

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
 - ▶ Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
 - ▶ maximum likelihood estimates
 - ▶ maximum a posteriori estimates

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
 - ▶ Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
 - ▶ maximum likelihood estimates
 - ▶ maximum a posteriori estimates

From definition to representation

Joint distribution

Let \mathbf{X}_V be a set of variable, then

$$\forall \mathbf{x}_V, p(\mathbf{x}_V) = p(x_1, \dots, x_v)$$

- ▶ A mapping: $\mathbf{x}_V \mapsto [0, 1]$
- ▶ $\sum_{\mathbf{x}_V} p(\mathbf{x}_V) = 1$
- ▶ Number of free parameters: $\prod_{i=1}^v r_i - 1 = r_V - 1$
- ▶ Exponential in the number of variables, v

From definition to representation

Marginal distribution

Let A and B be a partition of V . Then,

$$\forall \mathbf{x}_A, \mathbf{x}_B, p(\mathbf{x}_A) = \sum_{\mathbf{x}_B} p(\mathbf{x}_A, \mathbf{x}_B)$$

- ▶ Number of **free parameters**: $\prod_{i \in A} r_i - 1 = r_A - 1$

From definition to representation

Conditional distribution

Let A and B be a partition of V

$$\forall \mathbf{x}_A, p(\mathbf{x}_A | \mathbf{x}_B) = \frac{p(\mathbf{x}_A, \mathbf{x}_B)}{p(\mathbf{x}_B)} = \frac{p(\mathbf{x}_A, \mathbf{x}_B)}{\sum_{\mathbf{x}_A} p(\mathbf{x}_A, \mathbf{x}_B)}$$

- ▶ Family of distribution:
A marginal distribution for each value assignment $\mathbf{X}_B = \mathbf{x}_B$
 $\sum_{\mathbf{x}_A} p(\mathbf{x}_A | \mathbf{x}_B) = 1$
- ▶ Number of free parameters:
 $(\prod_{i \in A} r_i - 1) \cdot (\prod_{j \in B} r_j) = (r_A - 1) \cdot r_B$

Exercise

| x_1 | x_2 | $p(x_1, x_2)$ |
|-------|--------|---------------|
| - | banana | 0.1 |
| - | apple | 0.3 |
| - | pear | 0.2 |
| + | banana | 0.1 |
| + | apple | 0.2 |
| + | pear | 0.1 |

- Determine the domains of X_1 and X_2 .
- Obtain the marginal distributions $p(X_1)$ y $p(X_2)$
- Obtain the conditional distributions $p(X_1|X_2 = \text{apple})$ y $p(X_2|X_1 = +)$
- Compute the number of free parameters $p(X_1)$, $p(X_1|X_2)$ y $p(X_1, X_2)$

Conditional independence

A qualitative relationship between random variables

Let A, B, C be disjoint subsets of $V = \{1, \dots, v\}$. We say that X_A is independent from X_B given X_C if and only if for all (x_A, x_B, x_C) we have that $p(x_A|x_B, x_C) = p(x_A|x_C)$.

- ▶ Denoted by $X_A \perp\!\!\!\perp X_B | X_C$
- ▶ $p(x_A|x_B, x_C) = p(x_A|x_C)$: Knowing/observing/fixing x_C , the value x_B does not modify the probability of x_A
- ▶ Exercise: Prove that
$$X_A \perp\!\!\!\perp X_B | X_C \Rightarrow p(x_A, x_B | x_C) = p(x_A | x_C) \cdot p(x_B | x_C)$$

Conditional independence

- ▶ Allow to simplify the factorization given by the chain rule
- ▶ Choose an appropriate ordering that allows to apply the independence over a conditional distribution

Example

- ▶ $\mathbf{X} = X_1, \dots, X_5$
- ▶ $3 \perp\!\!\!\perp 4 | 1, 5$
- ▶ Ordering: 1, 4, 5, 3, 2

$$\begin{aligned} p(\mathbf{X}) &= p(\mathbf{X}_{1,4,5}) p(X_3 | \mathbf{X}_{1,4,5}) p(X_2 | \mathbf{X}_{1,3,4,5}) \\ &= p(\mathbf{X}_{1,4,5}) p(X_3 | \mathbf{X}_{1,5}) p(X_2 | \mathbf{X}_{1,3,4,5}) \end{aligned}$$

Exercise

True or false:

- ▶ $X_A \perp\!\!\!\perp X_B \Rightarrow p(\mathbf{x}_A|\mathbf{x}_B) = p(\mathbf{x}_A)$
- ▶ $X_A \perp\!\!\!\perp X_B \Rightarrow p(\mathbf{x}_A|\mathbf{x}_B, \mathbf{x}_C) = p(\mathbf{x}_A|\mathbf{x}_C)$
- ▶ $X_A \perp\!\!\!\perp X_B \Rightarrow p(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_C) = p(\mathbf{x}_A|\mathbf{x}_C)$
- ▶ $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_C) = p(\mathbf{x}_A|\mathbf{x}_C)p(\mathbf{x}_B|\mathbf{x}_C)$
- ▶ $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\mathbf{x}_A|\mathbf{x}_B, \mathbf{x}_C) = p(\mathbf{x}_A|\mathbf{x}_C)p(\mathbf{x}_B|\mathbf{x}_C)$
- ▶ $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_C) = p(\mathbf{x}_A|\mathbf{x}_C)p(\mathbf{x}_B|\mathbf{x}_A, \mathbf{x}_C)$

Counting the parameters

Conditional independences reduce the number of (free) parameters

- ▶ $X_A \perp\!\!\!\perp X_B$
 - ▶ $\forall \mathbf{x}: p(\mathbf{x}_A, \mathbf{x}_B) = p(\mathbf{x}_A)p(\mathbf{x}_B)$
 - ▶ From $r_A \cdot r_B - 1$ to $r_A - 1 + r_B - 1$

E.g., $r_A = 3, r_B = 4$

From $3 \times 4 - 1$ to $2 + 3$

- ▶ $X_A \perp\!\!\!\perp X_B | X_C$
 - ▶ $\forall \mathbf{x}: p(\mathbf{x}_A, \mathbf{x}_B | \mathbf{x}_C) = p(\mathbf{x}_A | \mathbf{x}_C)p(\mathbf{x}_B | \mathbf{x}_C)$
 - ▶ From $(r_A \cdot r_B - 1) \cdot r_C$ to $(r_A - 1 + r_B - 1) \cdot r_C$

E.g., $r_A = 3, r_B = 4, r_C = 3$

From $(3 \times 4 - 1) \times 3$ to $(2 + 3) * 3$

Counting the parameters, without independence

$p(A, B)$: three parameters

| | | | |
|---|-------|-------|-------|
| | + | - | |
| + | ■ | ■ | ■ + ■ |
| - | ■ | ■ | ■ + ■ |
| | ■ + ■ | ■ + ■ | |

Counting the parameters, without independence

$p(A, B)$: three parameters

| | | | |
|---|-------|-------------------|---------------|
| | + | - | |
| + | ■ | □ | ■ + □ |
| - | □ | $1 - (■ + □ + △)$ | $1 - (■ + □)$ |
| | ■ + □ | $1 - (■ + □)$ | |

Counting the parameters, with independence

$p(A, B)$: two parameters

| | | | |
|---|-------|-------|---|
| | + | - | |
| + | ■ × ■ | ■ × ■ | ■ |
| - | ■ × ■ | ■ × ■ | ■ |

■ ■

Counting the parameters, with independence

$p(A, B)$: two parameters

| | | | |
|---|------------------|------------------|-------|
| | + | - | |
| + | ■ × ■ (1-■) | ■ × (1-■) | ■ |
| - | (1-■) × (1-■) | (1-■) × (1-■) | (1-■) |

■ (1-■)

Conditional independence

Discarding parameters

The complexity of a statistical model can be understood as its **flexibility** for learning or its **requirement of memory**

Conditional independences:

- ▶ Reduce the **complexity** (i.e., # parameters) of the statistical model represented by the (simplified) chain rule
- ▶ Allow for **avoiding** to model (irrelevant) parameters associated to **soft conditional dependences**
- ▶ Help to deal with the **trade-off** between the **complexity** of the statistical model and amount of **train data**
This has crucial implications in statistical models, e.g., **overfitting**

Exercise

Let be $r_A = 6, r_B = 4, r_C = 8$.

- ▶ $\forall \mathbf{x}, p(\mathbf{X}) = p(\mathbf{X}_A, \mathbf{X}_B, \mathbf{X}_C)$
 - ▶ $\forall \mathbf{x}, p(\mathbf{X}) = p(\mathbf{X}_A)p(\mathbf{X}_B)p(\mathbf{X}_C)$
 - ▶ $\forall \mathbf{x}, p(\mathbf{X}) = p(\mathbf{X}_A|\mathbf{X}_B)p(\mathbf{X}_B)p(\mathbf{X}_C)$
 - ▶ $\forall \mathbf{x}, p(\mathbf{X}) = p(\mathbf{X}_A|\mathbf{X}_C)p(\mathbf{X}_B|\mathbf{X}_C)p(\mathbf{X}_C)$
 - ▶ $\forall \mathbf{x}, p(\mathbf{X}) = p(\mathbf{X}_A)p(\mathbf{X}_B|\mathbf{X}_A)p(\mathbf{X}_C|\mathbf{X}_A, \mathbf{X}_B)$
1. Calculate the number of free parameters
 2. Read conditional independences from factorizations

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
 - ▶ Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
 - ▶ maximum likelihood estimates
 - ▶ maximum a posteriori estimates

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
 - ▶ Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
 - ▶ maximum likelihood estimates
 - ▶ maximum a posteriori estimates

You should know

- ▶ Events

Discrete random variables, continuous random variables, compound events

- ▶ Axioms of probability

What defines a reasonable theory of uncertainty

- ▶ Conditional probabilities

- ▶ Chain rule

- ▶ Bayes rule

- ▶ Joint probability distribution

How to calculate other quantities from the joint distribution

- ▶ $P(Y|X)$: Facing practical problems

- ▶ Conditional independencies for model simplification

- ▶ Principles of parameter estimation

Notation

- ▶ Indices $V = \{1, \dots, n\}$, subsets of indices $A, B, C \subseteq V$
- ▶ Random variables $\mathbf{X} = (X_1, \dots, X_n)$, $\mathbf{X}_A = (X_i : i \in A)$
- ▶ Domain of X_i , Ω_i con $|\Omega_i| = r_i$
- ▶ Instance/case/sample: $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{x}_A = (X_i : i \in A)$
- ▶ Data set: $D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$
- ▶ Probability distribution: \mathbf{X} distributed according to $p(\mathbf{X})$
- ▶ Probability of observing \mathbf{x} : $p(\mathbf{x})$

Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

Bayesian Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Concepts

- ▶ Joint/Marginal/Conditional probability distribution
- ▶ Factorization
- ▶ Chain rule
- ▶ Bayes rule
- ▶ (Conditional) independence

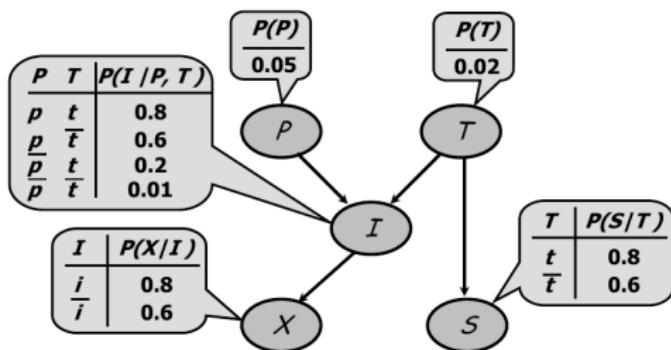
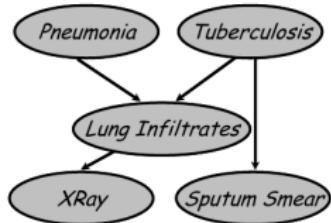
Example: conditional independence

Given a set of variables $\mathbf{X} = X_1, \dots, X_5$, and $3 \perp\!\!\!\perp 4 | 1, 5$

$$\begin{aligned} p(\mathbf{X}) &= p(X_2 | \mathbf{X}_{1,3,4,5}) p(\mathbf{X}_3 | \mathbf{X}_{1,4,5}) p(\mathbf{X}_{1,4,5}) \\ &= p(X_2 | \mathbf{X}_{1,3,4,5}) p(\mathbf{X}_3 | \mathbf{X}_{1,5}) p(\mathbf{X}_{1,4,5}) \end{aligned}$$

Bayesian networks

- ▶ What is a Bayesian network?
- ▶ Chain rule.
- ▶ What does it mean that a probability distribution factorizes over a graph \mathcal{G} ?



Bayesian network

Components

Directed acyclic graph $G = (V, E)$ + parameters $\Theta = (\Theta_1, \dots, \Theta_n)$

- ▶ Represents a joint probability distribution:

Vertices: related to random variables

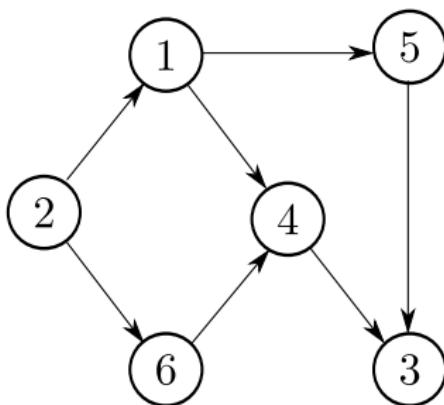
Edges: related to the simplification of the chain rule

Parameters: tables of the conditional probability distributions

Directed acyclic graph (DAG)

Formalism

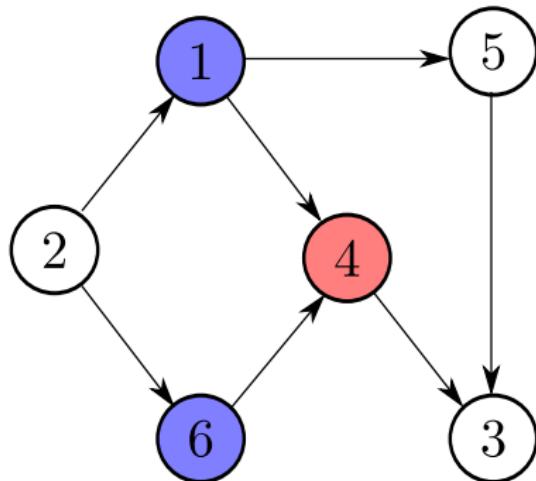
- ▶ A DAG G is a pair (V, E)
- ▶ $V = \{1, \dots, n\}$ represent the set of vertices
- ▶ $E = \{(u, v) : u, v \in V, u \neq v\}$ represent the set of arcs
- ▶ There are no directed cycles in G



DAG concepts

Definitions

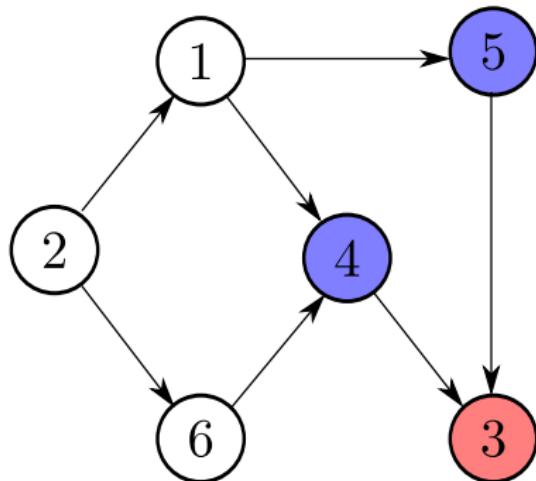
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

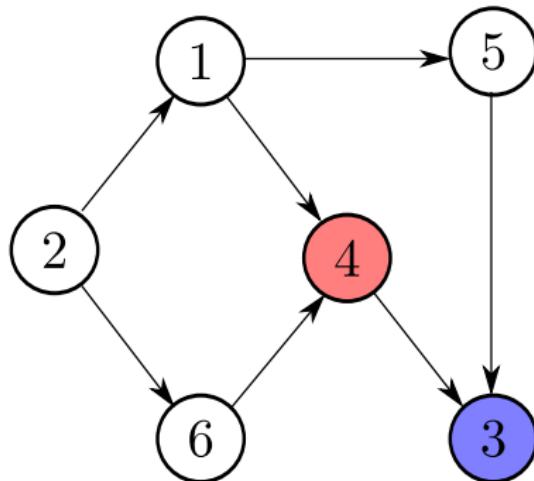
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

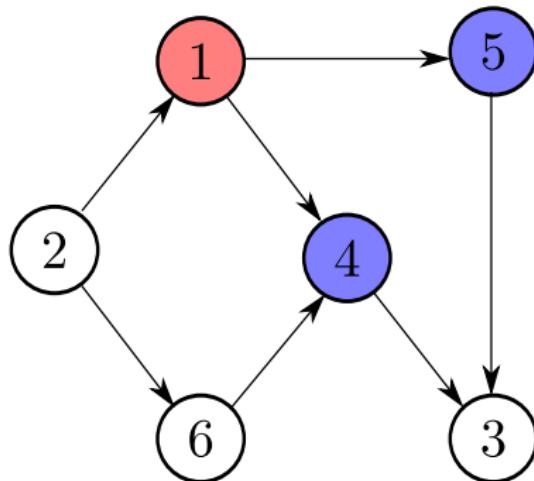
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

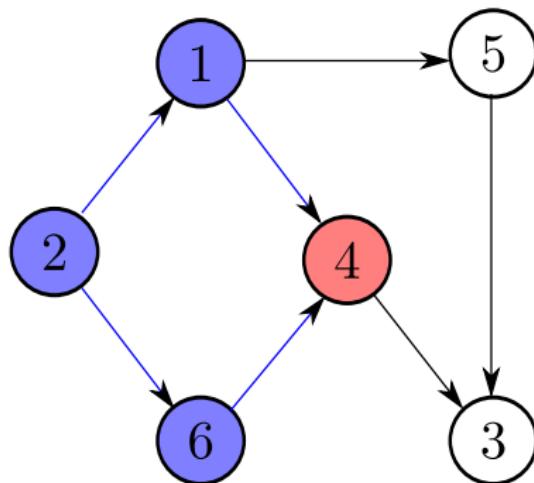
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

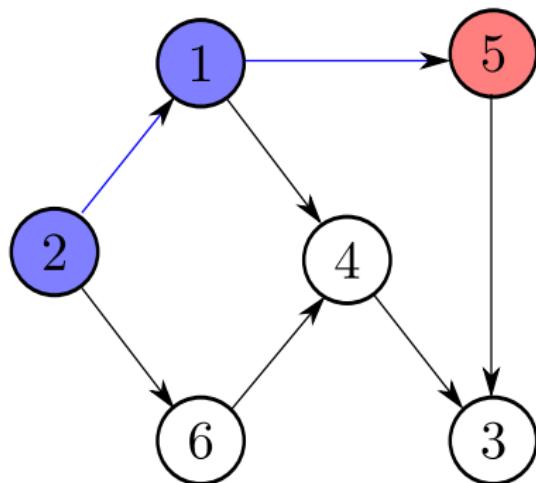
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

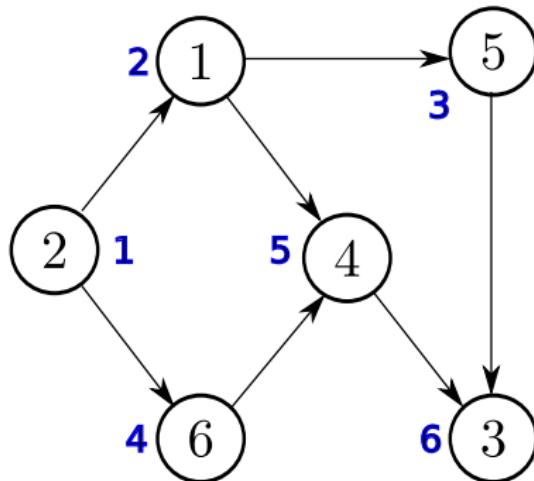
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

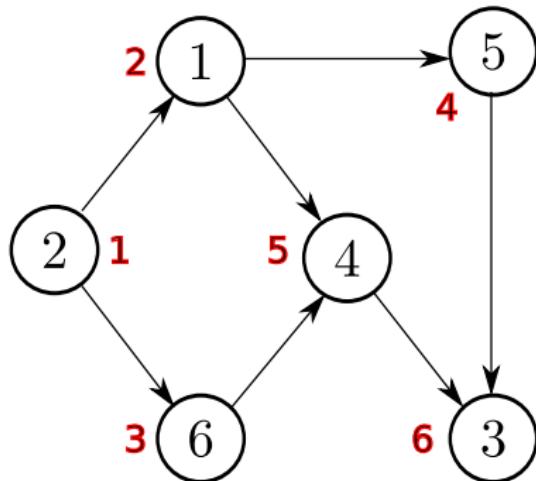
Parents, children, ancestral set, **ancestral ordering**, moral



DAG concepts

Definitions

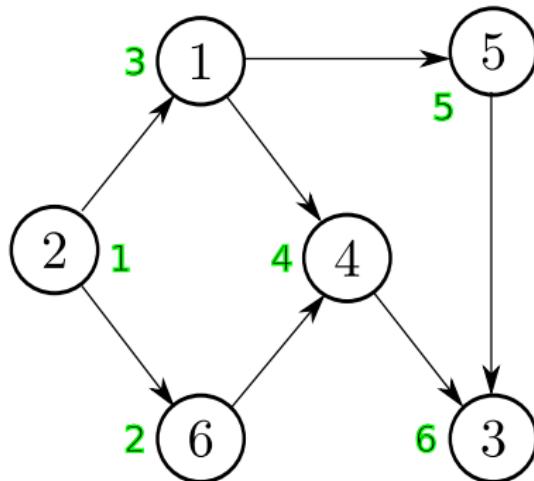
Parents, children, ancestral set, **ancestral ordering**, moral



DAG concepts

Definitions

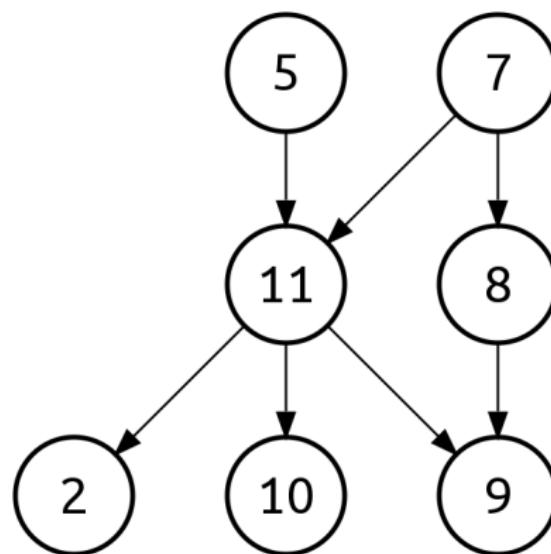
Parents, children, ancestral set, **ancestral ordering**, moral



DAG concepts

Definitions

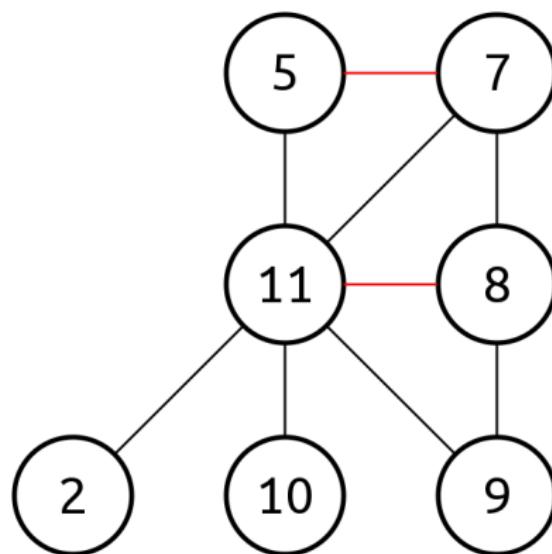
Parents, children, ancestral set, ancestral ordering, moral



DAG concepts

Definitions

Parents, children, ancestral set, ancestral ordering, moral



Exercise

Let be $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$

Let $G_1 = (\mathbf{X}, E_1)$, where

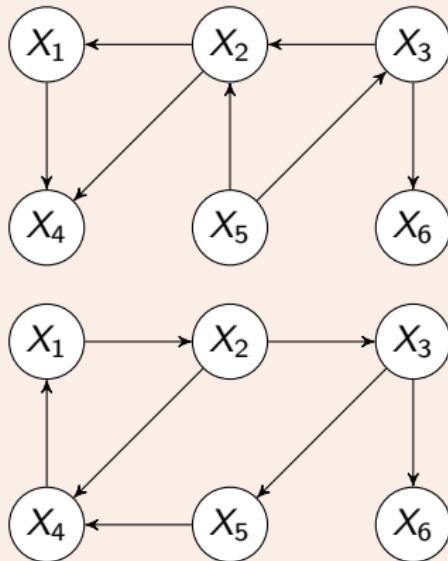
$$E_1 = \{(X_3, X_2), (X_5, X_3), (X_2, X_1), (X_5, X_2), (X_3, X_6), (X_2, X_4), (X_1, X_4)\}$$

Let $G_2 = (\mathbf{X}, E_2)$, where

$$E_2 = \{(X_2, X_3), (X_3, X_5), (X_3, X_6), (X_2, X_4), (X_5, X_4), (X_4, X_1), (X_1, X_2)\}$$

1. Are they DAGs?
2. Obtain the parents and children of X_3 and X_5
3. Find an ancestral ordering for G_1 and G_2

Exercise



Bayesian networks

Factorization

A Bayesian network $M = (G, \Theta)$ can be expressed as a product of conditional probability distribution:

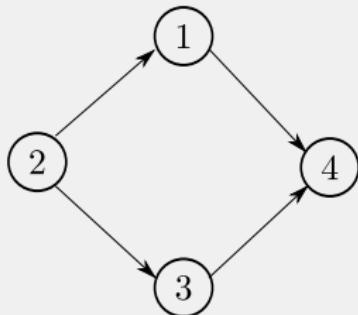
$$p_M(\mathbf{x}) = \prod_{i=1}^n p(x_i | \mathbf{pa}_i; \Theta_i)$$

where \mathbf{pa}_i relates to the set of variables with an edge towards X_i .

- ▶ Qualitative: G describes the **skeleton** of the factorization
- ▶ Quantitative: Θ determines the **shape** of the conditional distributions

Bayesian networks

Factorization



Bayesian network's factorization:

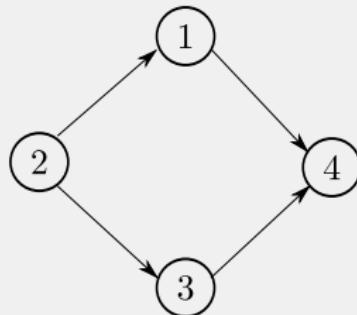
$$p(\mathbf{X}) = p(X_4|X_1, X_3) \cdot p(X_1|X_2) \cdot p(X_3|X_2) \cdot p(X_2)$$

Chain rule:

$$p(\mathbf{X}) = p(X_4|X_1, \textcolor{red}{X_2}, X_3) \cdot p(X_1|X_2, \textcolor{red}{X_3}) \cdot p(X_3|X_2) \cdot p(X_2)$$

Bayesian networks

Factorization



Conditional independencies

- ▶ $X_4 \perp\!\!\!\perp X_2 | X_1, X_3$
- ▶ $X_1 \perp\!\!\!\perp X_3 | X_2$

Bayesian network's factorization:

$$p(\mathbf{X}) = p(X_4|X_1, X_3) \cdot p(X_1|X_2) \cdot p(X_3|X_2) \cdot p(X_2)$$

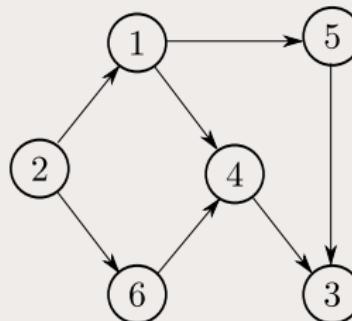
Chain rule:

$$p(\mathbf{X}) = p(X_4|X_1, \textcolor{red}{X_2}, X_3) \cdot p(X_1|X_2, \textcolor{red}{X_3}) \cdot p(X_3|X_2) \cdot p(X_2)$$

Exercise

Questions:

1. Find an ancestral ordering
2. Determine the factorized chain rule

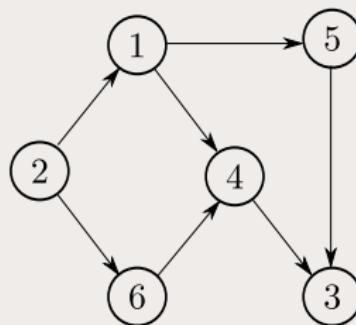


Exercise:

A possible answer

1. Find an ancestral ordering: 2, 1, 6, 4, 5, 3
2. Determine the factorized chain rule

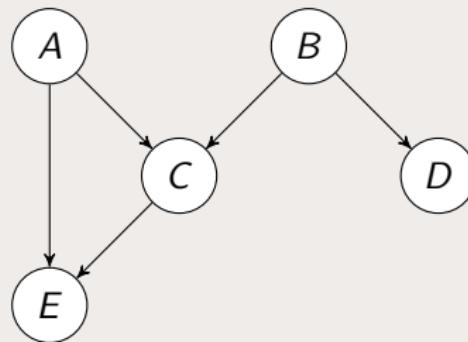
$$p_M(\mathbf{x}) = p(x_1|x_2) \cdot p(x_2) \cdot p(x_3|x_4, x_5) \cdot p(x_4|x_1, x_6) \cdot p(x_5|x_1) \cdot p(x_6|x_2)$$



Exercise:

Factorization

Given the model, which of these is an appropriate decomposition of the joint distribution?



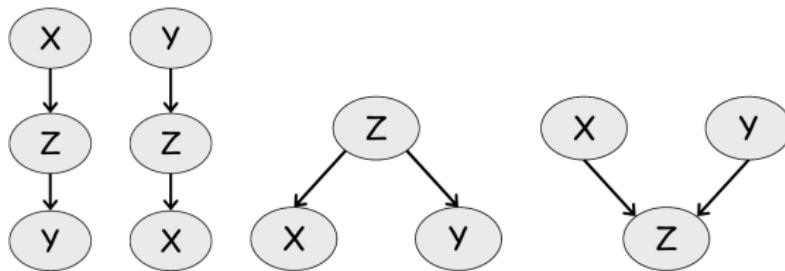
- a) $P(A, B, C, D, E) = P(A)P(B)P(A, B|C)P(B|D)P(A, C|E)$
- b) $P(A, B, C, D, E) = P(A)P(B)P(C|A)P(C|B)P(D|B)P(E|A)P(E|C)$
- c) $P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B)P(E|A, C)$
- d) $P(A, B, C, D, E) = P(A)P(B)P(C)P(D)P(E)$

Bayesian Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Flow of probabilistic influence

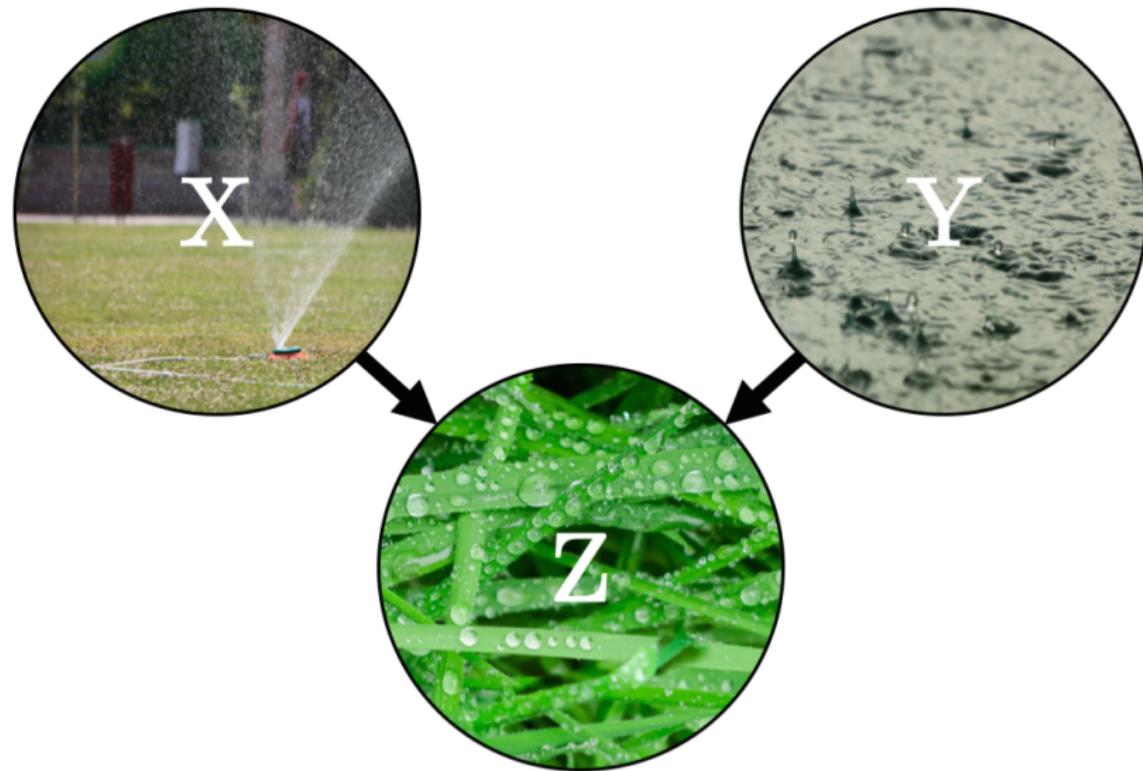


In Bayesian networks, influence flow is stopped by observed nodes and non-observed **v-structures**.

A v-structure is observed if Z or any of its descendants is observed.

Flow of probabilistic influence

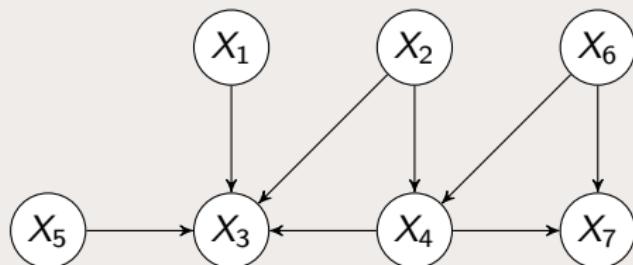
The wet grass example



Z indicates whether our garden is wet; Possible explanations are: X , we turned on the sprinklers; Y , rain. If grass is wet ($Z = \text{True}$) and we didn't turn on the sprinklers ($X = \text{False}$), then the probability of rain ($A = \text{True}$) increases!

Exercise

Flow of influence



True or false:

- ▶ $X_5 \perp\!\!\!\perp X_3$
- ▶ $X_6 \perp\!\!\!\perp X_3$
- ▶ $X_1 \perp\!\!\!\perp X_5$
- ▶ $X_3 \perp\!\!\!\perp X_7$
- ▶ $X_1 \perp\!\!\!\perp X_5 \mid X_3$
- ▶ $X_6 \perp\!\!\!\perp X_2 \mid X_3$

Flow of probabilistic influence

Active trail

- ▶ Let \mathcal{G} be a DAG
- ▶ Let $X_1 \rightleftharpoons \dots \rightleftharpoons X_m$ be a **trail** in \mathcal{G}
- ▶ A trail is **active** given a set of observed variables Z if:
 1. Whenever there is a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i or one of its descendants is in Z
 2. no other node along the trail is in Z

d -separation

- ▶ Let X , Y and Z be three disjoint sets of variables in \mathcal{G} .
- ▶ Z **d -separates** X from Y in \mathcal{G} if $X \perp\!\!\!\perp Y | Z$ holds in \mathcal{G}
- ▶ $X \perp\!\!\!\perp Y | Z$ holds in \mathcal{G} if there is no active trail between any variable in X and any variable in Y given Z .

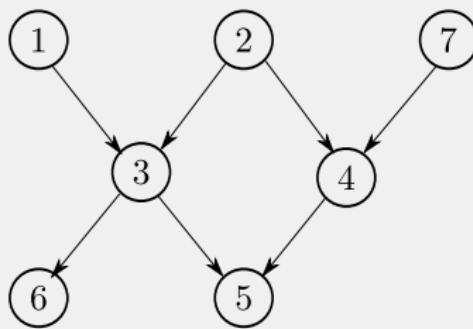
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



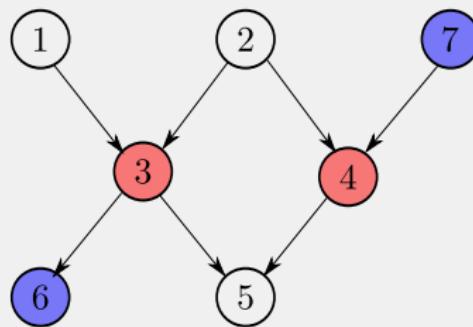
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$X_6 \perp\!\!\!\perp X_7 | X_3, X_4$?

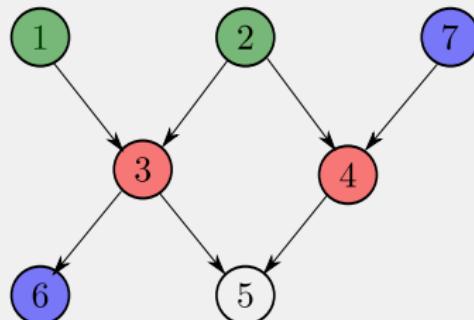
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_6 \perp\!\!\!\perp X_7 \mid X_3, X_4?$$

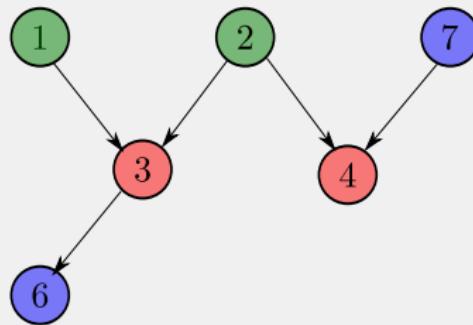
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_6 \perp\!\!\!\perp X_7 \mid X_3, X_4?$$

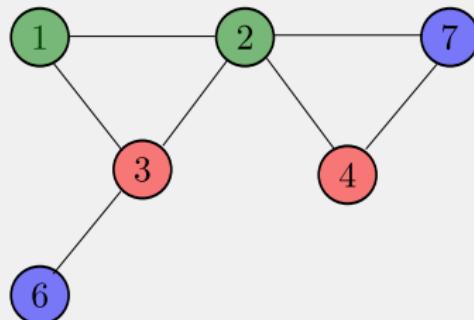
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_6 \perp\!\!\!\perp X_7 \mid X_3, X_4? \text{ true}$$

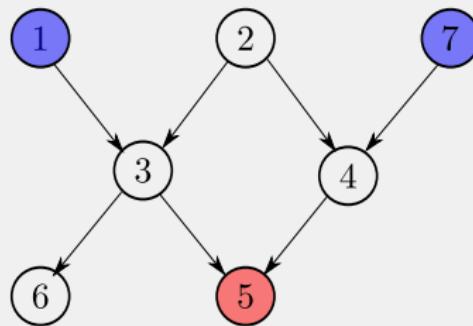
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_5?$$

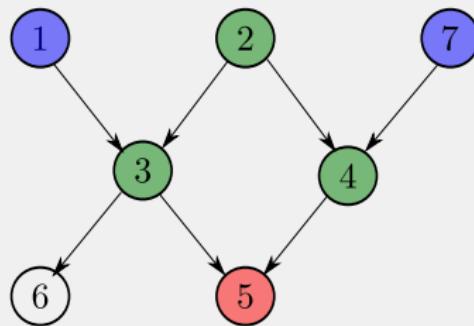
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_5?$$

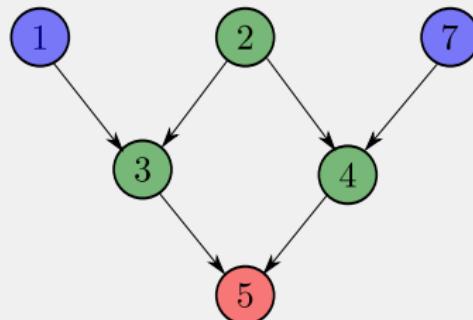
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_5?$$

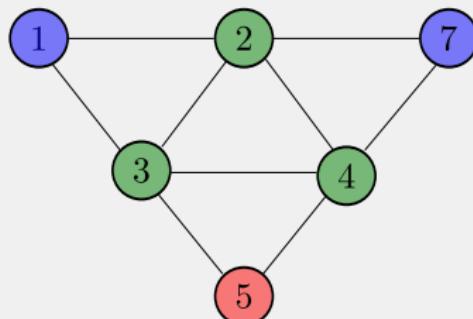
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$X_1 \perp\!\!\!\perp X_7 | X_5$? false

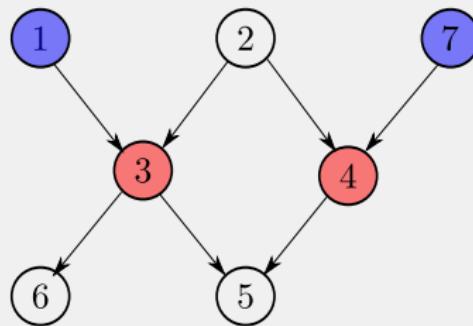
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$X_1 \perp\!\!\!\perp X_7 | X_3, X_4$?

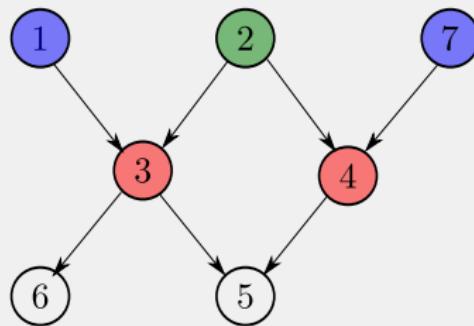
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_3, X_4?$$

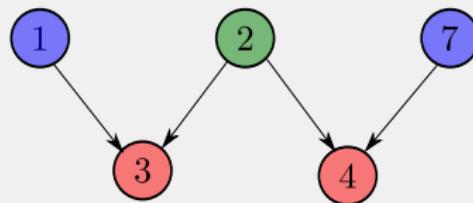
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_3, X_4?$$

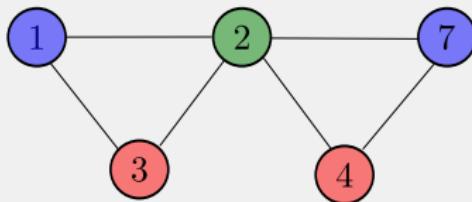
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$X_1 \perp\!\!\!\perp X_7 | X_3, X_4$? **false**

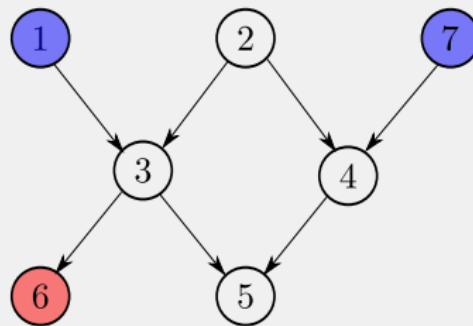
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_6?$$

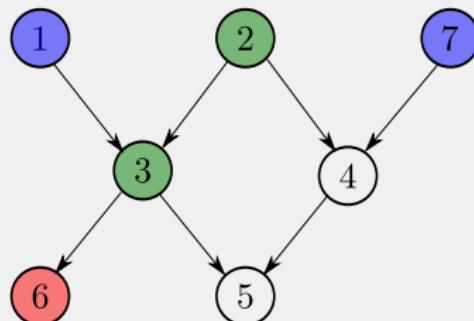
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_6?$$

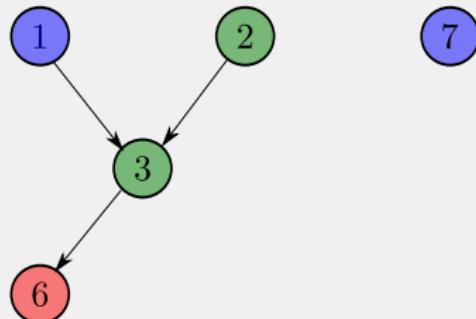
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_6?$$

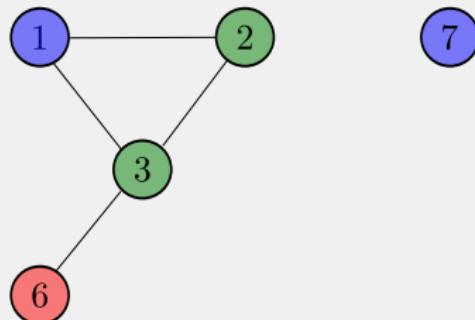
Graphical criteria of d -separation

Z d -separates X from Y ?

Three steps:

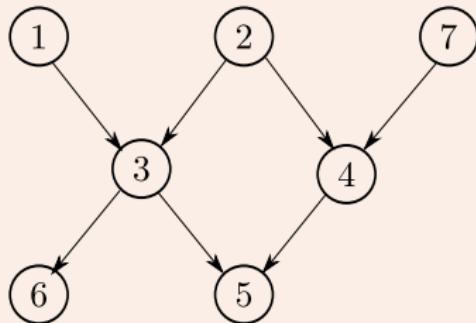
1. Identify the ancestors of X , Y and Z
2. Remove the rest of variables and moralize the left subgraph
3. Does Z block all the paths from X to Y ?

Example



$$X_1 \perp\!\!\!\perp X_7 \mid X_6? \text{ true}$$

Exercise

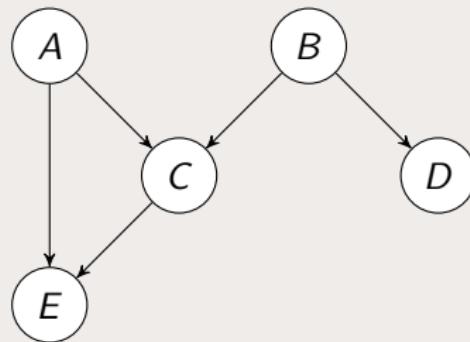


- Check if these terms are verified:
 - $X_1 \perp\!\!\!\perp X_4 | X_5$
 - $X_1 \perp\!\!\!\perp X_5 | X_3, X_4$
 - $X_6 \perp\!\!\!\perp X_7 | X_3$
 - $X_1 \perp\!\!\!\perp X_7 | X_5$
 - $X_1 \perp\!\!\!\perp X_4 | X_6$
 - $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$
- Identify the subsets of variables that verify:
 - $X_1 \perp\!\!\!\perp X_7 | ?$
 - $X_1 \perp\!\!\!\perp X_5 | ?$

Exercise

Independencies in a graph

Which pairs of variables are independent in this BN, given that none of them have been observed?

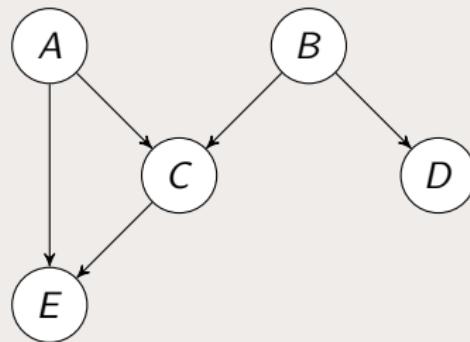


- a) $C \perp\!\!\!\perp D$
- b) $A \perp\!\!\!\perp B$
- c) $B \perp\!\!\!\perp E$
- d) $A \perp\!\!\!\perp C$

Exercise

Independencies in a graph

Now assume that the value of E is observed. Which pairs of variables are independent given E?



- a) $A \perp\!\!\!\perp D|E$
- b) $A \perp\!\!\!\perp C|E$
- c) $A \perp\!\!\!\perp B|E$
- d) $D \perp\!\!\!\perp C|E$
- e) $B \perp\!\!\!\perp D|E$

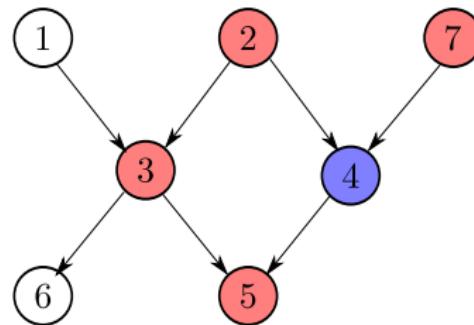
Markov blanket

Definition

- ▶ Parents, children and parents of the children
- ▶ Given X_j and \mathbf{Mb}_j , for any set of variables $\mathbf{X}_c \subseteq (\mathbf{V} \setminus \mathbf{Mb}_j)$:

$$\mathbf{X}_c \perp\!\!\!\perp X_j \mid \mathbf{Mb}_j$$

** Applications for structural learning, feature subset selection,...



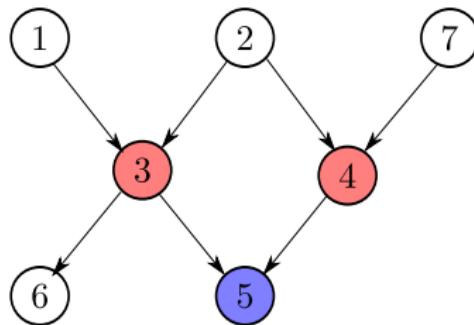
Markov blanket

Definition

- ▶ Parents, children and parents of the children
- ▶ Given X_j and \mathbf{Mb}_j , for any set of variables $\mathbf{X}_c \subseteq (\mathbf{V} \setminus \mathbf{Mb}_j)$:

$$\mathbf{X}_c \perp\!\!\!\perp X_j \mid \mathbf{Mb}_j$$

** Applications for structural learning, feature subset selection,...



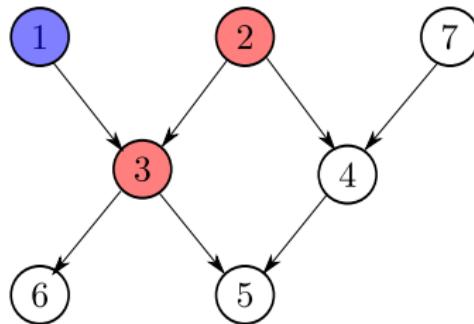
Markov blanket

Definition

- ▶ Parents, children and parents of the children
- ▶ Given X_j and \mathbf{Mb}_j , for any set of variables $\mathbf{X}_c \subseteq (\mathbf{V} \setminus \mathbf{Mb}_j)$:

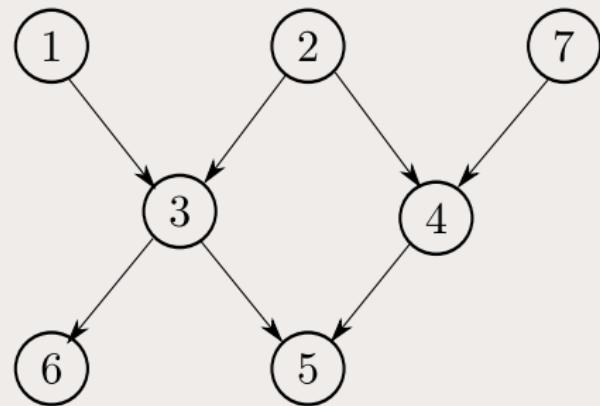
$$\mathbf{X}_c \perp\!\!\!\perp X_j \mid \mathbf{Mb}_j$$

** Applications for structural learning, feature subset selection,...

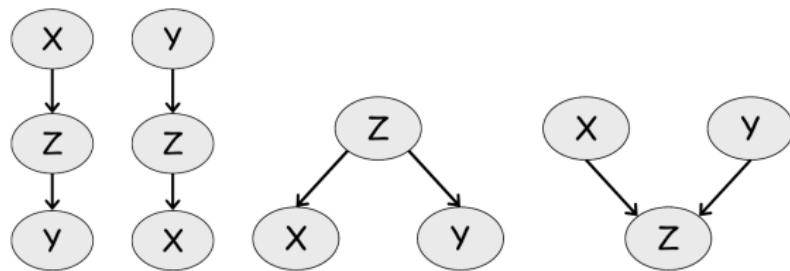


Exercise

Determine the Markov Blanket of X_3



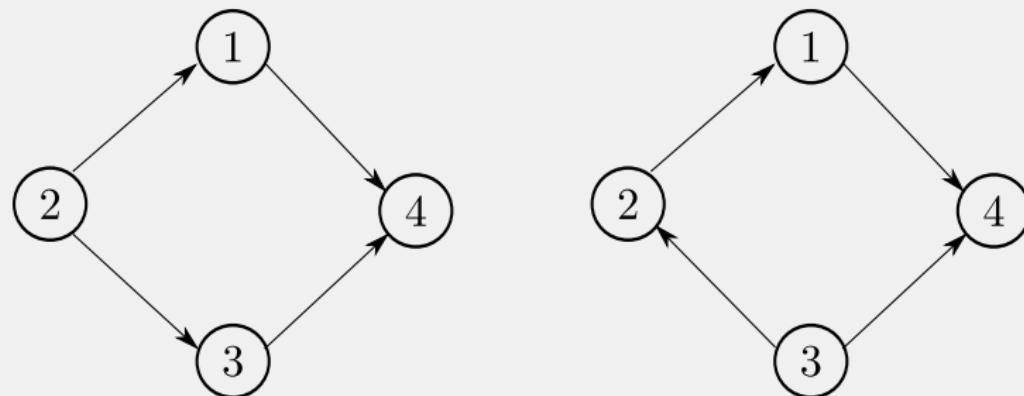
Equivalence classes



Equivalence classes

Redundancy

Two different DAGs can determine the same dependence model



The same undirected edges with the same **V-structures**

** Applications in structural learning

Independency Maps

I-maps

- ▶ $I(\mathcal{G})$ is the set of independence statements that hold on \mathcal{G}
Any distribution that factorizes over \mathcal{G} satisfies the ind. statements in $I(\mathcal{G})$
- ▶ $I(\mathcal{P})$ is the set of independence statements that hold for P
- ▶ \mathcal{G} is an **I-map** of P if $I(\mathcal{G}) \subseteq I(\mathcal{P})$
The other way around is not necessarily true (\sim wasted parameters)

D-separation and I-maps

- ▶ If P factorizes over \mathcal{G} and X and Y are d-separated in \mathcal{G} given Z then P satisfies $X \perp\!\!\!\perp Y|Z$

Exercise

I-maps

Which of the following statements about I?-maps are true? You may select 1 or more options, or none of them.

- a) The graph K that is the same as the graph G , except that all of the edges are oriented in the opposite direction as the corresponding edges in G , is always an I-map for G , regardless of the structure of G .
- b) A graph K is an I-map for a graph G if and only if K encodes all of the independences that G has, and more.
- c) An I-map is a function that maps a graph G to itself, i.e., $f(G) = G$.
- d) A graph K is an I-map for a graph G if and only if all of the independencies encoded by K are also encoded by G .

Connecting the three views

- ▶ Given a Bayesian network structure \mathcal{G} , and a probability distribution P that factorizes according to \mathcal{G} , then P satisfies the independencies that hold in \mathcal{G} .

P factorizes $\implies P$ satisfies independencies

- ▶ Given a Bayesian network structure \mathcal{G} , and a probability distribution P that satisfies the independencies that hold in \mathcal{G} , then P factorizes over \mathcal{G} .

P satisfies independencies $\implies P$ factorizes

Exercise

Model

Let X_1, X_2, \dots, X_5 be binary random variables. Given p such that it verifies the set of conditional independences $I = \{X_2 \perp\!\!\!\perp X_1; X_3 \perp\!\!\!\perp X_2|X_1; X_4 \perp\!\!\!\perp X_1, X_2|X_3; X_5 \perp\!\!\!\perp X_2, X_3|X_1, X_4\}$ and the ancestral ordering 1, 2, 3, 4, 5:

- ▶ Obtain the simplified chain rule
- ▶ Compute the number of free parameters of the model

Exercise

Ancestral ordering that minimizes the number of parameters

Given a BN over a set of variables X_1, \dots, X_5 ,

1. $I = \{X_5 \perp\!\!\!\perp X_1 | X_3; \quad X_2 \perp\!\!\!\perp X_1, X_5 | X_3; \quad X_3 \perp\!\!\!\perp X_1;$
 $X_4 \perp\!\!\!\perp X_3, X_5 | X_1, X_2\}$
2. $I = \{X_4 \perp\!\!\!\perp X_2, X_1 | X_3; \quad X_2 \perp\!\!\!\perp X_1; \quad X_5 \perp\!\!\!\perp X_1, X_4 | X_2, X_3\}$
3. $I = \{X_3 \perp\!\!\!\perp X_1 | X_2; \quad X_4 \perp\!\!\!\perp X_3 | X_2, X_1; \quad X_5 \perp\!\!\!\perp X_1, X_4 | X_2, X_3\}$

Exercise

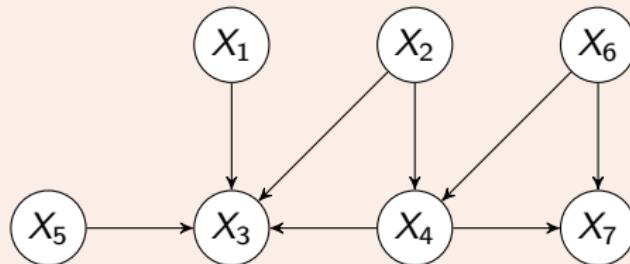
1. Let $P(X)$ be a distribution such that

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2)P(X_3|X_4, X_1)P(X_4|X_2)$$

Are these (conditional) independencies true?

- ▶ $X_1 \perp\!\!\!\perp X_4$
- ▶ $X_2 \perp\!\!\!\perp X_1$
- ▶ $X_2 \perp\!\!\!\perp X_1 | X_3$
- ▶ $X_1 \perp\!\!\!\perp X_2 | X_3, X_4$

2. Given that P satisfies the independencies in



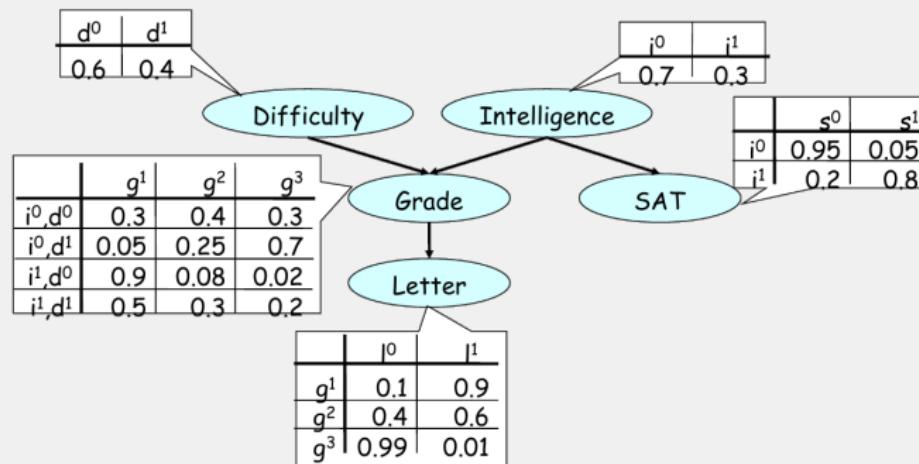
Provide a factorization for P .

Flow of probabilistic influence

Reasoning Patterns

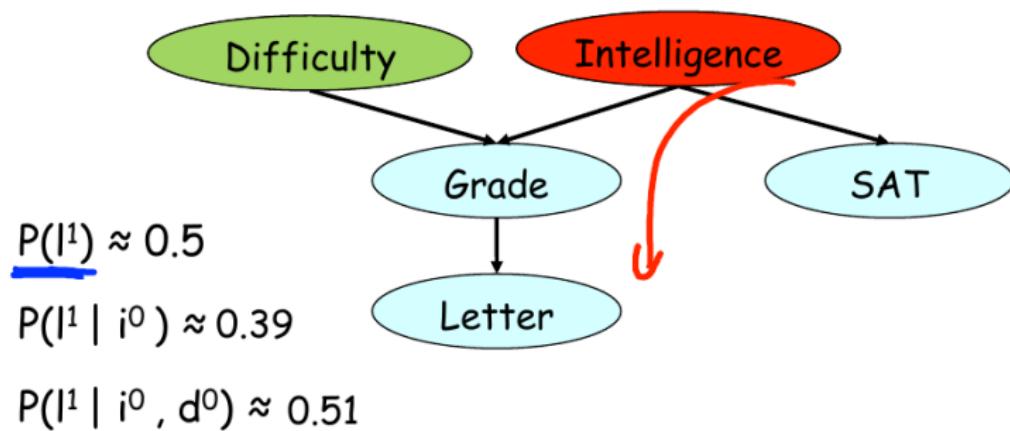
- ▶ Causal reasoning
- ▶ Evidential reasoning
- ▶ Intercausal reasoning

The student network



Flow of probabilistic influence

Reasoning Patterns



Causal reasoning

Flow of probabilistic influence

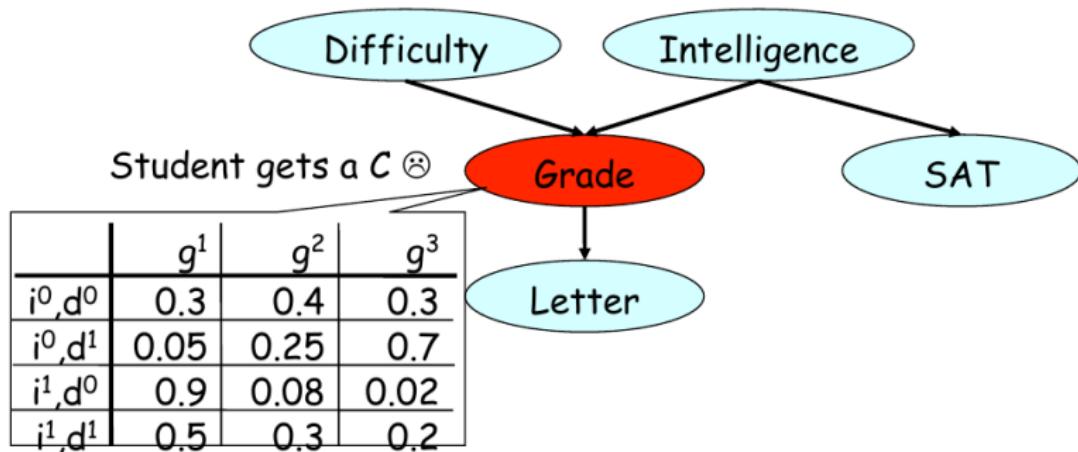
Reasoning Patterns

$$P(d^1) = 0.4$$

$$P(d^1 | g^3) \approx 0.63$$

$$P(i^1) = 0.3$$

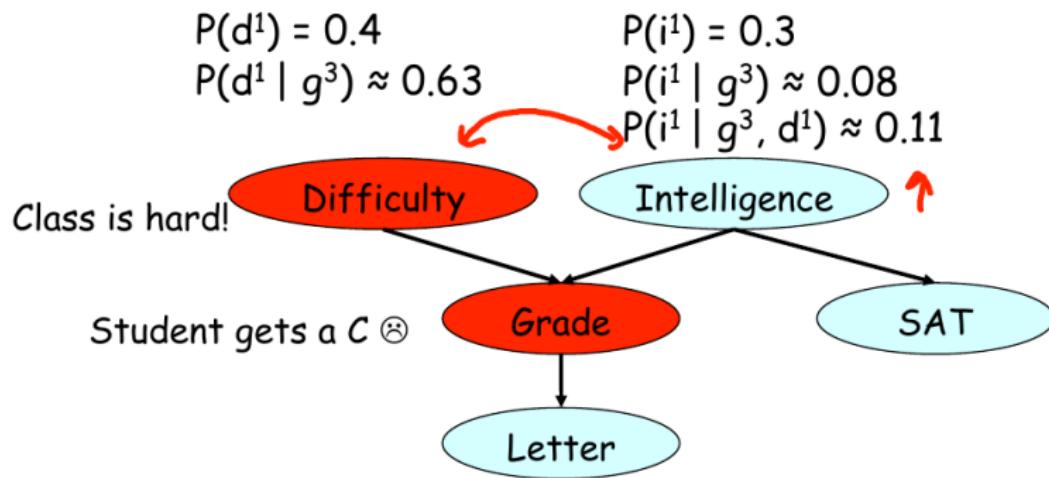
$$P(i^1 | g^3) \approx 0.08$$



Evidential reasoning

Flow of probabilistic influence

Reasoning Patterns

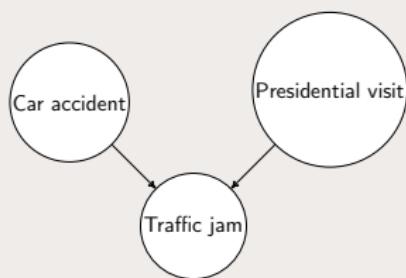


Intercausal reasoning

Exercise

Inter-causal reasoning

In this model for traffic jams in a small town, which we assume can be caused by a car accident, or by a visit from the president.



$$P(PV = \text{True}) = 0.01$$

$$\underline{\underline{P(CA = \text{True}) = 0.1}}$$

$$P(TJ = \text{True} \mid PV = \text{False}, CA = \text{False}) = 0.1$$

$$P(TJ = \text{True} \mid PV = \text{False}, CA = \text{True}) = 0.5$$

$$P(TJ = \text{True} \mid PV = \text{True}, CA = \text{False}) = 0.6$$

$$P(TJ = \text{True} \mid PV = \text{True}, CA = \text{True}) = 0.9$$

- ▶ Calculate $P(CA = \text{True} \mid TJ = \text{True})$
- ▶ Calculate $P(CA = \text{True} \mid TJ = \text{True}, PV = \text{True})$
where TJ, CA and PV stand for Traffic Jam, Car Accident and Presidential Visit respectively.

Bayesian Networks

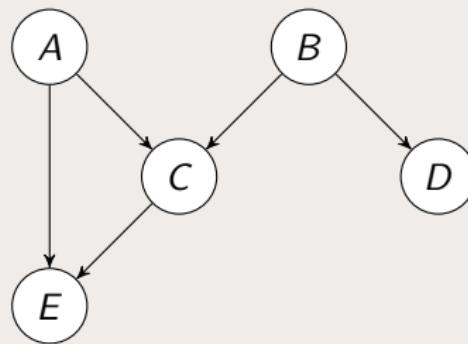
Probabilistic Graphical Models

Jerónimo Hernández-González

Exercise

Independent parameters

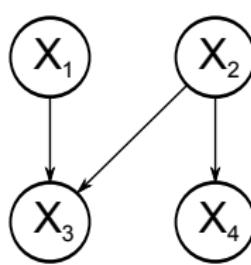
How many independent parameters are required to uniquely define the CPD of E if A, B, and D are binary, and C and E have three values each?



- ▶ 8
- ▶ 11
- ▶ 17
- ▶ 3
- ▶ 18
- ▶ 12
- ▶ 6

Bayesian networks

Number of parameters



Set of probability distributions, $p(X_i | \mathbf{PA}_i)$

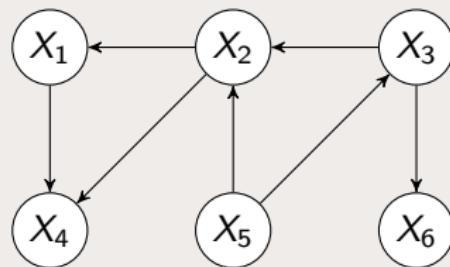
| X_i | $ \Omega_{X_i} $ | \mathbf{PA}_i | $ \Omega_{\mathbf{PA}_i} $ |
|-------|------------------|-----------------|----------------------------|
| X_1 | 3 | \emptyset | - |
| X_2 | 2 | \emptyset | - |
| X_3 | 2 | (X_1, X_2) | 6 |
| X_4 | 2 | X_2 | 2 |

$$\text{No. params} = \sum_i (|\Omega_{X_i}| - 1) \times |\Omega_{\mathbf{PA}_i}|$$

Exercise

Storing a Bayesian network

Let $P(X)$ be a Bayesian network over \mathcal{G} . How much memory do we need to store it?



Bayesian networks

Number of parameters

| X_j | Model parameters |
|-------|---|
| X_1 | $\theta_{1\emptyset 1} = p(x_{11}), \theta_{1\emptyset 2} = p(x_{12}), \theta_{1\emptyset 3} = p(x_{13})$ |
| X_2 | $\theta_{2\emptyset 1} = p(x_{21}), \theta_{2\emptyset 2} = p(x_{22})$ |
| X_3 | $\theta_{311} = p(x_{31} x_{11}, x_{21}), \theta_{321} = p(x_{31} x_{11}, x_{22}), \theta_{331} = p(x_{31} x_{12}, x_{21}),$ $\theta_{341} = p(x_{31} x_{12}, x_{22}), \theta_{351} = p(x_{31} x_{13}, x_{21}), \theta_{361} = p(x_{31} x_{13}, x_{22}),$ $\theta_{312} = p(x_{32} x_{11}, x_{21}), \theta_{322} = p(x_{32} x_{11}, x_{22}), \theta_{332} = p(x_{32} x_{12}, x_{21}),$ $\theta_{342} = p(x_{32} x_{12}, x_{22}), \theta_{352} = p(x_{32} x_{13}, x_{21}), \theta_{362} = p(x_{32} x_{13}, x_{22})$ |
| X_4 | $\theta_{411} = p(x_{41} x_{21}), \theta_{421} = p(x_{41} x_{22}),$ $\theta_{412} = p(x_{42} x_{21}), \theta_{422} = p(x_{42} x_{22})$ |

Bayesian networks

Number of parameters

| X_j | Model parameters |
|-------|---|
| X_1 | $\theta_{1\emptyset 1} = p(x_{11}), \theta_{1\emptyset 2} = p(x_{12}), \theta_{1\emptyset 3} = p(x_{13})$ |
| X_2 | $\theta_{2\emptyset 1} = p(x_{21}), \theta_{2\emptyset 2} = p(x_{22})$ |
| X_3 | $\theta_{311} = p(x_{31} x_{11}, x_{21}), \theta_{321} = p(x_{31} x_{11}, x_{22}), \theta_{331} = p(x_{31} x_{12}, x_{21}),$ $\theta_{341} = p(x_{31} x_{12}, x_{22}), \theta_{351} = p(x_{31} x_{13}, x_{21}), \theta_{361} = p(x_{31} x_{13}, x_{22}),$ $\theta_{312} = p(x_{32} x_{11}, x_{21}), \theta_{322} = p(x_{32} x_{11}, x_{22}), \theta_{332} = p(x_{32} x_{12}, x_{21}),$ $\theta_{342} = p(x_{32} x_{12}, x_{22}), \theta_{352} = p(x_{32} x_{13}, x_{21}), \theta_{362} = p(x_{32} x_{13}, x_{22})$ |
| X_4 | $\theta_{411} = p(x_{41} x_{21}), \theta_{421} = p(x_{41} x_{22}),$ $\theta_{412} = p(x_{42} x_{21}), \theta_{422} = p(x_{42} x_{22})$ |

Bayesian networks

Number of parameters

| X_j | Model parameters |
|-------|---|
| X_1 | $\theta_{1\emptyset 1} = p(x_{11}), \theta_{1\emptyset 2} = p(x_{12}), \theta_{1\emptyset 3} = p(x_{13})$ |
| X_2 | $\theta_{2\emptyset 1} = p(x_{21}), \theta_{2\emptyset 2} = p(x_{22})$ |
| X_3 | $\theta_{311} = p(x_{31} x_{11}, x_{21}), \theta_{321} = p(x_{31} x_{11}, x_{22}), \theta_{331} = p(x_{31} x_{12}, x_{21}),$ $\theta_{341} = p(x_{31} x_{12}, x_{22}), \theta_{351} = p(x_{31} x_{13}, x_{21}), \theta_{361} = p(x_{31} x_{13}, x_{22}),$ $\theta_{312} = p(x_{32} x_{11}, x_{21}), \theta_{322} = p(x_{32} x_{11}, x_{22}), \theta_{332} = p(x_{32} x_{12}, x_{21}),$ $\theta_{342} = p(x_{32} x_{12}, x_{22}), \theta_{352} = p(x_{32} x_{13}, x_{21}), \theta_{362} = p(x_{32} x_{13}, x_{22})$ |
| X_4 | $\theta_{411} = p(x_{41} x_{21}), \theta_{421} = p(x_{41} x_{22}),$ $\theta_{412} = p(x_{42} x_{21}), \theta_{422} = p(x_{42} x_{22})$ |

| X_3 | | | X_1 | | | X_2 | | | $p(X_3 X_1, X_2)$ | | | X_3 | | | X_1 | | | X_2 | | | $p(X_3 X_1, X_2)$ | | | X_4 | | | X_2 | | | $p(X_4 X_2)$ | | |
|-------|----------|--|-------|---|---|-------|---|---|-------------------|---|---|-------|---|---|-------|---|---|-------|---|---|-------------------|---|---|-------|---|---|-------|---|--------------|--------------|--|--|
| X_1 | $p(X_1)$ | | | a | a | a | b | a | b | a | b | b | b | b | a | b | b | a | c | a | b | c | a | b | a | b | a | b | $p(X_4 X_2)$ | | | |
| a | 0,4 | | | b | a | a | a | b | a | a | b | b | b | b | a | b | b | a | c | a | b | c | a | b | a | b | a | b | 0,25 | | | |
| b | 0,3 | | | a | a | b | a | a | b | a | c | a | a | a | b | c | a | b | c | a | b | c | a | b | a | b | a | b | 0,75 | | | |
| c | 0,3 | | | b | a | b | b | a | b | b | c | a | b | b | c | a | b | b | c | a | b | c | a | b | a | b | a | b | 0,66 | | | |
| | | | | a | b | a | a | b | a | a | c | b | a | a | c | b | a | a | c | b | b | c | b | b | a | b | a | b | 0,33 | | | |
| | | | | b | b | a | b | b | a | b | c | b | b | b | c | b | b | b | c | b | b | c | b | b | a | b | a | b | 0,25 | | | |

Bayesian networks

Number of parameters

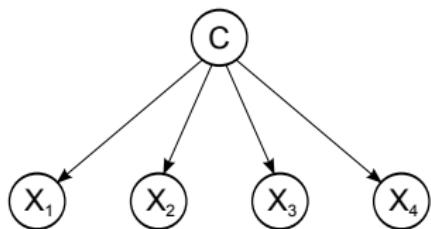
| X_j | Model parameters |
|-------|---|
| X_1 | $\theta_{1\emptyset 1} = p(x_{11}), \theta_{1\emptyset 2} = p(x_{12}), \theta_{1\emptyset 3} = p(x_{13})$ |
| X_2 | $\theta_{2\emptyset 1} = p(x_{21}), \theta_{2\emptyset 2} = p(x_{22})$ |
| X_3 | $\theta_{311} = p(x_{31} x_{11}, x_{21}), \theta_{321} = p(x_{31} x_{11}, x_{22}), \theta_{331} = p(x_{31} x_{12}, x_{21}),$ $\theta_{341} = p(x_{31} x_{12}, x_{22}), \theta_{351} = p(x_{31} x_{13}, x_{21}), \theta_{361} = p(x_{31} x_{13}, x_{22}),$ $\theta_{312} = p(x_{32} x_{11}, x_{21}), \theta_{322} = p(x_{32} x_{11}, x_{22}), \theta_{332} = p(x_{32} x_{12}, x_{21}),$ $\theta_{342} = p(x_{32} x_{12}, x_{22}), \theta_{352} = p(x_{32} x_{13}, x_{21}), \theta_{362} = p(x_{32} x_{13}, x_{22})$ |
| X_4 | $\theta_{411} = p(x_{41} x_{21}), \theta_{421} = p(x_{41} x_{22}),$ $\theta_{412} = p(x_{42} x_{21}), \theta_{422} = p(x_{42} x_{22})$ |

| X_1 | $p(X_1)$ | $X_3 \ X_1 \ X_2$ | $p(X_3 X_1, X_2)$ | $X_3 \ X_1 \ X_2$ | $p(X_3 X_1, X_2)$ | $X_4 \ X_2$ | $p(X_4 X_2)$ |
|-------|-------------------------------|-------------------|-----------------------|-------------------|-----------------------|-------------|-----------------------|
| a | $0,4 = \theta_{1\emptyset 1}$ | a a a | $0,55 = \theta_{311}$ | a b b | $0,30 = \theta_{341}$ | a a | $0,25 = \theta_{411}$ |
| b | $0,3 = \theta_{1\emptyset 2}$ | b a a | $0,45$ | b b b | $0,70$ | b a | $0,75$ |
| c | $0,3$ | a a b | $0,40 = \theta_{321}$ | a c a | $0,80 = \theta_{351}$ | a b | $0,66 = \theta_{421}$ |
| | | b a b | $0,60$ | b c a | $0,20$ | b b | $0,33$ |
| | | a b a | $0,35 = \theta_{331}$ | a c b | $0,25 = \theta_{361}$ | | |
| | | b b a | $0,65$ | b c b | $0,75$ | | |

Naive Bayes

Parameters ($|\Omega_C| = 2$)

$$\theta_C = p(C = 1)$$



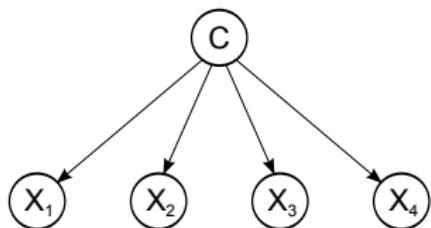
Naive Bayes

Parameters ($|\Omega_C| = 2$)

$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$



Naive Bayes

Parameters ($|\Omega_C| = 2$)

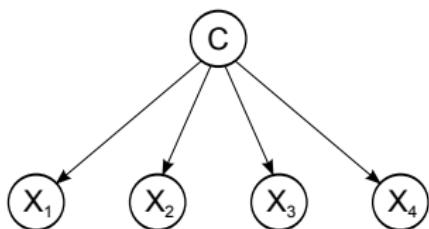
$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

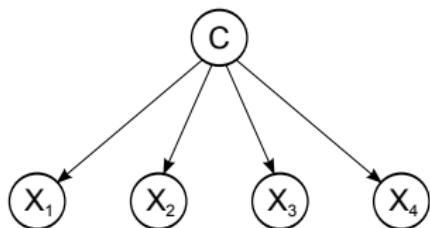
$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$



Naive Bayes

Parameters ($|\Omega_C| = 2$)



$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

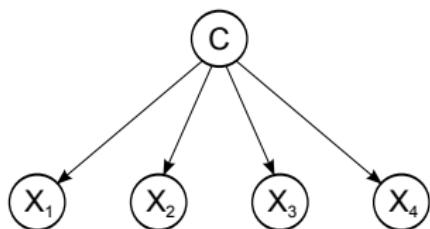
$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$

$$\theta_{X_3|C=0} = p(X_3 = 1|C = 0)$$

$$\theta_{X_3|C=1} = p(X_3 = 1|C = 1)$$

Naive Bayes

Parameters ($|\Omega_C| = 2$)



$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$

$$\theta_{X_3|C=0} = p(X_3 = 1|C = 0)$$

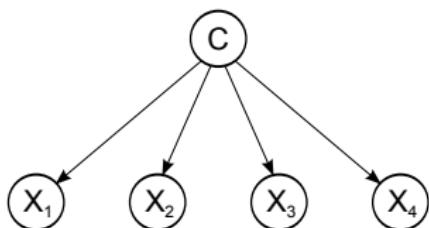
$$\theta_{X_3|C=1} = p(X_3 = 1|C = 1)$$

$$\theta_{X_4|C=0} = p(X_4 = 1|C = 0)$$

$$\theta_{X_4|C=1} = p(X_4 = 1|C = 1)$$

Naive Bayes

Parameters ($|\Omega_C| = 2$)



$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$

$$\theta_{X_3|C=0} = p(X_3 = 1|C = 0)$$

$$\theta_{X_3|C=1} = p(X_3 = 1|C = 1)$$

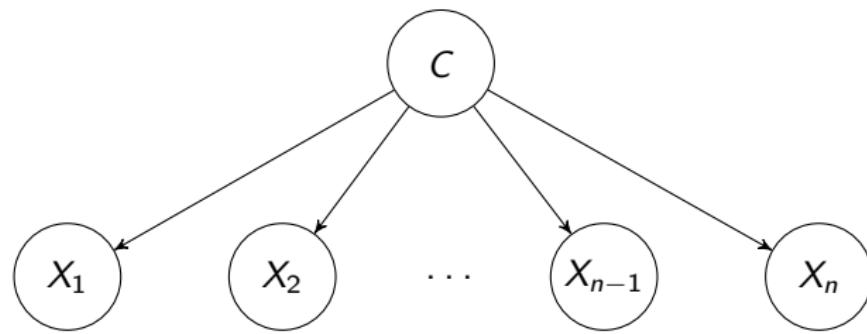
$$\theta_{X_4=1|C=0} = p(X_4 = 1|C = 0)$$

$$\theta_{X_4=2|C=0} = p(X_4 = 2|C = 0)$$

$$\theta_{X_4=1|C=1} = p(X_4 = 1|C = 1)$$

$$\theta_{X_4=2|C=1} = p(X_4 = 2|C = 1)$$

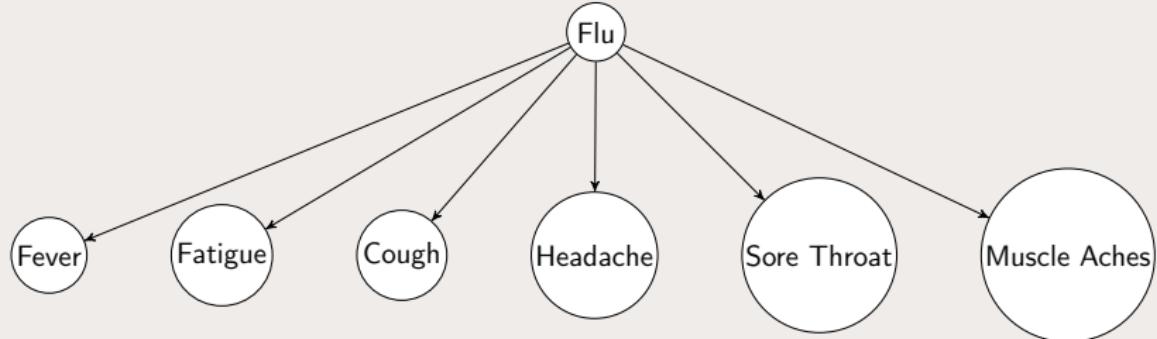
Naive Bayes



Exercise

Naive Bayes model for Flue diagnosis

Which of the following statements are true in this model?



Exercise

Naive Bayes model for Flue diagnosis

Which of the following statements are true in this model?

- a) Say we observe that 1000 people have a headache (and possibly other symptoms), out of which 500 people have the flu (and possibly other symptoms), and 500 people have a fever (and possibly other symptoms). We would expect that approximately 250 people with a headache also have both the flu and a fever.
- b) Say we observe that 500 people have a headache (and possibly other symptoms) and 500 people have a fever (and possibly other symptoms). We would expect that approximately 250 people have both a headache and fever.
- c) Say we observe that 1000 people have the flu, out of which 500 people have a headache (and possibly other symptoms) and 500 have a fever (and possibly other symptoms). We would expect that approximately 250 people with the flu also have both a headache and fever.
- d) Say we observe that 1000 people have the flu, out of which 500 people have a headache (and possibly other symptoms) and 500 have a fever (and possibly other symptoms). We can conclude that exactly 250 people with the flu also have both a headache and fever.

Bayesian networks summary

| | |
|-----------------------|--|
| Model name | Bayesian network |
| Graph type | Directed Acyclic |
| Factorization | factor per variable (variable and its parents) |
| Independence | determined by d-separation in the graph |
| $F \implies I$ | Always |
| $I \implies F$ | Always |
| Markov blanket | Parents, children and parents of children |

Bayesian Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Markov Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Concepts review

What we have already seen:

- ▶ The probabilistic approach to AI
- ▶ Factorization, Conditional independence
- ▶ What is a probabilistic graphical model
- ▶ What is a Bayesian network

Today:

- ▶ Factor algebra
- ▶ Markov networks



Markov networks

Misconception example

1. Organize four students **in pairs** to work on homework
 - 1.1 Alice and Bob are friends
 - 1.2 Bob and Charles study together
 - 1.3 Charles and Debbie work together but usually disagree
 - 1.4 Debbie and Alice study together
 - 1.5 Alice and Charles cannot stand each other
 - 1.6 Bob and Debbie had a relationship that ended badly
2. Some students find out a mistake: the teacher explained something wrong
3. Students transmit this finding to their study partners
4. Probability that all of them are aware of the mistake?

Markov networks

Misconception example

1. Organize four students **in pairs** to work on homework
 - 1.1 Alice and Bob are friends
 - 1.2 Bob and Charles study together
 - 1.3 Charles and Debbie work together but usually disagree
 - 1.4 Debbie and Alice study together
 - 1.5 Alice and Charles cannot stand each other
 - 1.6 Bob and Debbie had a relationship that ended badly
2. Some students find out a mistake: the teacher explained something wrong
3. Students transmit this finding to their study partners
4. Probability that all of them are aware of the mistake?

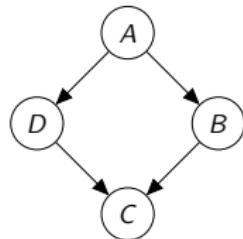
$$A \perp\!\!\!\perp C \mid D, B$$

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

Markov networks

Misconception example

Let us try with Bayesian networks

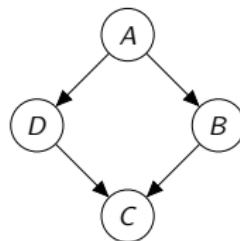


$$B \not\perp\!\!\!\perp D \mid A, C$$

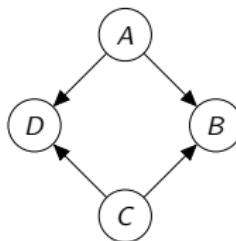
Markov networks

Misconception example

Let us try with Bayesian networks



$$B \not\perp\!\!\!\perp D \mid A, C$$

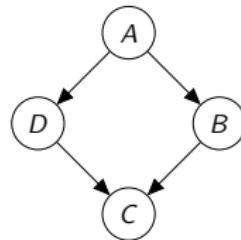


$$A \not\perp\!\!\!\perp C \mid D, B$$

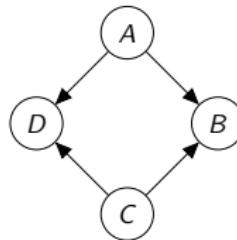
Markov networks

Misconception example

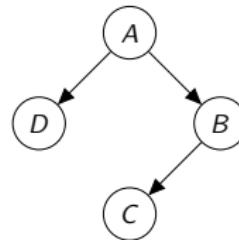
Let us try with Bayesian networks



$$B \not\perp\!\!\!\perp D \mid A, C$$



$$A \not\perp\!\!\!\perp C \mid D, B$$

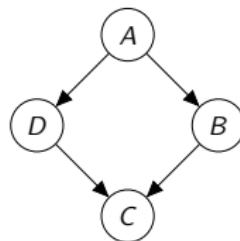


$$C \perp\!\!\!\perp D \mid A, B$$

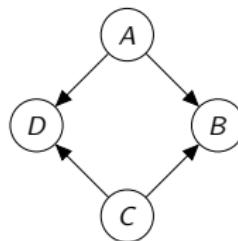
Markov networks

Misconception example

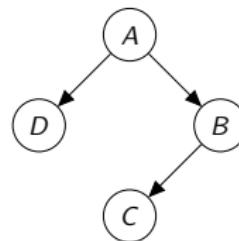
Let us try with Bayesian networks



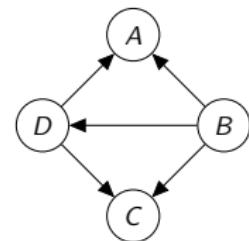
$$B \not\perp\!\!\!\perp D \mid A, C$$



$$A \not\perp\!\!\!\perp C \mid D, B$$



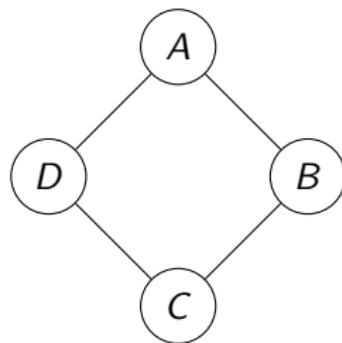
$$C \perp\!\!\!\perp D \mid A, B$$



$$B \not\perp\!\!\!\perp D \mid A, C$$

Markov networks

Misconception example

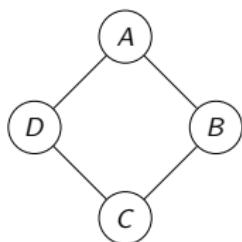


Markov networks vs. Bayesian Networks

| Model | Bayesian network | Markov network |
|-----------------------|---|--|
| Graph | Directed Acyclic | Undirected |
| Factorization | Distribution per variable (variable and its parents) | Factors over the cliques of the graph |
| Joint | | Requires normalization term (partition function) |
| Independence | Determined by d-separation in the graph | Determined by separation in the graph |
| $F \implies I$ | Always | Always |
| $I \implies F$ | Always | Only for positive distribu- tions |
| Markov blanket | Parents, children and pa- rents of children | The neighbors |

Markov networks

Misconception example



$$\phi_1(A, B)$$

$$\begin{array}{ccc} a^0 & b^0 & 30 \\ a^0 & b^1 & 5 \\ a^1 & b^0 & 1 \\ a^1 & b^1 & 10 \end{array}$$

$$\phi_2(B, C)$$

$$\begin{array}{ccc} b^0 & c^0 & 100 \\ b^0 & c^1 & 1 \\ b^1 & c^0 & 1 \\ b^1 & c^1 & 100 \end{array}$$

$$\phi_3(C, D)$$

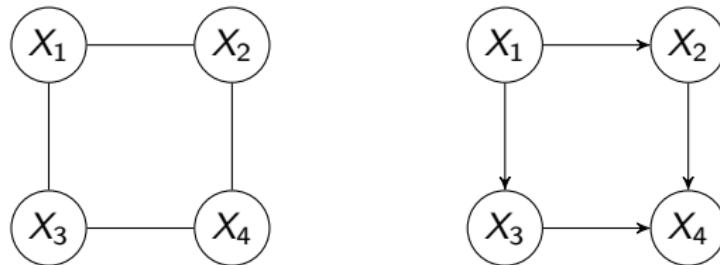
$$\begin{array}{ccc} c^0 & d^0 & 1 \\ c^0 & d^1 & 100 \\ c^1 & d^0 & 100 \\ c^1 & d^1 & 1 \end{array}$$

$$\phi_4(D, A)$$

$$\begin{array}{ccc} d^0 & a^0 & 100 \\ d^0 & a^1 & 1 \\ d^1 & a^0 & 1 \\ d^1 & a^1 & 100 \end{array}$$

Markov networks vs. Bayesian networks

1. Graph



2. Factorization

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^f \phi_i(\mathbf{x}_{\phi_i}) \quad P(\mathbf{x}) = \prod_{i=1}^v p(x_i | \mathbf{pa}_i)$$

There is no partition function

3. Factors

| X_4 | X_2 | $\phi(X_4, X_2)$ |
|-------|-------|------------------|
| a | a | 3,3 |
| b | a | 13,2 |
| a | b | 18,0 |
| b | b | 12,0 |

Factor potentials

| X_4 | X_2 | $p(X_4 X_2)$ |
|-------|-------|----------------|
| a | a | 0,25 |
| b | a | 0,75 |
| a | b | 0,66 |
| b | b | 0,33 |

Conditional prob. distributions

Markov networks vs. Bayesian Networks

Distributions and factors

- ▶ Distributions:
 - ▶ Joint distribution
 - ▶ Common operators:
 - ▶ Conditioning
 - ▶ Marginalization
- ▶ Factors:
 - ▶ How are they different from distributions?
 - ▶ Common operators:
 - ▶ Product
 - ▶ Reduction
 - ▶ Marginalization
 - ▶ Normalization

Markov Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Factor algebra

- ▶ $\mathbf{X} = (X_1, \dots, X_N)$ is a set of variables
- ▶ A factor ϕ is a **function**,

$$\phi : Val(\mathbf{X}_\phi) \rightarrow \mathbb{R},$$

over a subset of variables. This set of variables is called *scope*:

$$Scope(\phi) = \mathbf{X}_\phi \subseteq \mathbf{X}$$

- ▶ Each variable X_i has its own set of possible values, $Val(X_i)$ or Ω_{X_i} .
- ▶ Similarly, for a set of variables \mathbf{X}_ϕ , we define $Val(\mathbf{X}_\phi)$ or $\Omega_{\mathbf{X}_\phi}$ as the **cartesian product** of the sets of possible values of the variables in \mathbf{X}_ϕ .

** Remember: assuming that random variables are discrete

Exercise

Factor Scope

Let $\phi(c, e)$ be a factor in a graphical model, where c is a value of C and e is a value of E . Which is the scope of ϕ ?

- a) C, E
- b) A, C, E
- c) A, B, C, E
- d) C

Factor Algebra

Examples

| X | Y | ϕ |
|-----|-----|--------|
| 0 | 0 | 3 |
| 0 | 1 | 2 |
| 1 | 0 | 4 |
| 1 | 1 | 1 |

| Y | Z | ψ |
|-----|-----|--------|
| 0 | 0 | 5 |
| 0 | 1 | 4 |
| 0 | 2 | 1 |
| 1 | 0 | 2 |
| 1 | 1 | 0 |
| 1 | 2 | 6 |

Operations

- ▶ Product
- ▶ Reduction
- ▶ Marginalization
- ▶ Normalization

Factor algebra

Product

Product of factors

Given two factors ϕ and ψ , their product $\phi \times \psi$ is a new factor whose scope is the union of the scopes of ϕ and ψ ($\Omega_{X_\phi} \cup \Omega_{X_\psi}$) and whose value is the product of ϕ and ψ .

Factor algebra

Product

Product of factors

Given two factors ϕ and ψ , their product $\phi \times \psi$ is a new factor whose scope is the union of the scopes of ϕ and ψ ($\Omega_{X_\phi} \cup \Omega_{X_\psi}$) and whose value is the product of ϕ and ψ .

| X | Y | ϕ |
|---|---|--------|
| 0 | 0 | 3 |
| 0 | 1 | 2 |
| 1 | 0 | 4 |
| 1 | 1 | 1 |

| Y | Z | ψ |
|---|---|--------|
| 0 | 0 | 5 |
| 0 | 1 | 4 |
| 0 | 2 | 1 |
| 1 | 0 | 2 |
| 1 | 1 | 0 |
| 1 | 2 | 6 |

Scope ($\Omega_{X_\phi} \cup \Omega_{X_\psi}$) of
 $\phi \times \psi$?

Factor algebra

Product

Product of factors

Given two factors ϕ and ψ , their product $\phi \times \psi$ is a new factor whose scope is the union of the scopes of ϕ and ψ ($\Omega_{X_\phi} \cup \Omega_{X_\psi}$) and whose value is the product of ϕ and ψ .

| X | Y | ϕ |
|---|---|--------|
| 0 | 0 | 3 |
| 0 | 1 | 2 |
| 1 | 0 | 4 |
| 1 | 1 | 1 |

| Y | Z | ψ |
|---|---|--------|
| 0 | 0 | 5 |
| 0 | 1 | 4 |
| 0 | 2 | 1 |
| 1 | 0 | 2 |
| 1 | 1 | 0 |
| 1 | 2 | 6 |

| X | Y | Z | $\phi \times \psi$ |
|---|---|---|--------------------|
| 0 | 0 | 0 | |
| 0 | 0 | 1 | |
| 0 | 0 | 2 | |
| 0 | 1 | 0 | |
| 0 | 1 | 1 | |
| 0 | 1 | 2 | |
| 1 | 0 | 0 | |
| 1 | 0 | 1 | |
| 1 | 0 | 2 | |
| 1 | 1 | 0 | |
| 1 | 1 | 1 | |
| 1 | 1 | 2 | |

Factor algebra

Product

Product of factors

Given two factors ϕ and ψ , their product $\phi \times \psi$ is a new factor whose scope is the union of the scopes of ϕ and ψ ($\Omega_{X_\phi} \cup \Omega_{X_\psi}$) and whose value is the product of ϕ and ψ .

| X | Y | ϕ |
|---|---|--------|
| 0 | 0 | 3 |
| 0 | 1 | 2 |
| 1 | 0 | 4 |
| 1 | 1 | 1 |

| Y | Z | ψ |
|---|---|--------|
| 0 | 0 | 5 |
| 0 | 1 | 4 |
| 0 | 2 | 1 |
| 1 | 0 | 2 |
| 1 | 1 | 0 |
| 1 | 2 | 6 |

| X | Y | Z | $\phi \times \psi$ |
|---|---|---|--------------------|
| 0 | 0 | 0 | 15 |
| 0 | 0 | 1 | 12 |
| 0 | 0 | 2 | 3 |
| 0 | 1 | 0 | 4 |
| 0 | 1 | 1 | 0 |
| 0 | 1 | 2 | 12 |
| 1 | 0 | 0 | 20 |
| 1 | 0 | 1 | 16 |
| 1 | 0 | 2 | 4 |
| 1 | 1 | 0 | 2 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 2 | 6 |

Factor algebra

Product

Properties of the product

- ▶ The product of factors is **commutative**

$$\phi \times \psi = \psi \times \phi$$

- ▶ The product of factors is **associative**

$$(\phi \times \psi) \times \xi = \psi \times (\phi \times \xi)$$

- ▶ For each scope $\mathbf{U} \subseteq \mathbf{X}$, there is a **neutral element** that assigns $\phi(\mathbf{u}) = 1$

Factor algebra

Reduction

Reduction of a factor

The reduction of a factor ϕ for an assignment of values $\mathbf{U} = \mathbf{u}$ is a new factor $\phi[\mathbf{u}]$ whose scope is $\mathbf{V} = \mathbf{X}_\phi \setminus \mathbf{U}$ and whose value for the assignment $\mathbf{V} = \mathbf{v}$, $\phi[\mathbf{u}](\mathbf{v})$, is the value of ϕ for the joint assignment of \mathbf{u} and \mathbf{v} , $\phi[\mathbf{u}](\mathbf{v}) = \phi(\mathbf{u}, \mathbf{v})$.

Factor algebra

Reduction

Reduction of a factor

The reduction of a factor ϕ for an assignment of values $\mathbf{U} = \mathbf{u}$ is a new factor $\phi[\mathbf{u}]$ whose scope is $\mathbf{V} = \mathbf{X}_\phi \setminus \mathbf{U}$ and whose value for the assignment $\mathbf{V} = \mathbf{v}$, $\phi[\mathbf{u}](\mathbf{v})$, is the value of ϕ for the joint assignment of \mathbf{u} and \mathbf{v} , $\phi[\mathbf{u}](\mathbf{v}) = \phi(\mathbf{u}, \mathbf{v})$.

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

Scope of
 $\phi[X = 0]?$

Scope of
 $\phi[Y = 1, Z = 2]?$

Factor algebra

Reduction

Reduction of a factor

The reduction of a factor ϕ for an assignment of values $\mathbf{U} = \mathbf{u}$ is a new factor $\phi[\mathbf{u}]$ whose scope is $\mathbf{V} = \mathbf{X}_\phi \setminus \mathbf{U}$ and whose value for the assignment $\mathbf{V} = \mathbf{v}$, $\phi[\mathbf{u}](\mathbf{v})$, is the value of ϕ for the joint assignment of \mathbf{u} and \mathbf{v} , $\phi[\mathbf{u}](\mathbf{v}) = \phi(\mathbf{u}, \mathbf{v})$.

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

| Y | Z | $\phi[X = 0]$ |
|---|---|---------------|
| 0 | 0 | |
| 0 | 1 | |
| 0 | 2 | |
| 1 | 0 | |
| 1 | 1 | |
| 1 | 2 | |

| X | $\phi[Y = 1, Z = 2]$ |
|---|----------------------|
| 0 | |
| 1 | |

Factor algebra

Reduction

Reduction of a factor

The reduction of a factor ϕ for an assignment of values $\mathbf{U} = \mathbf{u}$ is a new factor $\phi[\mathbf{u}]$ whose scope is $\mathbf{V} = \mathbf{X}_\phi \setminus \mathbf{U}$ and whose value for the assignment $\mathbf{V} = \mathbf{v}$, $\phi[\mathbf{u}](\mathbf{v})$, is the value of ϕ for the joint assignment of \mathbf{u} and \mathbf{v} , $\phi[\mathbf{u}](\mathbf{v}) = \phi(\mathbf{u}, \mathbf{v})$.

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

| Y | Z | $\phi[X = 0]$ |
|---|---|---------------|
| 0 | 0 | 4 |
| 0 | 1 | 3 |
| 0 | 2 | 5 |
| 1 | 0 | 11 |
| 1 | 1 | 2 |
| 1 | 2 | 1 |

| X | $\phi[Y = 1, Z = 2]$ |
|---|----------------------|
| 0 | 1 |
| 1 | 9 |

Factor algebra

Product and reduction

Relationship between reduction and product of factors

Let ϕ_1 and ϕ_2 be two factors, and $\mathbf{U} = \mathbf{u}$ an assignment of values to variables:

$$(\phi_1 \times \phi_2)[\mathbf{U} = \mathbf{u}] = \phi_1[\mathbf{U} = \mathbf{u}] \times \phi_2[\mathbf{U} = \mathbf{u}]$$

Factor algebra

Product and reduction

Relationship between reduction and product of factors

Let ϕ_1 and ϕ_2 be two factors, and $\mathbf{U} = \mathbf{u}$ an assignment of values to variables:

$$(\phi_1 \times \phi_2)[\mathbf{U} = \mathbf{u}] = \phi_1[\mathbf{U} = \mathbf{u}] \times \phi_2[\mathbf{U} = \mathbf{u}]$$

This rule can help us reducing work!

$$p(E = e) \quad \text{or} \quad p(X \mid E = e)$$

Factor algebra

Marginalization

Marginal

Given a factor ϕ and a set of variables \mathbf{V} to remove, the **marginal** $\sum_{\mathbf{V}} \phi$ is a factor ψ with scope $\mathbf{U} = \mathbf{X}_\phi \setminus \mathbf{V}$, defined by

$$\psi(\mathbf{u}) = \sum_{\mathbf{v}} \phi(\mathbf{u}, \mathbf{v}) \quad \text{**sometimes written as } \phi^{\downarrow \mathbf{U}}$$

Factor algebra

Marginalization

Marginal

Given a factor ϕ and a set of variables V to remove, the **marginal** $\sum_V \phi$ is a factor ψ with scope $U = X_\phi \setminus V$, defined by
 $\psi(u) = \sum_v \phi(u, v)$

**sometimes written as $\phi^{\downarrow U}$

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

Scope of $\sum_Y \phi$?

Scope of $\sum_{Y,Z} \phi$?

Factor algebra

Marginalization

Marginal

Given a factor ϕ and a set of variables V to remove, the **marginal** $\sum_V \phi$ is a factor ψ with scope $U = X_\phi \setminus V$, defined by
 $\psi(u) = \sum_v \phi(u, v)$

**sometimes written as $\phi^{\downarrow U}$

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

| X | Z | $\sum_Y \phi$ |
|---|---|---------------|
| 0 | 0 | |
| 0 | 1 | |
| 0 | 2 | |
| 1 | 0 | |
| 1 | 1 | |
| 1 | 2 | |

| X | $\sum_{Y,Z} \phi$ |
|---|-------------------|
| 0 | |
| 1 | |

Factor algebra

Marginalization

Marginal

Given a factor ϕ and a set of variables V to remove, the **marginal** $\sum_V \phi$ is a factor ψ with scope $U = X_\phi \setminus V$, defined by
 $\psi(u) = \sum_v \phi(u, v)$

**sometimes written as $\phi^{\downarrow U}$

| X | Y | Z | ϕ |
|---|---|---|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

| X | Z | $\sum_Y \phi$ |
|---|---|---------------|
| 0 | 0 | 15 |
| 0 | 1 | 5 |
| 0 | 2 | 6 |
| 1 | 0 | 8 |
| 1 | 1 | 6 |
| 1 | 2 | 21 |

| X | $\sum_{Y,Z} \phi$ |
|---|-------------------|
| 0 | 26 |
| 1 | 35 |

Factor algebra

Product and marginalization

Relationship between marginalization and product of factors

Let ϕ_1 and ϕ_2 be two factors, if $X \notin \text{Scope}(\phi_1)$:

$$\sum_X (\phi_1 \times \phi_2) = \phi_1 \times \sum_X \phi_2$$

Factor algebra

Product and marginalization

Relationship between marginalization and product of factors

Let ϕ_1 and ϕ_2 be two factors, if $X \notin \text{Scope}(\phi_1)$:

$$\sum_X (\phi_1 \times \phi_2) = \phi_1 \times \sum_X \phi_2$$

This rule can help us reducing work!

$$p(X) = \frac{1}{\theta} \sum_{V \setminus X} \phi_1 \times \cdots \times \phi_f$$

Factor algebra

Product and marginalization

This rule can help us reducing work!

The order of summation matters!

Factor algebra

Product and marginalization

This rule can help us reducing work!

The order of summation matters!

Whenever we have to assess a factor algebra expression:

1. For each reduction operation:
 - 1.1 Move it towards the affected factors and do them
2. For each marginalization operation:
 - 2.1 Select a **good ordering** of the variables
 - 2.2 Move it towards the affected factors
 - 2.3 Perform marginalization operations “from the inside out”

Exercise

Reduction and marginalization

| X | Y | Z | ϕ |
|-----|-----|-----|--------|
| 0 | 0 | 0 | 4 |
| 0 | 0 | 1 | 3 |
| 0 | 0 | 2 | 5 |
| 0 | 1 | 0 | 11 |
| 0 | 1 | 1 | 2 |
| 0 | 1 | 2 | 1 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 0 | 2 | 12 |
| 1 | 1 | 0 | 4 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 2 | 9 |

$$p(X|Y=1)$$

$$p(Z|X=0)$$

Exercise

Reduction and marginalization

Let $\{A, B, C, D\}$ be a set of binary variables, where

$$p(A, B, C, D) = \frac{1}{\theta} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)$$

- ▶ $p(A, C|B = b)$
- ▶ $p(A|B = b)$
- ▶ $p(D|C = c)$
- ▶ $p(B|D = d)$

Exercise

Reduction and marginalization

| $\phi_1(A, B)$ | | | $\phi_2(B, C)$ | | | $\phi_3(C, D)$ | | | $\phi_4(D, A)$ | | |
|----------------|-------|----|----------------|-------|-----|----------------|-------|-----|----------------|-------|-----|
| a^0 | b^0 | 30 | b^0 | c^0 | 100 | c^0 | d^0 | 1 | d^0 | a^0 | 100 |
| a^0 | b^1 | 5 | b^0 | c^1 | 1 | c^0 | d^1 | 100 | d^0 | a^1 | 1 |
| a^1 | b^0 | 1 | b^1 | c^0 | 1 | c^1 | d^0 | 100 | d^1 | a^0 | 1 |
| a^1 | b^1 | 10 | b^1 | c^1 | 100 | c^1 | d^1 | 1 | d^1 | a^1 | 100 |

$$p(A, B, C, D) = \frac{1}{\theta} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)$$

$$p(D|C = c^1) =$$

Factor algebra

Normalization

Marginal

Given a factor ϕ , its normalization

$$\text{Norm}(\phi)(\mathbf{x}) = \frac{\phi(\mathbf{x})}{Z_\phi}$$

where $Z_\phi = \sum_{\mathbf{x}} \phi(\mathbf{x})$

Factor algebra

Normalization

Marginal

Given a factor ϕ , its normalization

$$\text{Norm}(\phi)(\mathbf{x}) = \frac{\phi(\mathbf{x})}{Z_\phi}$$

where $Z_\phi = \sum_{\mathbf{x}} \phi(\mathbf{x})$

| Y | Z | ψ | $\text{Norm}(\psi)$ |
|-----|-----|--------|---------------------|
| 0 | 0 | 2 | |
| 0 | 1 | 5 | |
| 0 | 2 | 2 | |
| 1 | 0 | 4 | |
| 1 | 1 | 3 | |
| 1 | 2 | 1 | |

| Z | ϕ | $\text{Norm}(\phi)$ |
|-----|--------|---------------------|
| 0 | 8 | |
| 1 | 2 | |

Factor algebra

Normalization

Marginal

Given a factor ϕ , its normalization

$$\text{Norm}(\phi)(\mathbf{x}) = \frac{\phi(\mathbf{x})}{Z_\phi}$$

where $Z_\phi = \sum_{\mathbf{x}} \phi(\mathbf{x})$

| Y | Z | ψ | $\text{Norm}(\psi)$ |
|-----|-----|--------|---------------------|
| 0 | 0 | 2 | 2/17 |
| 0 | 1 | 5 | 5/17 |
| 0 | 2 | 2 | 2/17 |
| 1 | 0 | 4 | 4/17 |
| 1 | 1 | 3 | 3/17 |
| 1 | 2 | 1 | 1/17 |

| Z | ϕ | $\text{Norm}(\phi)$ |
|-----|--------|---------------------|
| 0 | 8 | 8/10=0.8 |
| 1 | 2 | 2/10=0.2 |

Factor algebra

Summary

- ▶ Product: $\phi_1 \times \phi_2$
- ▶ Reduction: $\phi[\mathbf{E} = \mathbf{e}]$
- ▶ Marginalization: $\sum_X \phi$
- ▶ Normalization: $\text{Norm}(\phi)$
- ▶ Product and reduction interaction:

$$(\phi_1 \times \phi_2)[\mathbf{U} = \mathbf{u}] = \phi_1[\mathbf{U} = \mathbf{u}] \times \phi_2[\mathbf{U} = \mathbf{u}]$$

- ▶ Product and marginalization interaction:

$$\sum_X (\phi_1 \times \phi_2) = \phi_1 \sum_X \phi_2$$

where $X \notin \text{Scope}(\phi_1)$

Exercise

Factors in Markov Network

Let $\phi_1(A, B)$, $\phi_2(B, C)$, and $\phi_3(A, C)$ be the factors of a MN.

What is

$$\sum_{A,B,C} \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(A, C)?$$

- a) Always less than or equal to $\phi_1(a, b) \times \phi_2(b, c) \times \phi_3(a, c)$, where a is a value of A , b is a value of B , and c is a value of C .
- b) Always greater than or equal to $\phi_1(a, b) \times \phi_2(b, c) \times \phi_3(a, c)$, where a is a value of A , b is a value of B , and c is a value of C .
- c) Always greater than or equal to 0
- d) Always greater than or equal to 1
- e) Always equal to the partition function, Z
- f) Always equal to 1

** More than 1 option might be valid

Markov Networks

Probabilistic Graphical Models

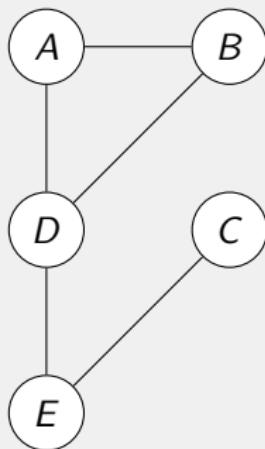
Jerónimo Hernández-González

Markov networks

Cliques: relationship between graph and factorization

Clique or complete subgraph

Subgraph where every two vertices are connected to each other



► 1-vertex cliques:

$$C_1 = \{A\}, C_2 = \{B\}, C_3 = \{C\}, \\ C_4 = \{D\}, C_5 = \{E\}$$

► 2-vertex cliques:

$$C_6 = \{A, B\}, C_7 = \{A, D\}, \\ C_8 = \{B, D\}, C_9 = \{D, E\}, \\ C_{10} = \{E, C\}$$

► 3-vertex cliques:

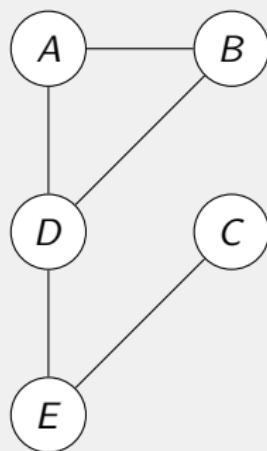
$$C_{11} = \{A, B, D\}$$

Markov networks

Cliques: relationship between graph and factorization

Clique or complete subgraph

Subgraph where every two vertices are connected to each other



► 1-vertex cliques:

$$C_1 = \{A\}, C_2 = \{B\}, C_3 = \{C\}, \\ C_4 = \{D\}, C_5 = \{E\}$$

► 2-vertex cliques:

$$C_6 = \{A, B\}, C_7 = \{A, D\}, \\ C_8 = \{B, D\}, C_9 = \{D, E\}, \\ C_{10} = \{E, C\}$$

► 3-vertex cliques:

$$C_{11} = \{A, B, D\}$$

Maximal clique

There is no any other larger clique that contains it

Markov networks

Factorization

Let \mathcal{H} be an undirected graph. A distribution $P(\mathbf{X})$ factorizes according to \mathcal{H} iff

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^f \phi_i(\mathbf{x}_{\phi_i})$$

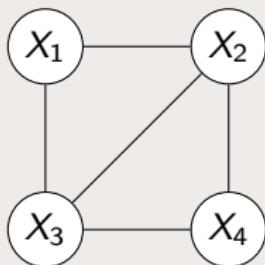
where, for each ϕ_i , there is an edge in \mathcal{H} between each pair of variables in $\text{Scope}(\phi_i)$ (i.e., $\text{Scope}(\phi_i)$ forms a **clique** in \mathcal{H})

and $Z = \sum_{\mathbf{x}} \prod_{i=1}^n \phi_i(\mathbf{x}_{\phi_i})$ is the **normalization constant**, known as **partition function**.

Exercise

Factorization in MNs

Which of these factorizations is valid for the following graph?



$$P(X) \propto \phi_1(X_4)\phi_2(X_1, X_2, X_3)$$

$$P(X) \propto \phi_1(X_1, X_2, X_4)$$

$$P(X) \propto \phi_1(X_1, X_2, X_3)\phi_2(X_2, X_3, X_4)$$

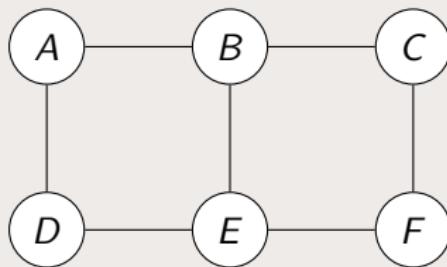
$$P(X) \propto \phi_1(X_1)\phi_2(X_1, X_2)\phi_3(X_3)\phi_4(X_3, X_4)$$

$$P(X) \propto \phi_1(X_1, X_2)\phi_2(X_1, X_3)\phi_3(X_3, X_4)\phi_4(X_2, X_4)\phi_5(X_2, X_3)$$

Exercise

Factorization in MNs

Which of the following sets of factors could factorize over this undirected graph?



- a) $\phi(A), \phi(B), \phi(C), \phi(D), \phi(E), \phi(F)$
- b) $\phi(A, B, D), \phi(A, B), \phi(C, D, E), \phi(E, F), \phi(F)$
- c) $\phi(A, B, D, C), \phi(C, D, E, F)$
- d) $\phi(A, B), \phi(C, D), \phi(E, F)$

Markov Networks

Probabilistic Graphical Models

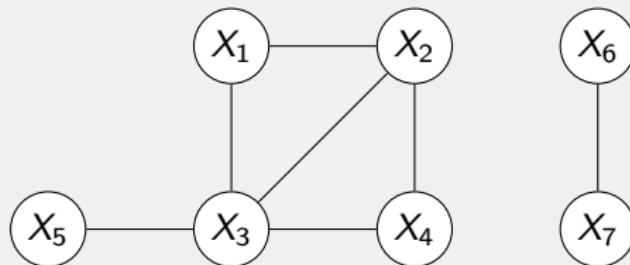
Jerónimo Hernández-González

Markov network

Independencies

Independence

Given a Markov network structure \mathcal{H} , two variables X and Y are independent in \mathcal{H} if there is no path between them in \mathcal{H}

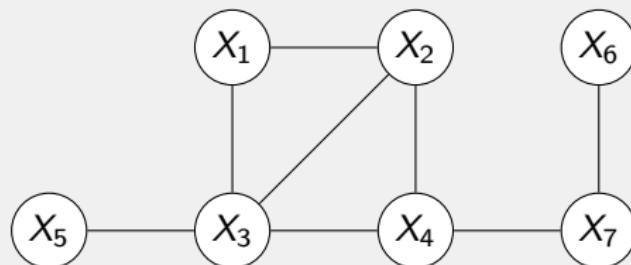


Markov network

Independencies

Conditional independence

Given a Markov network structure \mathcal{H} , two variables X and Y are independent given a third variable Z if Z separates X from Y in \mathcal{H}



Intuitively, probabilistic influence flow is stopped at observed nodes

Flow of probabilistic influence

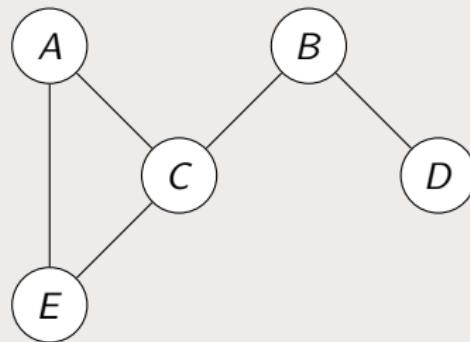
Active trail

- ▶ Let \mathcal{H} be an undirected graph
- ▶ Let $X_1 \rightleftharpoons \dots \rightleftharpoons X_m$ be a **trail** in \mathcal{H}
- ▶ A trail is **active** given a set of observed variables Z if no node along the trail is in Z

Exercise

Independence in MNs

Which pairs of variables are independent in this network?

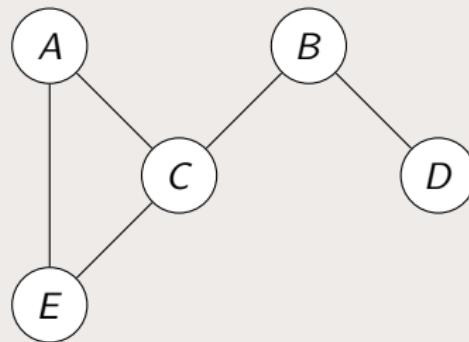


- a) $A \perp\!\!\!\perp C$
- b) $C \perp\!\!\!\perp D$
- c) $D \perp\!\!\!\perp E$
- d) $A \perp\!\!\!\perp D$

Exercise

Independence in MNs

Which conditional independence statements hold in this network?



- a) $A \perp\!\!\!\perp C|E$
- b) $C \perp\!\!\!\perp D|B$
- c) $D \perp\!\!\!\perp E|C$
- d) $A \perp\!\!\!\perp D|E$

Markov network vs. Bayesian network

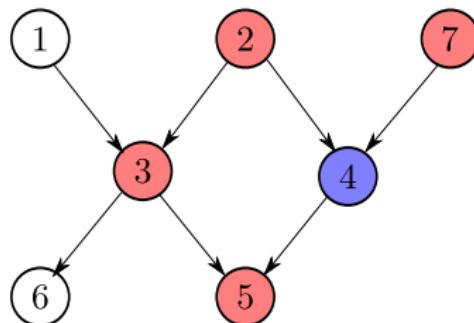
Markov blanket

Definition

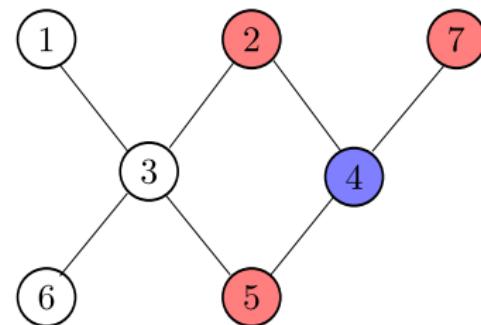
Given X_j and \mathbf{Mb}_j , for any set of variables $\mathbf{X}_c \subseteq (\mathbf{V} \setminus \mathbf{Mb}_j)$:

$$\mathbf{X}_c \perp\!\!\!\perp X_j \mid \mathbf{Mb}_j$$

Parents, children and parents
of the children

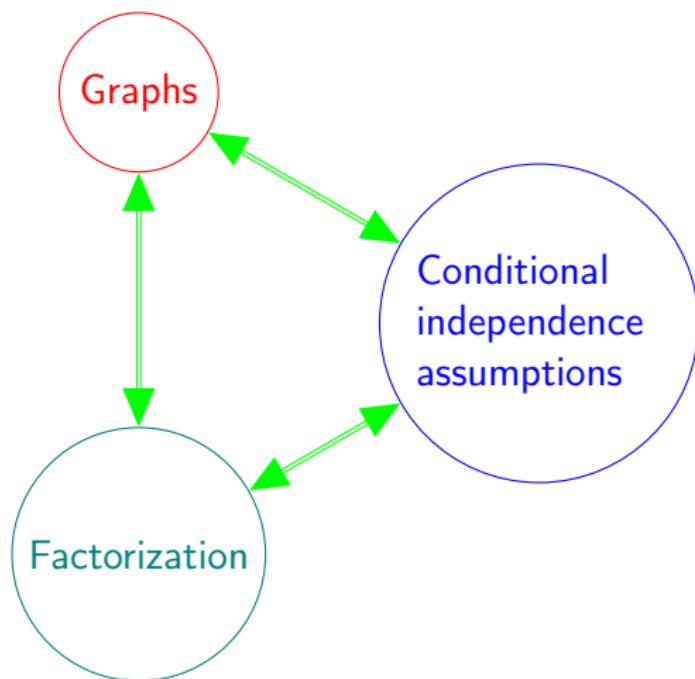


Neighbors



Markov networks

Connection between the three views



Markov networks

Hammersley Clifford Theorem

- ▶ Given a Markov network structure \mathcal{H} , and a probability distribution P that factorizes over \mathcal{H} , then P satisfies the independencies that hold in \mathcal{H} .

P factorizes $\implies P$ satisfies independencies

- ▶ Given a Markov network structure \mathcal{H} , and a *positive* probability distribution P that satisfies the independencies that hold in \mathcal{H} , then P factorizes over \mathcal{H} .

P satisfies independencies
 P positive } $\implies P$ factorizes

Exercise

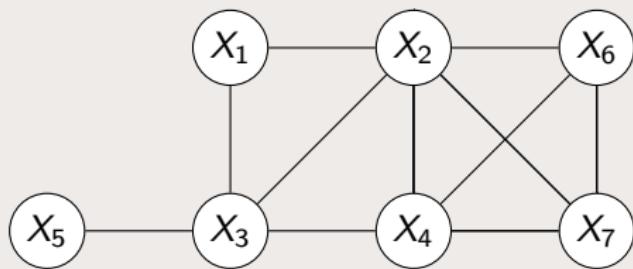
Let P be a distribution such that

$$P(X_1, X_2, X_3, X_4) \propto \phi_1(X_1)\phi_2(X_3, X_4)\phi_3(X_1, X_4)\phi_4(X_4, X_2)$$

- ▶ $X_1 \perp\!\!\!\perp X_3 \mid X_4$?
- ▶ $X_2 \perp\!\!\!\perp X_3 \mid X_1$?
- ▶ $X_2, X_3 \perp\!\!\!\perp X_1$?

Exercise

Given that P is positive and satisfies the independencies in:



Provide a factorization for P .

Markov Networks

Probabilistic Graphical Models

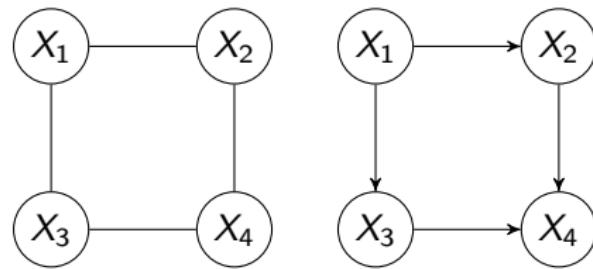
Jerónimo Hernández-González

Markov networks summary

| Model name | Markov network |
|---------------------------|---|
| Graph type | Undirected |
| Factorization | Over the cliques of the graph |
| Probability distr. | Requires a normalization term (partition funct.) |
| Independence | Determined by separation in the graph |
| $F \implies I$ | Always |
| $I \implies F$ | Only for positive distributions |
| Markov blanket | The neighbors |

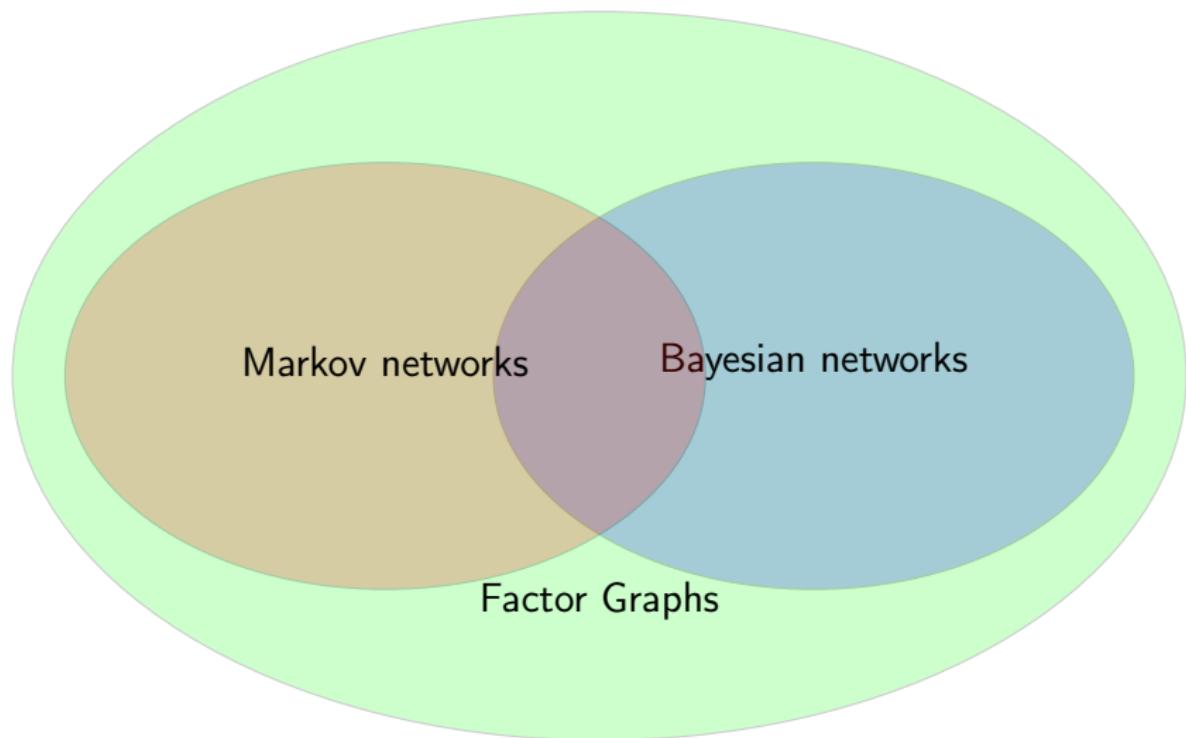
Expressiveness

Models represented by Markov networks and Bayesian networks



Expressiveness

Models represented by Markov networks and Bayesian networks



** The intersection covers the decomposable models: BNs without V-structures _{42 / 45}

Exercise

Storing a Markov network

Let $P(X)$ be a probability distribution that factorizes according to a Markov network \mathcal{H} . How much memory do we need?

Markov Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Markov networks

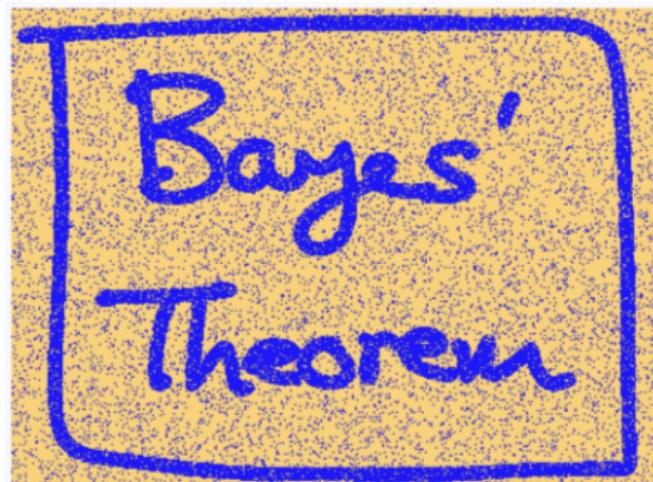
Uses



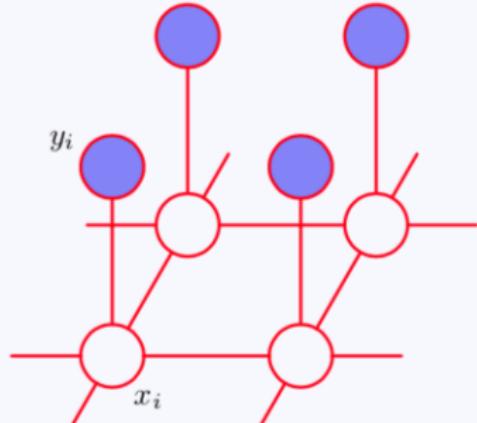
from Christopher Burger, Christian Schuler and Stefan Harmeling. CVPR 2012.

Markov networks

Uses



Noisy image



Markov Network

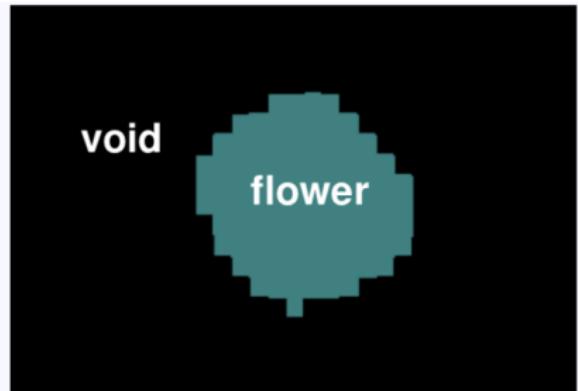
Markov networks

Uses



Markov networks

Uses



Markov networks

Uses



Markov Networks

Probabilistic Graphical Models

Jerónimo Hernández-González

Plate models

Probabilistic Graphical Models

Jerónimo Hernández-González

Plate models

Example: Latent Dirichlet Allocation (LDA) for topic modeling

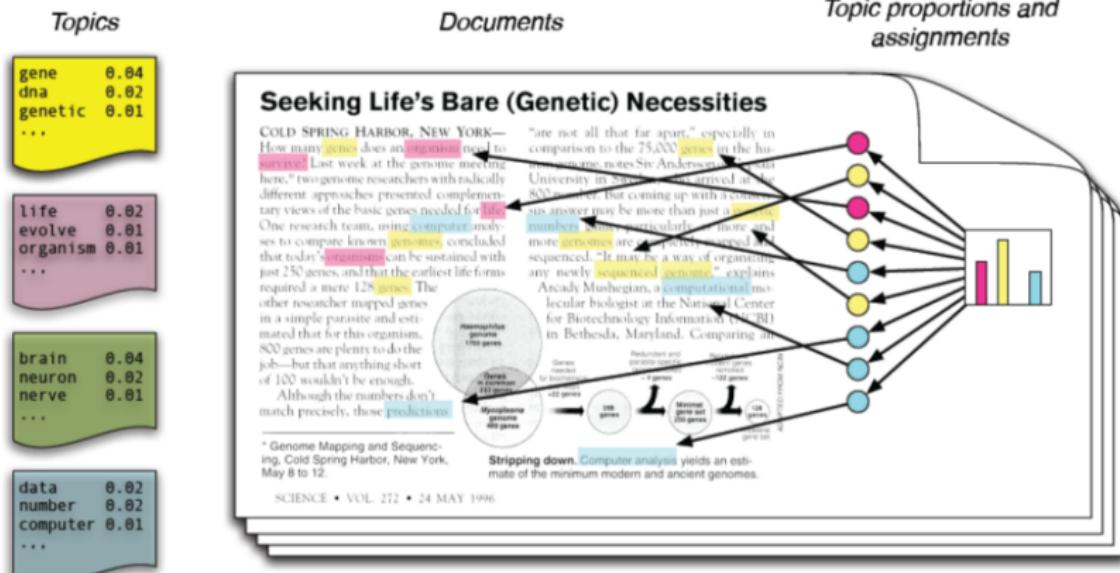
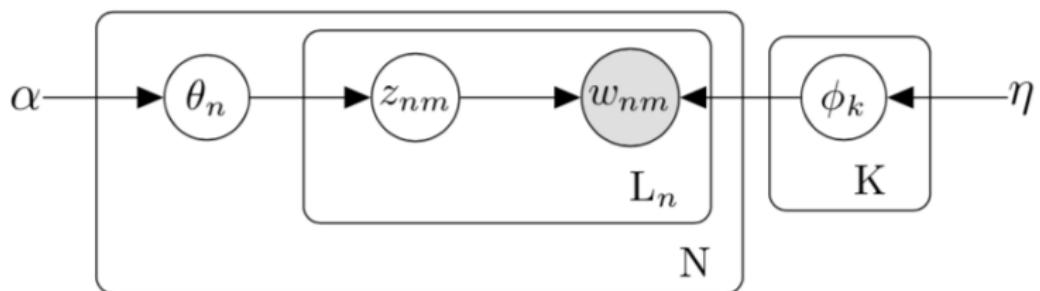


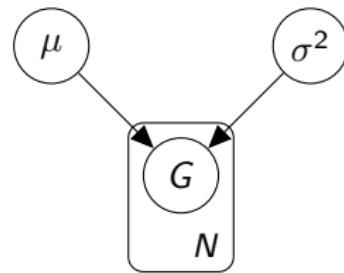
Plate models

Example: Latent Dirichlet Allocation (LDA) for topic modeling



BN with repeated structure

Student-grade

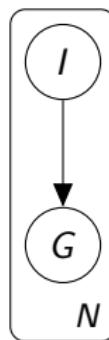


N students;

for each student n , her grade (G_n) is a sample from a Normal distribution $\mathcal{N}(\mu, \sigma^2)$

BN with repeated structure

Student-grade II

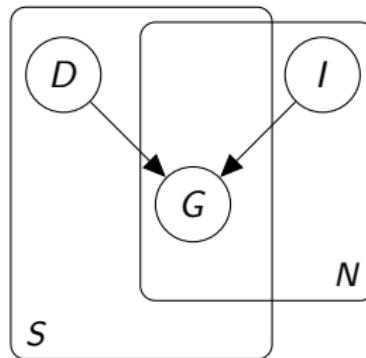


N students;

for each student n , her intelligence (I_n) determines her grade (G_n)

BN with repeated structure

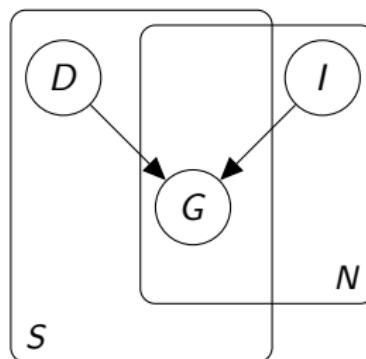
Intersection: Difficulty-Intelligence-Grade



S subjects and N students
for each subject s with a specific difficulty D_s ,
and for each student n with a specific intelligence I_n ,
both the difficulty of the subject (D_s) and her intelligence (I_n)
determine her grade (G_{sn})

BN with repeated structure

Intersection: Difficulty-Intelligence-Grade

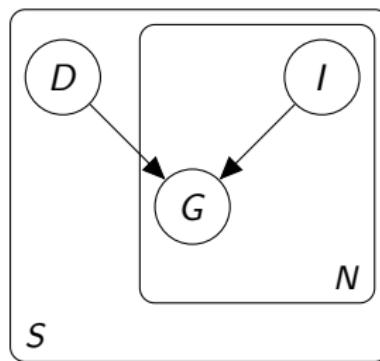


S subjects and N students
for each subject s with a specific difficulty D_s ,
and for each student n with a specific intelligence I_n ,
both the difficulty of the subject (D_s) and her intelligence (I_n)
determine her grade (G_{sn})

General intelligence

BN with Repeated Structure

Nesting: Difficulty-Intelligence-Grade

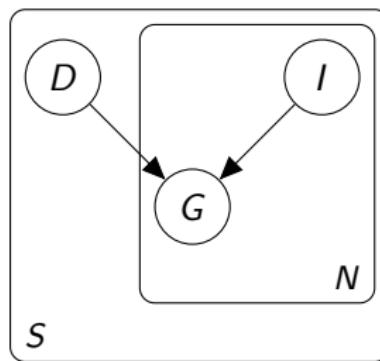


S subjects with N students

for **each subject** s with a specific difficulty D_s , there are N students;
for **each student** n her intelligence (I_{sn}) and the difficulty of the
subject (D_s) determine her grade (G_{sn})

BN with Repeated Structure

Nesting: Difficulty-Intelligence-Grade



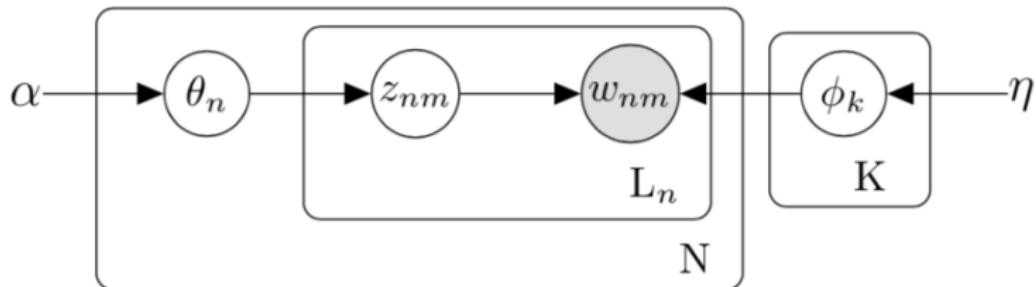
S subjects with N students

for each subject s with a specific difficulty D_s , there are N students;
for each student n her intelligence (I_{sn}) and the difficulty of the
subject (D_s) determine her grade (G_{sn})

Subject-specific intelligence

Plate models

Example: Latent Dirichlet Allocation (LDA) for topic modeling



For each topic $k = 1 \dots K$

$$\phi_k \sim \text{Dir}(\eta)$$

For each document $n = 1 \dots N$

$$\theta_n \sim \text{Dir}(\alpha)$$

For each word $m = 1 \dots L_n$

$$z_{nm} \sim \text{Cat}(\theta_n)$$

$$w_{nm} \sim \text{Cat}(\phi_{z_{nm}})$$

Exercise

Plate Semantics

Let A and B be random variables inside a common plate indexed by i . Which statement is true?

[You may select 1 or more options]

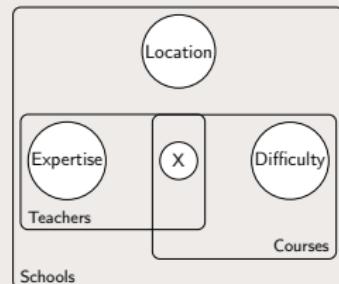
- a) For each i , $A(i)$ and $B(i)$ have different CPDs.
- b) For each i , $A(i)$ and $B(i)$ have edges connecting them to the same variables outside of the plate.
- c) For each i , $A(i)$ and $B(i)$ have the same CPDs.
- d) There is an instance of A and an instance of B for every i .

Exercise

Plate Interpretation

Consider this plate model with edges removed. What might possibly represent an instance of X in the grounded model?

[You may select 1 or more options]

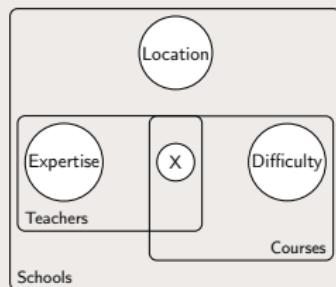


- a) Whether a specific teacher T taught a specific course C at school S
- b) Whether someone with expertise E taught something of difficulty D at a place in location L
- c) Whether a specific teacher T is a tough grader
- d) None
- e) Whether a teacher with expertise E taught a course of difficulty D

Exercise

Grounded Plates

Consider this plate model and assume that there are s schools, t teachers and c courses in each school. How many instances of the Difficulty variable are there?



- a) c
- b) $s \cdot c$
- c) Not enough information to answer
- d) $s \cdot t$

Plate models

Probabilistic Graphical Models

Jerónimo Hernández-González

Plate models

Limitations

- ▶ cannot have edges between two instances of the same variable
(e.g., position at time t depends on position at time $t - 1$)
- ▶ cannot have edges between particular pairs selected by some other relation
(e.g., Genotype(U1) depends on Genotype(U2), where U2 is mother of U1)

Alternatives

- ▶ Dynamic Bayesian Networks (DBNs)
Specific to repetitions over time
- ▶ Probabilistic relational models
More flexibility

Template models

What is a template model?

- ▶ X takes different values at each (discrete) time step
 $X(t)$ is the random variable at time t
- ▶ Markov assumption

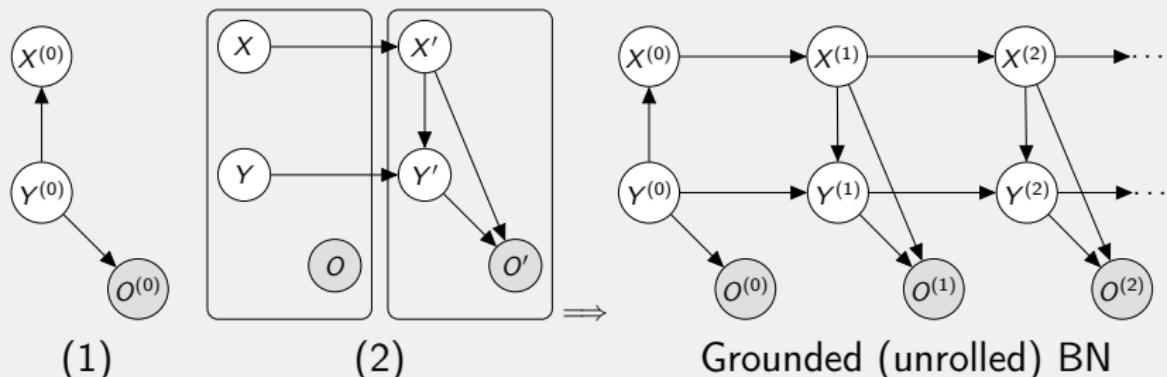
$$\mathbf{X}(t+1) \perp\!\!\!\perp \mathbf{X}(0), \dots, \mathbf{X}(t-1) \mid \mathbf{X}(t)$$

- ▶ Stationary assumption (Time invariance or homogenous)
 $P(\mathbf{X}(t+1) \mid \mathbf{X}(t))$, the same for all t
- ▶ Use **conditional** Bayesian network to define $P(\mathbf{X}(t+1) \mid \mathbf{X}(t))$
2-time slice Bayesian network, Dynamic Bayesian network, Hidden Markov models

Temporal models

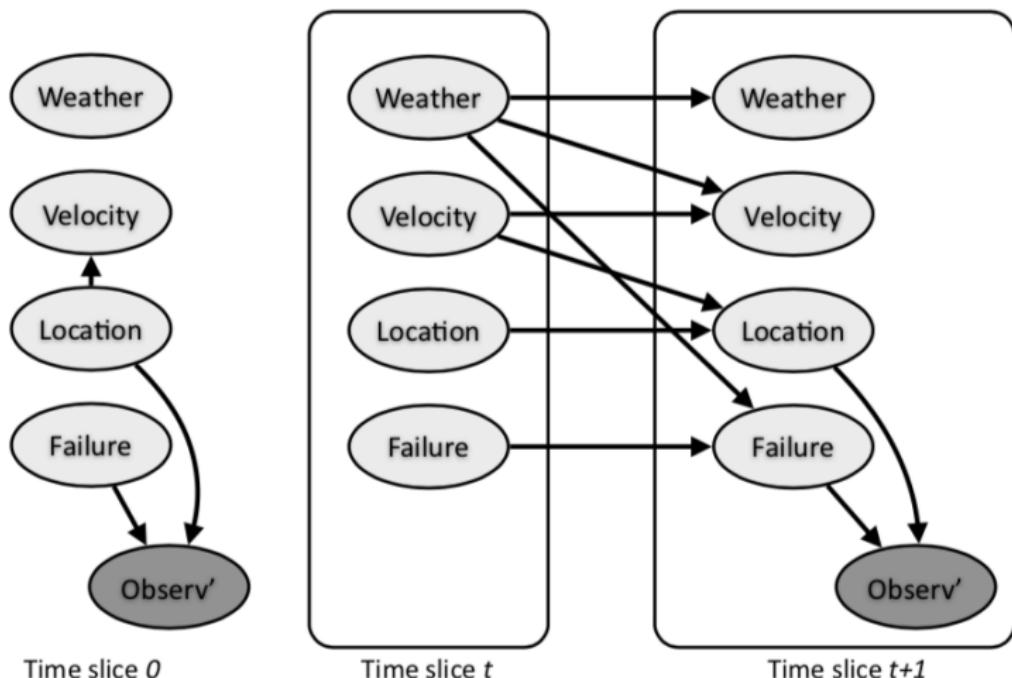
Dynamic Bayesian Network

1. Bayesian network over $\mathbf{X}(0)$
2. Conditional BN for $\mathbf{X}(t + 1)$ given $\mathbf{X}(t)$ (2-time-slice)



Temporal models

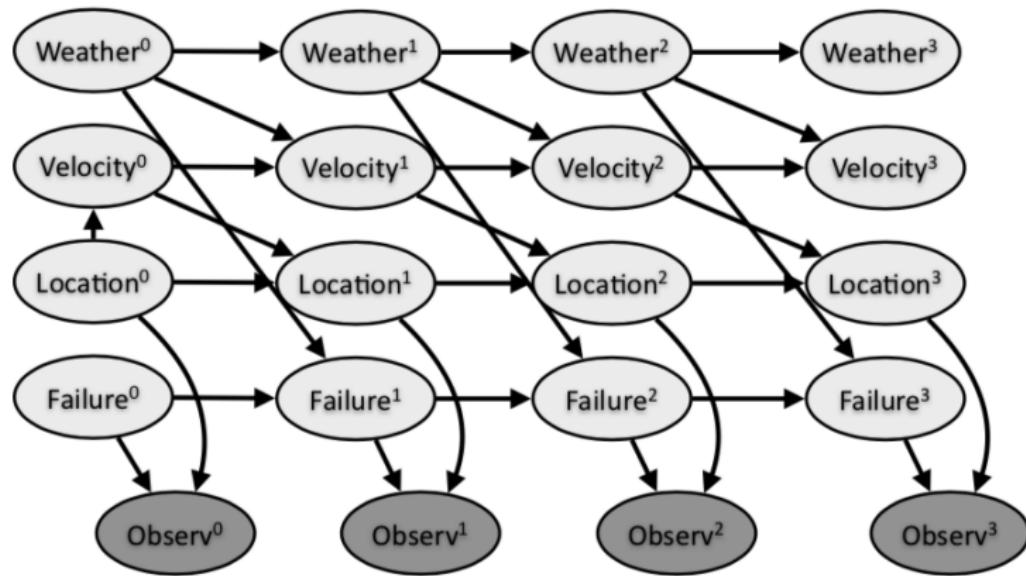
Example: Dynamic Bayesian Network (DBN) for vehicle position



$$P(W', V', L', F', O' | W, V, L, F) = p(O' | F', L') p(F' | F, W) p(L' | L, V) p(V' | V, W) p(W' | W)$$

Temporal models

Example: Dynamic Bayesian Network (DBN) for vehicle position



Grounded (unrolled) BN

Exercise

Markov Assumption

If a dynamic system X satisfies the Markov assumption for all time $t \geq 0$, which statement is true?

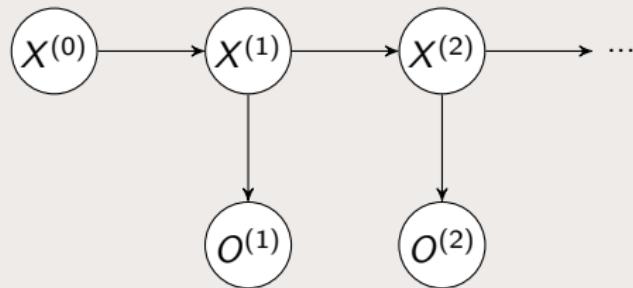
- a) $X^{(t+1)} \perp\!\!\!\perp X^{(t)}$
- b) $X^{(t+1)} \perp\!\!\!\perp X^{(t)} | X^{(t-1)}$
- c) $X^{(t+1)} \perp\!\!\!\perp X^{(0:(t-1))} | X^{(t)}$

Exercise

Independencies in DBNs

In this DBN, which of these independence statements are true?

[You may select 1 or more options]



- a) $O^{(t)} \perp\!\!\!\perp X^{(t+1)} | X^{(t)}$
- b) $O^{(t)} \perp\!\!\!\perp X^{(t-1)} | X^{(t)}$
- c) $O^{(t)} \perp\!\!\!\perp O^{(t-1)}$
- d) $O^{(t)} \perp\!\!\!\perp O^{(t-1)} | X^{(t)}$

Exercise

Applications of DBNs

For which of the following applications might one use a DBN?

- a) Modeling data taken at different locations along a road, where the data at each location is influenced by the data at many other locations.
- b) Predicting the probability that today will be a snow day (school will be closed because of the snow), when this probability depends only on whether yesterday was a snow day.
- c) Predicting the probability that today will be a snow day (school will be closed because of the snow), when this probability depends only on whether yesterday, the day before yesterday, and 2 Mondays ago were snow days.
- d) Modeling time-series data, where the events at each time-point are influenced by only the events at the one time-point directly before it

Plate models

Probabilistic Graphical Models

Jerónimo Hernández-González