

Introduction to Probabilistic Models

Silja Renooij

Department of Information and Computing Sciences
Utrecht University
s.renooij@uu.nl

Trondheim, June 12 2023

Motivation



"I always wondered how it would be if a Superior species landed on Earth and showed us How they played chess. Now I know it."

Peter Heine Nielsen.
Chess Grand Master and Magnus Carlsen's coach

The question is,

- Can we (**humans**) learn (**interpret**) anything from it?

Examples: Disaster Victim Identification



Processing and linking large numbers of DNA profiles of victims and relatives

Examples: Self-driving cars



Can I predict an event in advance so that I can avoid it?

Examples: Image generation

A majestic oil painting of a raccoon Queen wearing red French royal gown. The painting is hanging on an ornate wall decorated with wallpaper.



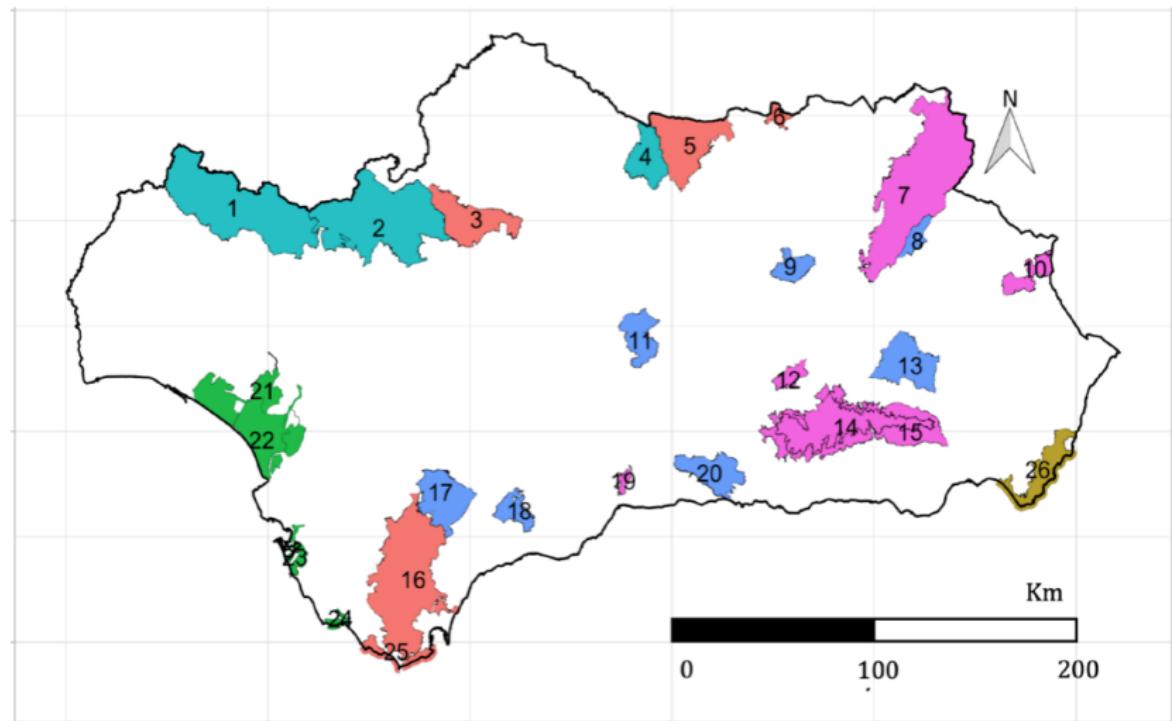
Imagen

unprecedented photorealism × deep level of language understanding

Google Research, Brain Team

Generating images from written descriptions

Examples: Land use



Monitoring protected areas

Probabilistic models

All the previous examples:

- Operate in environments where large amounts of data are available
- However, data don't cover all the possible scenarios ⇒ **UNCERTAINTY**
- Use a **probabilistic model**, typically learnt from data
- Use inference algorithms to carry out **prediction** and **structure analysis**

Probabilistic models offer:

- Principled quantification of uncertainty
- Natural way of dealing with missing data
- Interpretability

Uncertainty

You have probably heard about two types of uncertainty:

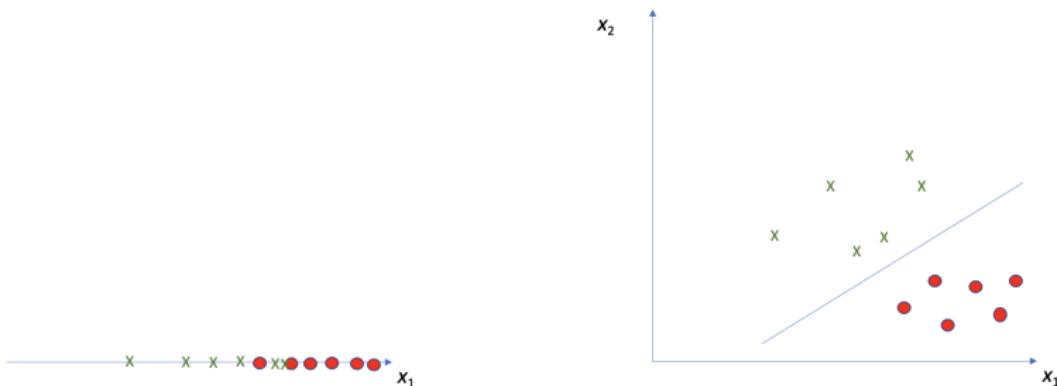
- **Aleatoric:** Due to (pure) randomness, i.e. the variability in the outcome of an experiment due to random effects
- **Epistemic:** Due to lack of knowledge

Example

- Assume we want to predict Y from X
- We estimate a joint distribution $p(x, y)$ [EPISTEMIC][REDUCIBLE]
- We predict Y using $p(y|x) = p(x, y)/p(x)$
- We observe $X = x$, what is our model prediction for Y ?
[ALEATORIC][IRREDUCIBLE]

Uncertainty

- Epistemic uncertainty can be reduced gathering more data, but also increasing the number of features.



Probabilistic graphical models

What we need from probabilistic models:

- Ability to operate in **high dimensional** spaces
- Support **efficient** inference and learning

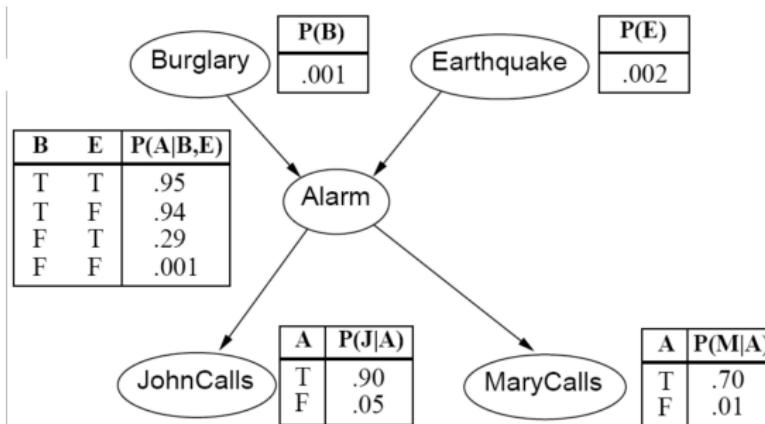
Probabilistic graphical models offer:

- **Structured** specification of high dimensional distributions in terms of low dimensional factors
- **Efficient** inference and learning taking advantage of the structure
- **Graphical** representation interpretable by humans

Bayesian network: definition

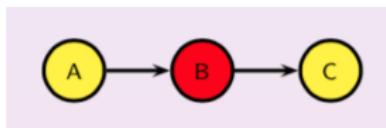
A Bayesian network over random variables X_1, \dots, X_n consists of

- A DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{X_1, \dots, X_n\}$
- A set of local conditional distributions $\mathcal{P} = \{p(X_i|pa(X_i)), X_i \in \mathcal{V}\}$ where $pa(X_i)$ denotes the parents of X_i according to \mathcal{E}

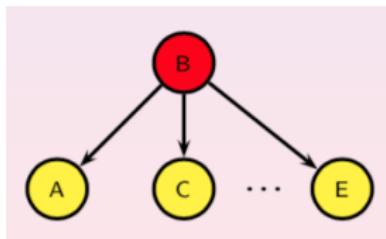


Interpreting Bayesian network structures: d -separation

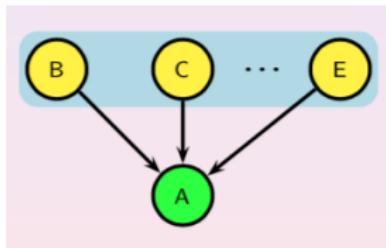
- Serial connection



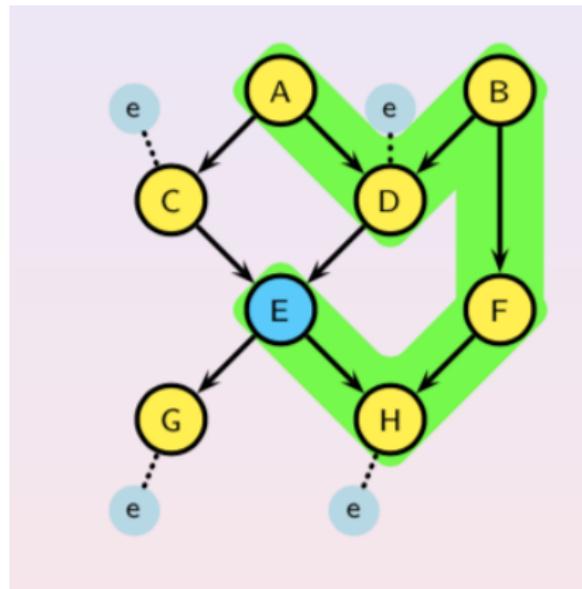
- Diverging connection



- Converging connection

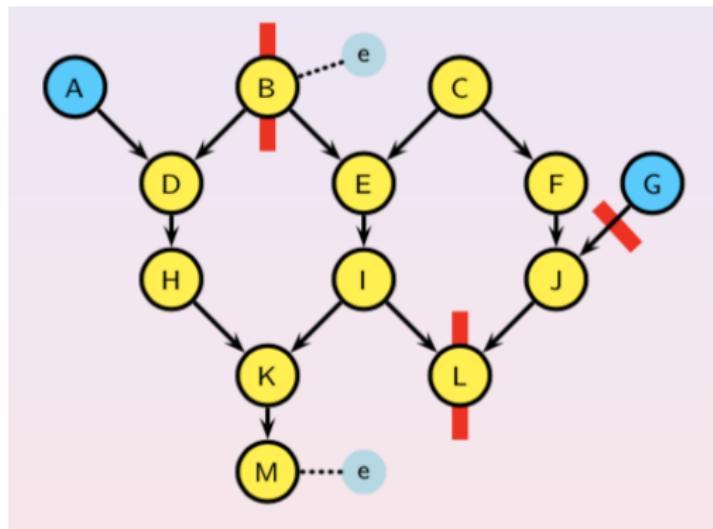


d -separation example



Which variables are **not** d -separated from E given the evidence (e)?

d -separation example

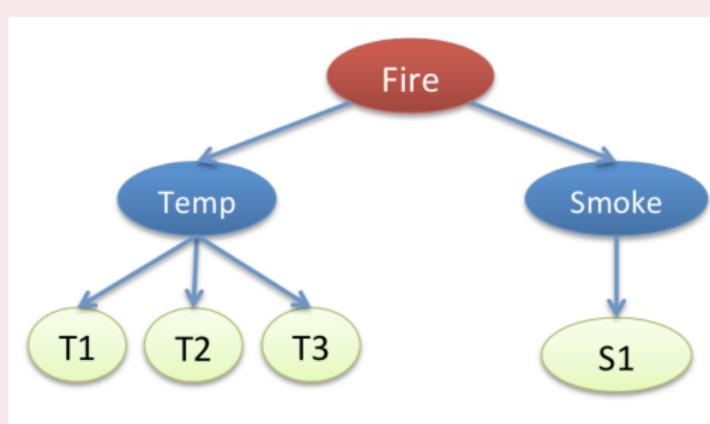


Are variables A and G d -separated given the evidence (e)?

Bayesian networks: compact representation of the joint

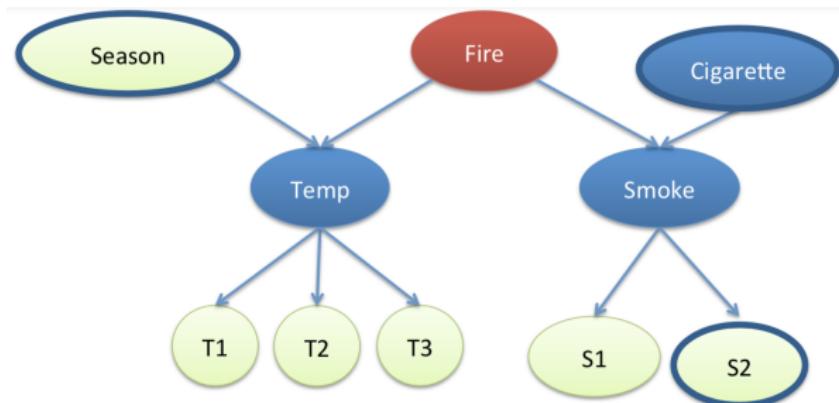
d-separation is used to capture **independences** among the variables;
as a result, every Bayesian network encodes a joint distribution **factorized** as

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | pa(X_i))$$



$$p(f, t, s, t_1, t_2, t_3, s_1) = p(t_1|t)p(t_2|t)p(t_3|t)p(s_1|s)p(t|f)p(s|f)p(f)$$

Bayesian networks: modular structure

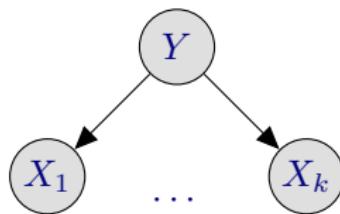


$$p(\text{se}, f, c, t, s, t_1, t_2, t_3, s_1) = p(t_1|t)p(t_2|t)p(t_3|t)p(s_1|s)p(s_2|s) \\ p(t|se, f)p(s|f, c)p(se)p(f)p(c)$$

A simple example

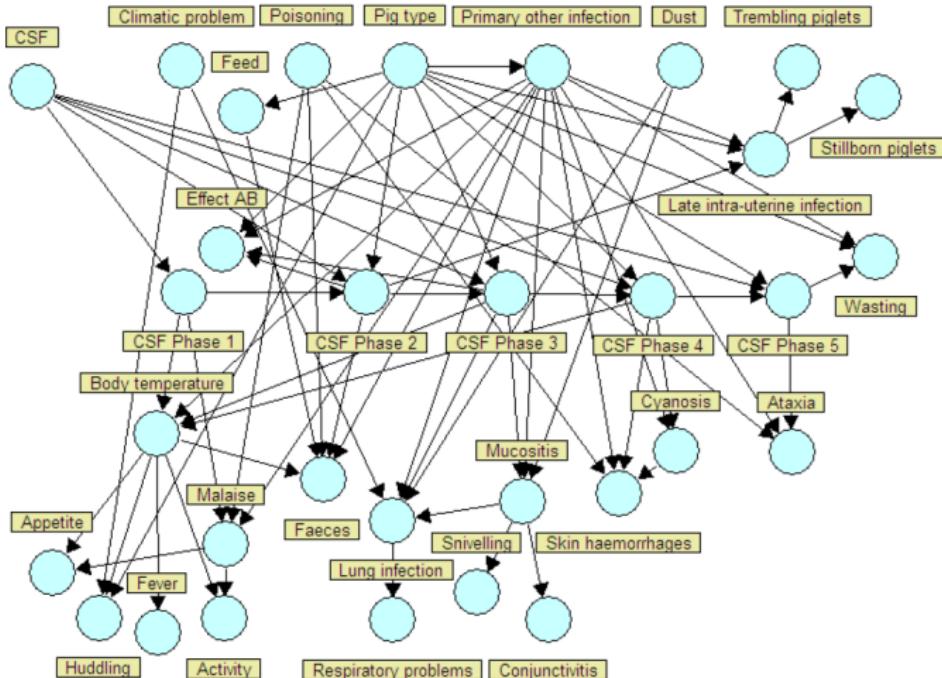
Naive Bayes

- Predicting the value of categorical variable Y from a set of features X_1, \dots, X_k



$$\begin{aligned} p(y|x_1, \dots, x_k) &\stackrel{\text{Bayes}}{=} \frac{p(x_1, \dots, x_k|y)p(y)}{p(x_1, \dots, x_k)} \\ &\propto p(x_1, \dots, x_k|y)p(y) = p(y) \prod_{i=1}^k p(x_i|y) \end{aligned}$$

A more complex example



L.C. van der Gaag, J. Bolt, W. Loeffen, A. Elbers (2010). "Modelling patterns of evidence in Bayesian networks: a case study in Classical Swine Fever"

IPMU 2010, LNCS, vol. 6178, Springer.

DIY: Monty Hall problem

You are given the choice between 3 doors. One has a real prize behind it, the other two joke prizes.



You choose a door; the host opens another one and offers you the choice to switch.

- C : your choice of door
- H : door openend by host Monty
- P : door with real prize

C

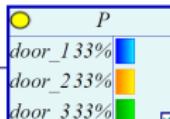
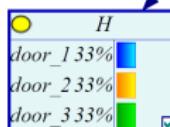
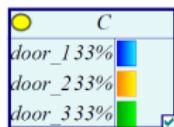
H

P

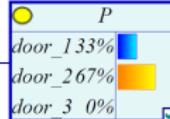
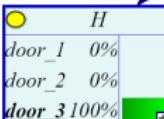
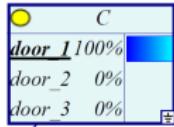
Inference in Bayesian networks

From the **joint distribution** $p(X_1, \dots, X_n)$ we can infer a.o.

- the **prior distribution** of any $X_i \in \mathcal{V}$, $p(X_i)$
- the **posterior distribution** of any $X_i \in \mathcal{V}$ given evidence for $\mathbf{x}_E \subseteq \mathcal{V} \setminus \{X_i\}$,
 $p(X_i | \mathbf{x}_E)$



$$p(H) = \sum_{c,p} p(H|c,p)p(c)p(p)$$



$$p(P|C = \text{door}_1, H = \text{door}_3) =$$

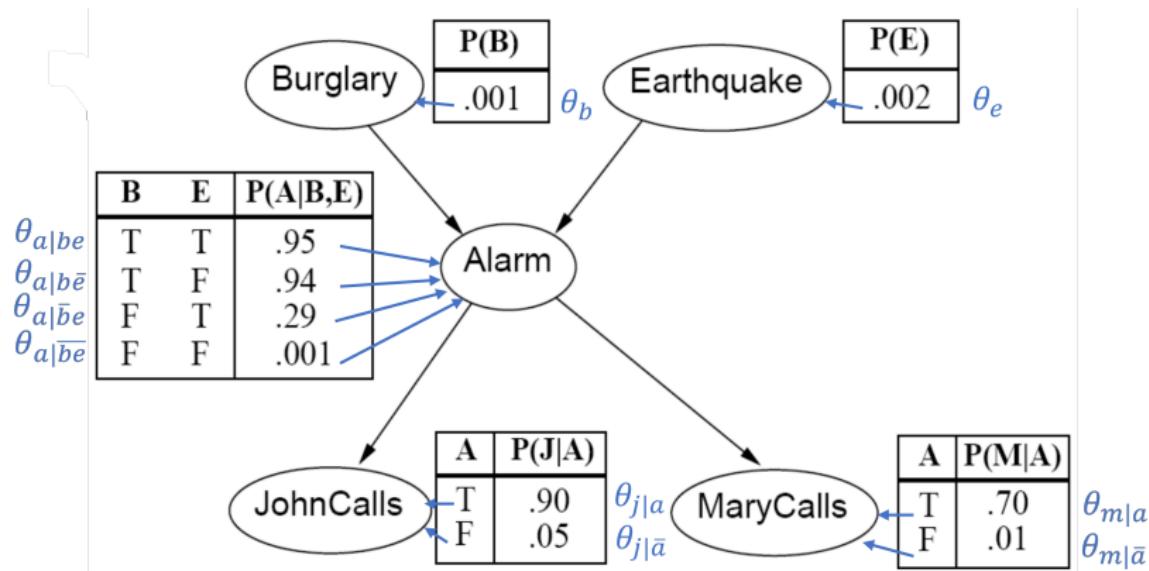
$$\frac{p(H = \text{door}_3|C = \text{door}_1, P)p(P)}{p(H = \text{door}_3)}$$

Probabilistic graphical models

Recall that probabilistic graphical models offer:

- Structured specification of high dimensional distributions in terms of low dimensional factors
- **Efficient inference and learning taking advantage of the structure**
- Graphical representation interpretable by humans

Model parameters



Probabilistic Machine Learning

What is machine learning?

"We say that a computer program P learns from experience E with respect to some class of tasks T and a performance measure R , if its performance on the tasks in T , measured in terms of R , improves with experience E ".

(Tom Mitchell, 1997)

Easy example: linear regression

- We have data about two variables X and Y
- We want to predict the value of Y from the value of X
- To solve this task, we decide to use a **linear regression model**

$$\hat{y} = a + bx$$

- As *performance measure*, we use

$$\text{rmse}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

with $\mathbf{y} = \{y_1, \dots, y_n\}$ denoting the data and $\hat{\mathbf{y}} = \{\hat{y}_1, \dots, \hat{y}_n\}$ denoting the model estimates

Is this really ML???

Easy example: linear regression

- The linear regression model we have used **IS NOT** a probabilistic model
- We'll see later how it can be approached from a probabilistic point of view

Learning probabilistic models from data

Model:

- a random variable X
- a probability density/mass function f associated with X
- a parameter θ of theoretical distribution f

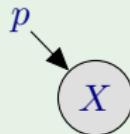
Learning:

- assume f is known except for parameter θ , and denote this fact as $f(x; \theta)$ or $f(x|\theta)$
- for a sample X_1, \dots, X_n drawn from $f(x|\theta)$, regard the joint density/mass $f(x_1, \dots, x_n|\theta)$ as a function of parameter θ , called the likelihood function:

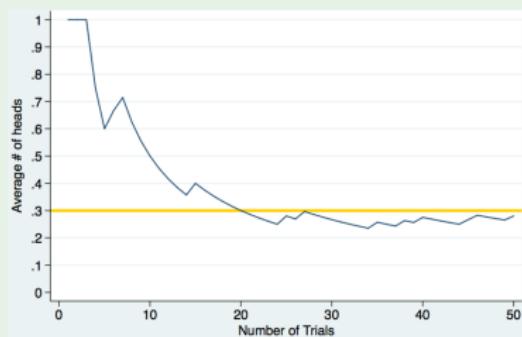
$$l(\theta|x_1, \dots, x_n) \stackrel{\text{def}}{=} f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Learning parameters from data: frequentist approach

- Parameter θ has a **fixed but unknown** value



p is a probability:
relative frequency
“in the long run”



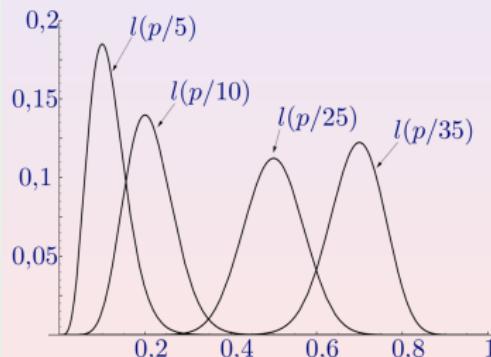
$HHHTTH \dots \dots TH$

$$\hat{p} = \frac{15}{50}$$

Assume a sample $X \sim \mathcal{B}(50, p)$

The likelihood function is

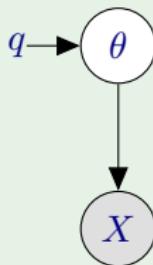
$$l(p|x) = \binom{50}{x} p^x (1-p)^{50-x}$$



Learning parameters from data: Bayesian approach

- Parameters are modelled as random variables and information about them can be included prior to observing data
- By Bayes' rule, the prior information is combined with the likelihood, yielding a posterior distribution
- As such, inferences about the parameter allow for its updating

Bayesian networks for Bayesian learning



- Random variables (and parameters) inside circles
- Grey if observable; white if hidden
- Fixed quantities without circle

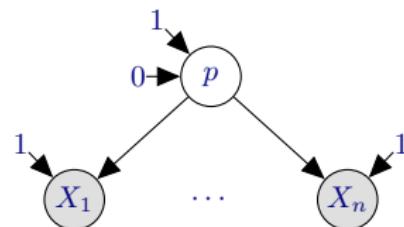
Learning from data: Bayesian approach

Distributions in a Bayesian model

- The prior distribution of θ , $\boxed{\pi(\theta)}$
- The joint distribution of (X, θ) , $\boxed{\psi(x, \theta) = f(x|\theta)\pi(\theta)}$
- The posterior distribution of θ given x , $\boxed{\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\theta} f(x|\theta)\pi(\theta) d\theta}}$
- The prior predictive distribution of X , $\boxed{m(x) = \int_{\theta} f(x|\theta)\pi(\theta) d\theta}$
- The (posterior) predictive distribution given $\mathbf{x} = \{x_1, \dots, x_n\}$:
$$\boxed{f(x_{n+1}|\mathbf{x}) = \int_{\theta} f(x_{n+1}|\theta, \mathbf{x})\pi(\theta|\mathbf{x})d\theta = \int_{\theta} f(x_{n+1}|\theta)\pi(\theta|\mathbf{x})d\theta}$$

Learning from data: Example of Bayesian approach

- Assume a sample $X_1, X_2, \dots, X_n \sim \mathcal{B}(1, p)$ and $p \sim \mathcal{U}(0, 1)$



- Then the likelihood and the prior are,

$$f(x_1, \dots, x_n | p) = p^{\sum x_i} (1-p)^{n - \sum x_i}, \quad \text{with } x_i = 0, 1; \quad p \in (0, 1),$$

$$\pi(p) = \frac{1}{1-0} = 1, \quad \text{if } p \in (0, 1)$$

Learning from data: Example of Bayesian approach

Assume a sample $X_1, X_2, \dots, X_n \sim \mathcal{B}(1, p)$ and $p \sim \mathcal{U}(0, 1)$

- Then the likelihood and the prior are,

$$f(x_1, \dots, x_n | p) = p^{\sum x_i} (1-p)^{n - \sum x_i}, \quad \text{with } x_i = 0, 1; \quad p \in (0, 1),$$
$$\pi(p) = 1, \quad \text{if } p \in (0, 1)$$

- The posterior distribution is

$$\pi(p|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | p) \pi(p)}{\int_0^1 f(x_1, \dots, x_n | p) \pi(p) dp} = \frac{p^{\sum x_i} (1-p)^{n - \sum x_i}}{\int_0^1 p^{\sum x_i} (1-p)^{n - \sum x_i} dp}$$

Learning from data: Example of Bayesian approach

Pattern matching: the Beta distribution $Be(\alpha, \beta)$

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$\begin{aligned} & \int_0^1 p^{\sum x_i} (1-p)^{n-\sum x_i} dp = \\ &= \int_0^1 \frac{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)}{\Gamma(n+2)} \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} p^{\sum x_i} (1-p)^{n-\sum x_i} dp \\ &= \frac{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)}{\Gamma(n+2)} \int_0^1 \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} p^{\sum x_i} (1-p)^{n-\sum x_i} dp \\ &= \frac{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)}{\Gamma(n+2)} \cdot 1 \end{aligned}$$

Learning from data: Example of Bayesian approach

Assume a sample $X_1, X_2, \dots, X_n \sim \mathcal{B}(1, p)$ and $p \sim \mathcal{U}(0, 1) = Be(1, 1)$

- Then the likelihood and the prior are,

$$f(x_1, \dots, x_n | p) = p^{\sum x_i} (1-p)^{n-\sum x_i}, \quad \text{with } x_i = 0, 1; \quad p \in (0, 1),$$
$$\pi(p) = 1, \quad \text{if } p \in (0, 1)$$

- The posterior distribution is

$$\begin{aligned}\pi(p|x_1, \dots, x_n) &= \frac{f(x_1, \dots, x_n | p)\pi(p)}{\int_0^1 f(x_1, \dots, x_n | p)\pi(p) dp} = \frac{p^{\sum x_i} (1-p)^{n-\sum x_i}}{\int_0^1 p^{\sum x_i} (1-p)^{n-\sum x_i} dp} \\ &= \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} p^{\sum x_i} (1-p)^{n-\sum x_i}\end{aligned}$$

which corresponds to $\boxed{Be\left(\sum x_i + 1, n - \sum x_i + 1\right)}$

Very easy to compute for some models

Conjugate priors and likelihoods

Likelihood	Prior	Posterior
$\mathcal{B}(1, \theta)$	$Be(\alpha, \beta)$	$Be(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$
$\mathcal{NB}(r, \theta)$	$Be(\alpha, \beta)$	$Be(\alpha + rn, \beta - nr + \sum_{i=1}^n x_i)$
$\mathcal{G}(\theta)$	$Be(\alpha, \beta)$	$Be(\alpha + n, \beta + \sum_{i=1}^n x_i)$
$\mathcal{MN}(n, \theta_1, \dots, \theta_k)$	$Dir(\alpha_1, \dots, \alpha_k)$	$Dir(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
$P(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + \sum_{i=1}^n x_i, \beta + n)$
$Exp(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + n, \beta + \sum_{i=1}^n x_i)$
$\mathcal{N}(\mu, \tau)$	$\mathcal{N}(\mu_0, \tau_0)$	$\mathcal{N}(\frac{\tau_0\mu_0 + n\tau\bar{x}}{\tau_0 + n\tau}, \tau_0 + n\tau)$
$\mathcal{N}(\underline{\mu}, \tau)$	$\Gamma(\alpha_0, \beta_0)$	$\Gamma(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2)$

Learning from data: Example Bayesian approach

- The prior predictive distribution is

$$m(x) = \int_0^1 p^x (1-p)^{1-x} dp = \frac{\Gamma(x+1)\Gamma(2-x)}{\Gamma(3)} = \frac{x!(1-x)!}{2} = \boxed{\frac{1}{2}} \text{ with } x = 0, 1$$

- The (posterior) predictive distribution is

$$\begin{aligned} f(x|x_1, \dots, x_n) &= \\ &= \int_0^1 p^x (1-p)^{1-x} \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} p^{\sum x_i} (1-p)^{n - \sum x_i} dp \\ &= \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} \int_0^1 p^{x + \sum x_i} (1-p)^{n+1-(x+\sum x_i)} dp \\ &= \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} \frac{\Gamma(x+1 + \sum x_i)\Gamma(n+2 - (x + \sum x_i))}{\Gamma(n+3)} \end{aligned}$$

Learning from data: Bayesian approach

- The method above is known as *fully Bayesian* approach
- Sometimes, we don't need to compute the denominator of the posterior distribution, in which case θ can be estimated as

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} f(x_1, \dots, x_n, \theta) \\ &= \arg \max_{\theta} f(x_1, \dots, x_n | \theta) \pi(\theta) \\ &= \arg \max_{\theta} \{\log f(x_1, \dots, x_n | \theta) + \log \pi(\theta)\}\end{aligned}$$

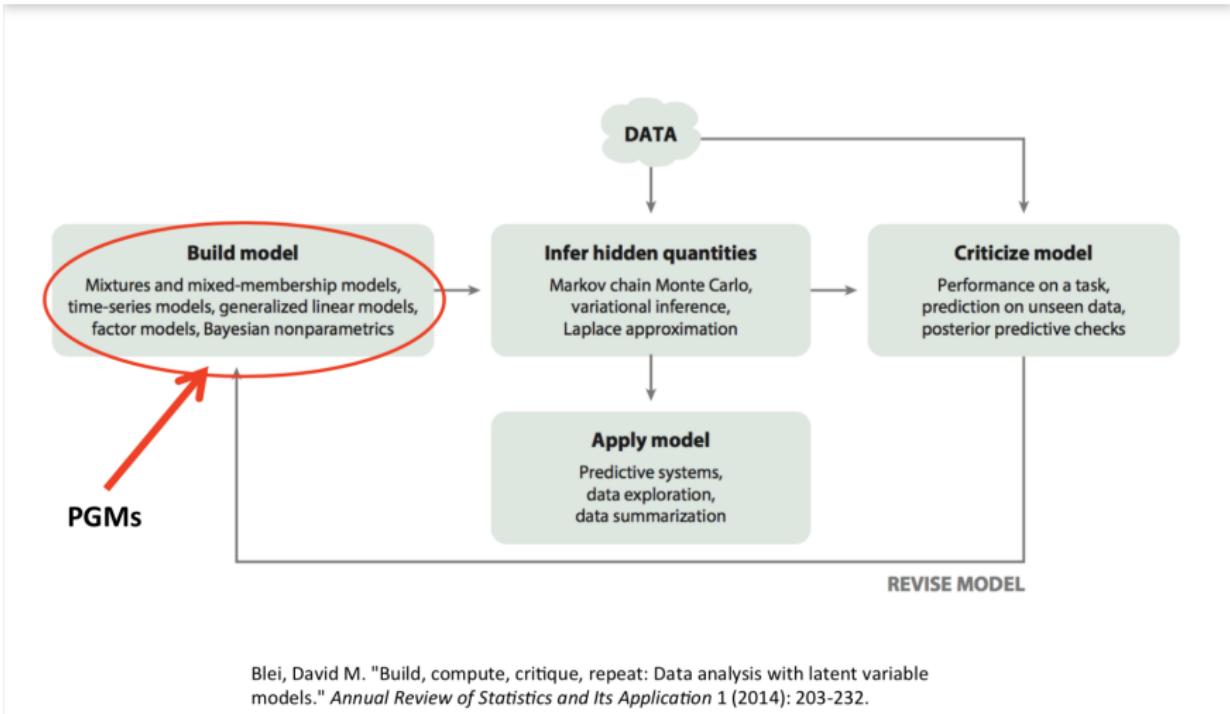
known as the **MAP (Maximum A Posteriori)** estimator

- Note that we could also choose

$$\hat{\theta} = \arg \max_{\theta} \log f(x_1, \dots, x_n | \theta)$$

which is actually the **MLE (Maximum Likelihood Estimator)**

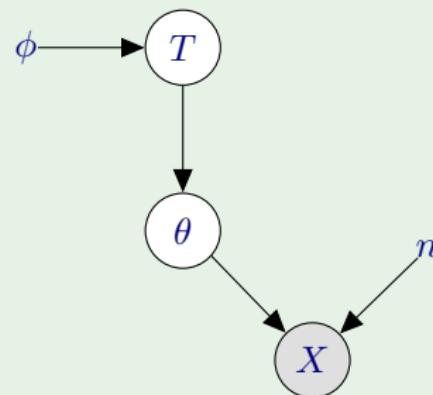
The probabilistic modelling cycle



Some simple examples

Tossing a biased(?) coin

- X : result of n coin tosses with some $p(\text{heads})$
- random variables?
- fixed quantities?
- hidden variables?
- coin is possibly biased towards tails



Some simple examples

Two coins

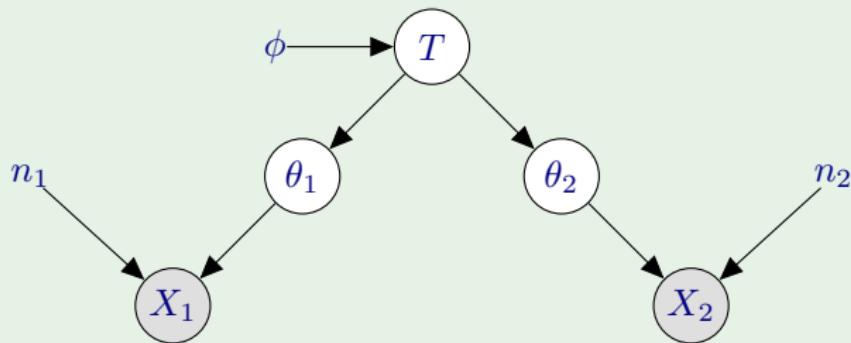


Plate notation

The idea is to avoid repeated substructures

Example: independent data points

- Assume the elements in a sample X_1, \dots, X_N are independent if the parameter θ is known

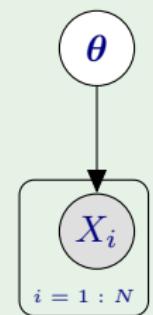
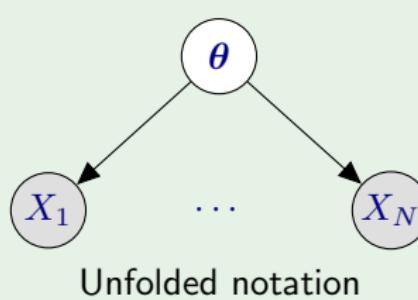


Plate notation

Plate notation: linear regression revisited

Example: linear regression (fully probabilistic)

- $Y_i | \{w, x_i\} = w^\top x_i + \epsilon_i$ with $x_i = [1, x_i]^\top$
- $\epsilon_i \sim \mathcal{N}(0, 1/\gamma)$ with precision parameter γ known
- $w \sim \mathcal{N}(\mu_0 = \mathbf{0}, \Sigma_0 = \mathbf{I}_{2 \times 2})$

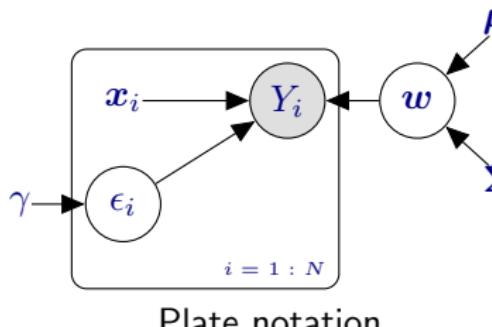
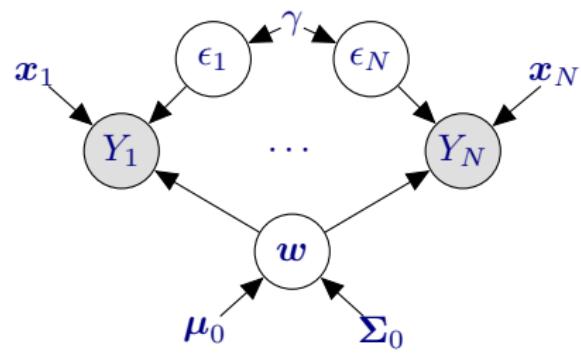


Plate notation



Unfolded notation

- Underlying model: $y_i = w_0 + w_1 x_i + \epsilon_i$

Generative and discriminative models

Predicting Y from X

Generative model

- Learn $p(x, y) = p(x|y)p(y)$ from data
- Compute $p(y|x)$ using Bayes rule

- Naive Bayes, Bayesian networks in general, ...
- Can be used to generate synthetic data
- Higher asymptotic error but reached more quickly

Discriminative model

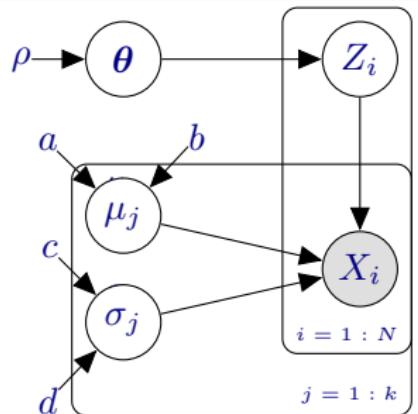
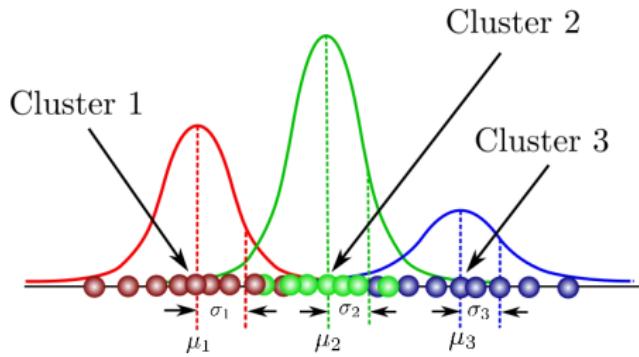
- Estimate $p(y|x)$ directly from data

- Logistic regression, NNs, ...
- Lower asymptotic error but reached more slowly

Latent variable models

Example: mixture of Gaussians: popular model for clustering

- In general, latent variable models are regarded as probabilistic models where some of the variables cannot be observed
- Model formulation:
 - k Gaussians with frequencies $\theta = (\theta_1, \dots, \theta_k)$ (sum to 1)
 - N observations generated by
 - $Z_i \sim \text{Multinom}(\theta)$, $i = 1, \dots, N$ (indicates which Gaussian)
 - $X_i | z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \dots, N$
 - Bayesian setting: priors on the parameters

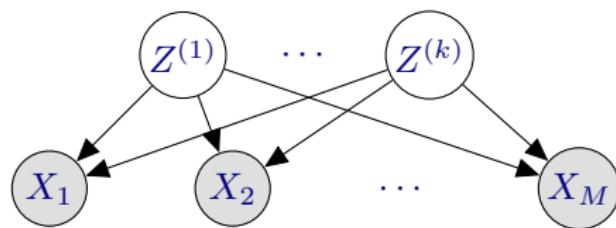
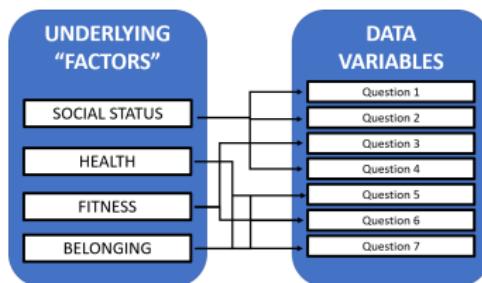


Latent variable models

Example: factor analysis model

- FA summarizes a high-dimensional observation \mathbf{X} of M correlated variables by a smaller set of factors \mathbf{Z} assumed independent a priori

Local model for single observation \mathbf{X}



$$x_1 = w_{1,1} \cdot z^{(1)} + \dots + w_{1,k} \cdot z^{(k)} + \theta_1$$
$$\vdots$$
$$x_M = w_{M,1} \cdot z^{(1)} + \dots + w_{M,k} \cdot z^{(k)} + \theta_M$$

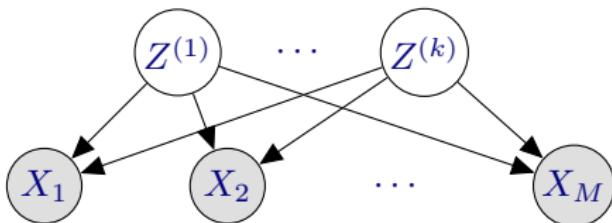
https://golden.com/wiki/Factor_analysis-RWGG9

Latent variable models

Example: factor analysis model

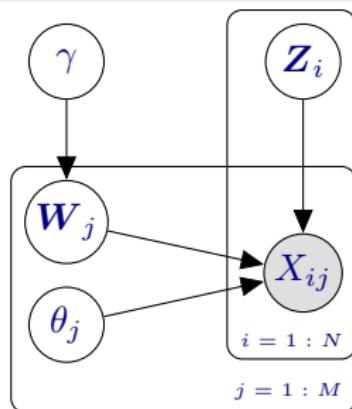
- FA summarizes a high-dimensional observation \mathbf{X} of M correlated variables by a smaller set of factors \mathbf{Z} assumed independent a priori
- Model formulation:
 - N observations \mathbf{X}_i generated by:
 - ★ $\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\Sigma}_0 = \mathbf{I})$, $i = 1, \dots, N$
 - ★ $X_{i,j} | \{\mathbf{z}_i, \mathbf{w}_j, \theta_j\} \sim \mathcal{N}(\mathbf{w}_j^\top \mathbf{z}_i, 1/\theta_j)$, $i = 1, \dots, N, j = 1, \dots, M$

Local model for single observation \mathbf{X}_i



$$x_{i,1} = w_{1,1} \cdot z_i^{(1)} + \dots + w_{1,k} \cdot z_i^{(k)} + \theta_1$$

etc.



Inference in Bayesian networks

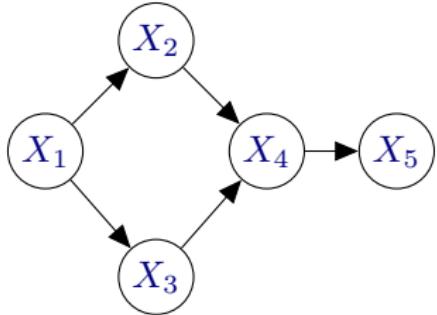
Assume a Bayesian network over variables $\mathbf{X} = \{X_1, \dots, X_n\}$

$$\left. \begin{array}{c} \text{Bayesian network} \\ + \\ \text{Evidence } (\mathbf{X}_E) \end{array} \right\} \Rightarrow P(\mathbf{X}_I | \mathbf{X}_E)?$$

Inference methods

- Exact
 - Brute force: compute $P(\mathbf{X}, \mathbf{X}_E)$ and marginalize out $\mathbf{X} \setminus \mathbf{X}_I$
 - Take advantage of the network structure
- Approximate
 - Sampling
 - Deterministic

Exact inference: Variable elimination



- We are interested in X_5
- All variables are discrete
- $E = \emptyset$

$$\begin{aligned} p(x_5) &= \sum_{x_1, \dots, x_4} p(x_1, x_2, x_3, x_4, x_5) \\ &= \sum_{x_1, \dots, x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_4) \\ &= \sum_{x_2, \dots, x_4} \sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_4) \\ &= \sum_{x_2, \dots, x_4} p(x_4|x_2, x_3)p(x_5|x_4) \boxed{\sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1)} \\ &= \sum_{x_2, \dots, x_4} p(x_4|x_2, x_3)p(x_5|x_4) \textcolor{red}{h}(x_2, x_3) \end{aligned}$$

We have reached a similar problem as initially, but with **one variable less**.

Variable elimination algorithm

Input

\mathcal{P} : Conditional distributions in the network

\mathbf{X} : Variables in the network

X_I : Target variable

\mathbf{X}_E : Observed variables

\mathbf{Y} : All variables in \mathbf{X} except X_I ($\mathbf{Y} = \mathbf{X} \setminus \{X_I\}$).

- ① Restrict all the distributions in \mathcal{P} to the evidence $\mathbf{X}_E = \mathbf{x}_E$
- ② For each $Y \in \mathbf{Y}$
 - ① Let \mathcal{P}_Y be the set of distributions in \mathcal{P} that contain Y
 - ② $q := \prod_{p \in \mathcal{P}_Y} p \implies \text{COMPLEXITY}$
 - ③ $r := \sum_y q$
 - ④ $\mathcal{P} := \mathcal{P} \setminus \mathcal{P}_Y \cup \{r\}$
- ③ $p(x_I) = \prod_{p \in \mathcal{P}} p$
- ④ Normalize p

Considerations about exact inference

- Product of functions raises complexity
 - Exponentially in the case of **discrete** variables
- Complexity also depends on the **elimination order**
- Representation of densities turns out to be relevant
 - **Closed-form** solutions to product and marginalization are preferable

Approximate inference

- **sampling:** Monte Carlo techniques, e.g. importance sampling, MCMC
 - accurate with enough samples
 - sampling can be computationally demanding
- **deterministic**, e.g. variational approaches
 - uses analytical approximations to the posterior
 - some techniques scale well

Monte Carlo inference algorithms

- A Bayesian network is a representation of a joint probability distribution over \mathbf{X} \Rightarrow it describes some **population** consisting of all the possible configurations of \mathbf{X}
- If the entire population was available, the **inference problem** could be solved exactly, basically by **counting cases**
- **Problem:** Population size can be huge or even infinite.
- **Monte Carlo** methods operate by drawing an artificial **sample** from the population using some random mechanism
- The sample (**much smaller than the population**), is used to estimate the distribution of each variable of interest.

Key issues in a Monte Carlo inference algorithm:

- ① The sampling mechanism
- ② The functions (estimators) which compute the probabilities from the sample

Conclusions

- PGMs provide a well founded way of handling uncertainty
- From a Bayesian point of view, inference and learning are connected tasks
- If scalability is important, approximate inference is needed
- Interpretability is a key issue



Thanks to Antonio Salmerón for sharing his material!