Probabilistic inference for (Bayesian) statistical inference

Silja Renooij

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

Slides are part of the ProbAI school slidedeck, containing slides contributed by Antonio Salmerón.

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

Probabilistic inference & Statistical inference

The phrase 'probabilistic inference' is often used in the PGM literature and considered synonymous to or a special case of statistical inference.
I like the following distinction:

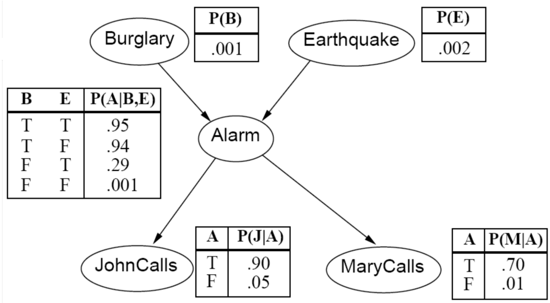
- **Probabilistic inference:** *calculate* (deduce) probabilities (or probability distributions) from a *modelled distribution with known parameters*, using *probability theory*;
- **Statistical inference:** *estimate* (infer) parameters (or other unknowns) *from data* for a hypothesized theoretical distribution, using *statistical tools*.
  - Frequentist statistics: works with point estimates; requires a lot of data;
  - Bayesian statistics: treats parameters as random variables with a distribution; already works with limited to no data.

Important observation: we can use probabilistic inference for (Bayesian) parameter estimation.

Bayesian network: definition

A Bayesian network over random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ consists of

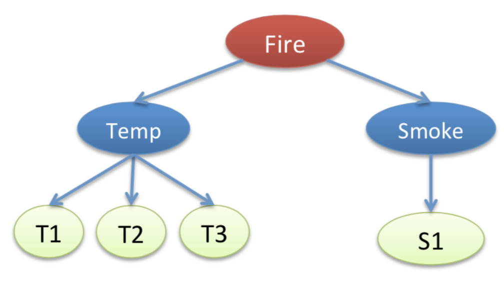
- A DAG $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ with $\mathcal{V} = \boldsymbol{X}$
- A set of local conditional distributions $\mathcal{P} = \{\ \Pr(X_i \mid \sigma(X_i)) \mid X_i \in \boldsymbol{X}\}$ where $\sigma(X_i)$ denotes the parents of X_i according to $\mathcal{A}$



Bayesian networks: compact represention 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

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \sigma(X_i))$$



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

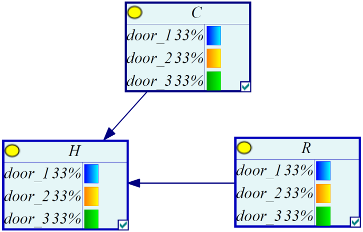
Monty Hall problem

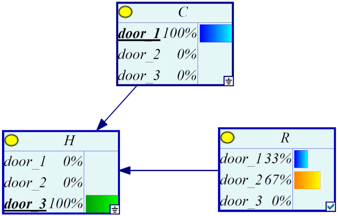
You are given the choice between 3 doors. One has a real prize behind it, the other two joke prizes.

You choose a door; the host then opens a door and offers you the choice to switch to a closed door.

Would you switch?

Probabilistic Inference





$$\Pr(H) = \sum_{c,r} \Pr(H \mid c, r) \Pr(c) \Pr(r)$$

$$\Pr(R \mid C = door_1, H = door_3) = \frac{\Pr(H = door_3 \mid C = door_1, R) \Pr(R)}{\Pr(H = door_3)}$$

From the joint distribution $\boxed{\Pr(X_1, \ldots, X_n)}$ we can infer (calculate) a.o.

- the prior distribution $\boxed{\Pr(X_i)}$ of any X_i,
- the posterior distribution $\boxed{\Pr(X_i \mid \boldsymbol{x}_E)}$ of any X_i given evidence for $\boldsymbol{x}_E$,

Note: interpretation of terms is slightly different when we consider learning!

Inference in Bayesian networks

Assume a Bayesian network over variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$

$$\left.\begin{array}{c}\text{Bayesian network,}\\ \text{variable(s) of interest } (\boldsymbol{X}_I)\\ +\\ \text{Evidence } (\boldsymbol{x}_E)\end{array}\right\} \Rightarrow P(\boldsymbol{X}_I|\boldsymbol{x}_E)?$$

Inference methods

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

Exact inference

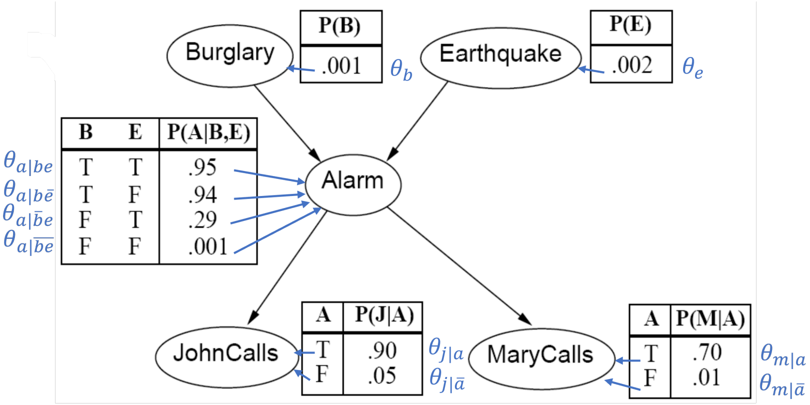
Considerations about exact inference:

- 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

Bayesian network model parameters



The probabilistic modelling cycle

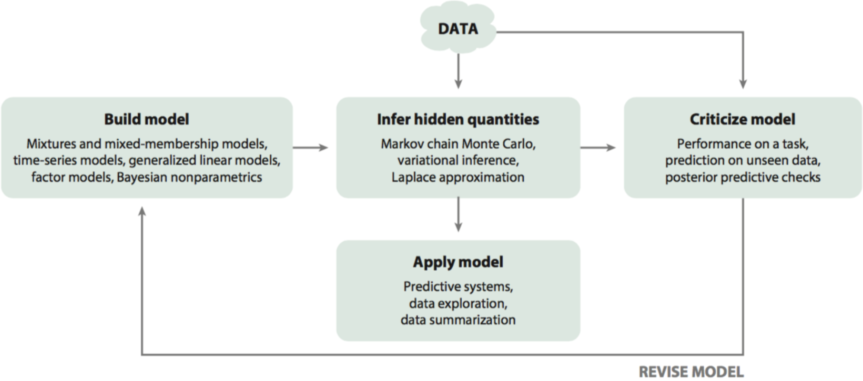


Image from: David M. Blei (2014) "Build, compute, critique, repeat: Data analysis with latent variable models." *Annual Review of Statistics and its Applications 1, 303–323.*

Learning probabilistic models from data

Model (simple):

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

Learning problem:

- We assume f is known except for parameter θ
- This is denoted as $f(x;\theta)$ or $f(x \mid \theta)$
- Goal: estimate θ

Tools:

- for a sample $X_1, \ldots, X_n$ drawn from $f(x \mid \theta)$, the likelihood function is:

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

  i.e. the joint density/mass regarded as a function of parameter θ

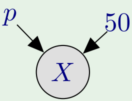
Learning parameters from data: frequentist approach

- POV: parameter θ has a fixed but unknown value

Consider tossing a (fair?) coin

Goal: estimate $p(heads)$

Frequentist POV:
probability = relative frequency
"in the long run"



After 50 trials:

$HHHTTH\ldots\ldots TH$

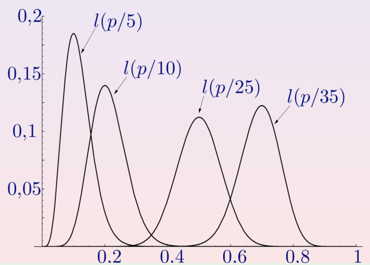
$\hat{p} = \frac{15}{50}$

What is underlying theoretical model $f(x \mid p)$? $\Rightarrow$

Assume a sample of size 1,
$X \sim \mathcal{B}(50, p)$ (Binomial, 50 trials)

The likelihood function is

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



Learning parameters from data: Bayesian approach

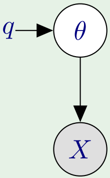
- POV: parameters are modelled as random variables → information about them can be included prior to observing data

- Additional tools: using Bayes' rule, the prior information is combined with the likelihood, yielding a posterior distribution

- The posterior then becomes the new prior

- As such, inferences about the parameter allow for its updating; to this end we can use *existing* algorithms for exact or approximate probabilistic inference!

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 - I

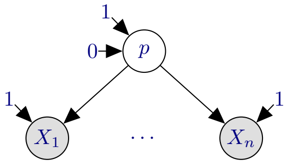
For learning:

- 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 denominator of the posterior is often a problem to compute, since we have to integrate out θ. Exception: if prior and posterior are from the same family, then exact computation is easy. Otherwise: approximate.

Learning from data: Example of Bayesian approach

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



- Then the likelihood and the prior are,

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

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

- The posterior distribution is

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

Learning from data: Example of Bayesian approach

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

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}; \qquad \int_0^1 f(p)\, dp = 1$$

$$\int_0^1 p^{\sum x_i}(1-p)^{n-\sum x_i}\, dp =$$

$$= \int_0^1 \frac{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}{\Gamma(n+2)} \frac{\Gamma(n+2)}{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}\, p^{\sum x_i}(1-p)^{n-\sum x_i}\, dp$$

$$= \frac{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}{\Gamma(n+2)} \int_0^1 \frac{\Gamma(n+2)}{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}\, p^{\sum x_i}(1-p)^{n-\sum x_i}\, dp$$

$$= \frac{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}{\Gamma(n+2)} \cdot 1$$

Learning from data: Example of Bayesian approach

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

- Then the likelihood and the prior are,

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

- The posterior distribution is

$$\pi(p|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|p)\pi(p)}{\int_0^1 f(x_1, \ldots, x_n|p)\pi(p)\; dp} = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\int_0^1 p^{\sum x_i}(1-p)^{n-\sum x_i}\; dp}$$
$$= \frac{\Gamma(n+2)}{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)} p^{\sum x_i}(1-p)^{n-\sum x_i}$$

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

Prior and likelihood are called conjugate, if prior and posterior are from same family.

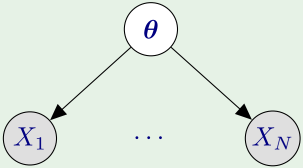
Likelihood	Prior	Posterior
$\mathcal{B}(1,\theta)$	$Be(\alpha,\beta)$	$Be\left(\alpha+\sum_{i=1}^n x_i, \beta+n-\sum_{i=1}^n x_i\right)$
$\mathcal{NB}(r,\theta)$	$Be(\alpha,\beta)$	$Be\left(\alpha+rn, \beta-nr+\sum_{i=1}^n x_i\right)$
$\mathcal{G}(\theta)$	$Be(\alpha,\beta)$	$Be\left(\alpha+n, \beta+\sum_{i=1}^n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

Plate notation

The idea is to avoid repeated substructures

Example: independent data points

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



Unfolded notation

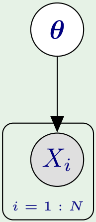


Plate notation

Learning from data: Bayesian approach

Distributions in a Bayesian model - II

For validation and use:

- The prior predictive distribution of X, $\boxed{m(x) = \int_\theta f(x|\theta)\pi(\theta)\ d\theta}$

- The (posterior) predictive distribution given $\boldsymbol{x} = \{x_1, \ldots, x_n\}$:

$$\boxed{f(x_{n+1}|\boldsymbol{x}) = \int_\theta f(x_{n+1}|\theta, \boldsymbol{x})\pi(\theta|\boldsymbol{x})d\theta = \int_\theta f(x_{n+1}|\theta)\pi(\theta|\boldsymbol{x})d\theta}$$

Example Bayesian approach, continued

- The prior predictive distribution is

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

- The (posterior) predictive distribution is

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

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

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

  known as the MAP (Maximum A Posteriori) estimator

- Note that we could also choose

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

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