Elective Course
Course Code: CS4103
Autumn 2025-26

Lecture #32

Artificial Intelligence for Data Science

Week-9:

Introduction to Probabilistic Reasoning [Part-IV]

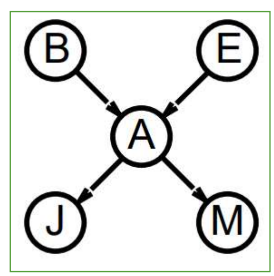
Bayesian Network Inference, Parameter Learning, Structure Learning

Course Instructor:
Dr. Monidipa Das
Assistant Professor
Department of Computational and Data Sciences
Indian Institute of Science Education and Research Kolkata, India 741246

Inference in Bayesian Networks

2

- Exact inference
  - Inference by enumeration
  - Inference by variable elimination



Simple query :

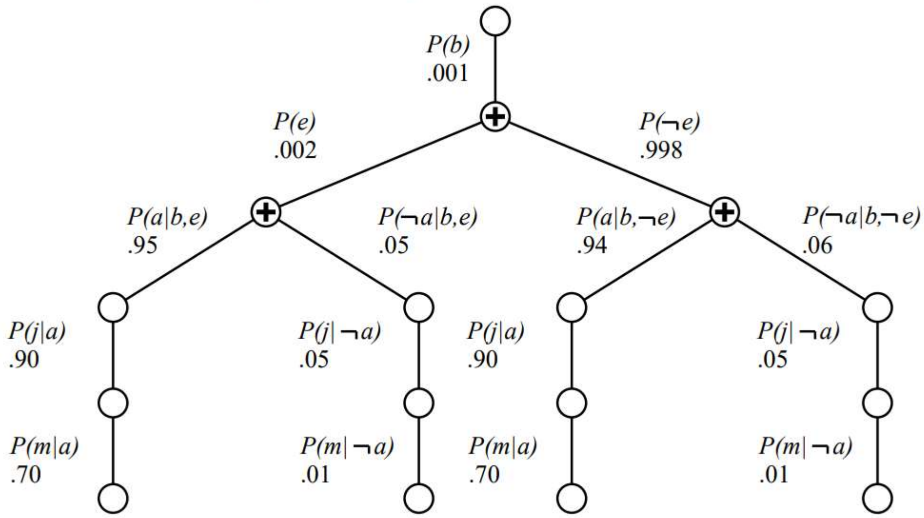
$$\mathbf{P}(B \mid j, m) = \mathbf{P}(B, j, m)/P(j, m) = \alpha \mathbf{P}(B, j, m) = \alpha \sum_e \sum_a \mathbf{P}(B, e, a, j, m)$$

Rewrite full joint entries using product of CPT entries:

$$\mathbf{P}(B \mid j, m) = \alpha \sum_e \sum_a \mathbf{P}(B) P(e) \mathbf{P}(a \mid B, e) P(j \mid a) P(m \mid a) = \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a \mid B, e) P(j \mid a) P(m \mid a)$$

Evaluation tree

$$\alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a \mid B, e) P(j \mid a) P(m \mid a)$$



Inference in Bayesian Networks

- Approximate inference

Inference by stochastic simulation

Basic idea:

1. Draw N samples from a sampling distribution S

2. Compute an approximate posterior probability $\hat{P}$
3. Show this converges to the true probability P

Inference by stochastic simulation

5

Direct Sampling

- Basic sampling: sampling with no evidence
- Rejection sampling: reject samples disagreeing with evidence
- Likelihood weighting: use evidence to weight samples

Markov chain simulation

- Markov chain Monte Carlo (MCMC): sample from a stochastic process whose stationary distribution is the true posterior

Approximate inference using MCMC

6

- "State" of network = current assignment to all of its variables
- Generate next state by sampling one var. given its Markov Blanket
- Sample each variable in turn, keeping evidence fixed

```
function GIBBS-ASK(X, e, bn, N) returns an estimate of P(X|e)
   local vars: N, a vector of counts for each value of X, initially zero
               Z, the nonevidence variables in bn
               x, the current state of the network, initially copied from e
   initialize x with random values for the variables in Z
   for j = 1 to N do
      for each Z_i in Z do
         set the value of Z_i in x by sampling from P(Z_i | MB(Z_i))
         N[x] ← N[x] + 1 where x is the value of X in x
   return NORMALIZE(N)
```

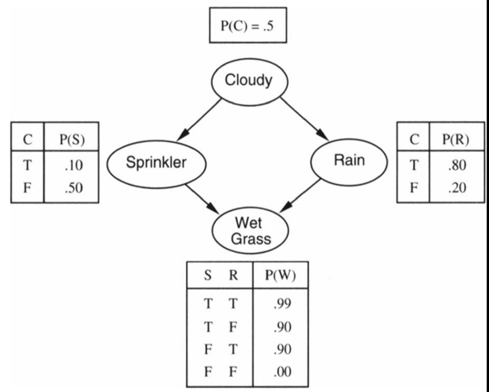
MCMC example

To estimate $\mathbf{P}(Rain \mid Sprinkler = true, WetGrass = true)$

1. Apply the Gibbs sampling algorithm with $Sprinkler$ and $WetGrass$ both fixed to $true$
2. Count number of times $Rain$ is $true$ and $false$ in the samples

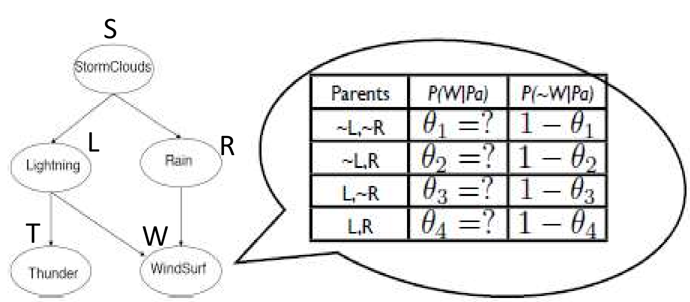
Example:
Visit 100 states; 31 have $Rain = true$, 69 have $Rain = false$

$$\hat{\mathbf{P}}(Rain \mid Sprinkler = true, WetGrass = true) = \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$$



Learning Parameters

- Parameters (θ)
  - Probabilities in the CPTs for all the variables in the network
- Learning parameters
  - Infer θ from data, given G



Learning Parameters

9

- G is a given DAG over N variables
- Goal: Estimate θ from i.i.d data $D = X = (x_1, \dots, x_M)$,
  where M is the number of records
  Each record: $x_m = \{x_m^1, x_m^2, \dots, x_m^N\}$
- Complete Observability (no missing values)

Bayes' formula as touchstone

10

$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})}$$

$$\boxed{posterior = \frac{likelihood \cdot prior}{evidence}}$$

Parameter Estimation Methods

11

- Parameter Estimation Methods
  - Frequentist Method
    - E.g.: MLE
  - Bayesian Method
    - E.g.: MAP, Bayesian Estimation

Frequentist vs. Bayesian

12

Bayesian Method	Frequentist Method
Uses probabilities for both hypotheses and data	Never uses or gives the probability of a hypothesis (no prior or posterior)
Depends on the prior and likelihood of observed data	Depends on the likelihood for both observed and unobserved data.
Requires one to know or construct a 'subjective prior'	Does not require a prior.
Dominated statistical practice before the 20th century	Dominated statistical practice during the 20th century.
May be computationally intensive due to integration over many parameters	Tends to be less computationally intensive.

MLE: Maximum Likelihood Estimator

13

- We seek that value for Θ which maximizes the likelihood (i.e. $prob(X|\Theta)$)
- We denote such a value of Θ by $\widehat{\Theta}_{ML}$
- $\widehat{\Theta}_{ML}$ maximizes $$\prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)$$ (assuming independent observations)
- Simply, $\widehat{\Theta}_{ML}$ maximizes $$\mathcal{L} = \sum_{\mathbf{x}_i \in \mathcal{X}} log\ prob(\mathbf{x}_i|\Theta)$$
- Hence,

$$\widehat{\Theta}_{ML} = \underset{\Theta}{\text{argmax}}\ \mathcal{L} \qquad \frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \qquad \forall\ \theta_i \in \Theta$$

MAP: Maximum A Posteriori estimate

14

- We seek that value for Θ which maximizes the posterior (i.e. $prob(\Theta|X)$)
- We denote such a value of Θ by $\widehat{\Theta}_{MAP}$

$$\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\text{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\text{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\text{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}$$

$$\boxed{\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left(\sum_{\mathbf{x}_i \in \mathcal{X}} log\ prob(\mathbf{x}_i|\Theta)\ +\ log\ prob(\Theta) \right)}$$

Bayesian Estimation

15

- $\widehat{\Theta}_B$ maximizes full posterior

$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})}$$

- The denominator in the Bayes' Rule cannot be ignored.
- The denominator, known as the probability of evidence, is related to the other probabilities that make their appearance in the Bayes' Rule by

$$prob(\mathcal{X}) = \int_{\Theta} prob(\mathcal{X}|\Theta) \cdot prob(\Theta) \, d\Theta$$

MLE vs. MAP vs. Bayesian Estimation

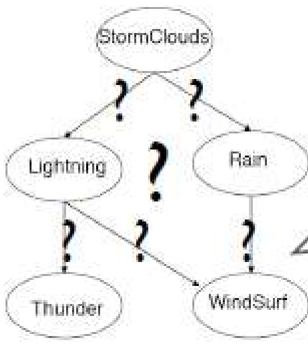
16

- MLE does NOT allow us to inject our prior beliefs about the likely values for Θ in the estimation calculations.
- MAP allows for the fact that the parameter vector Θ can take values from a distribution that expresses our prior beliefs regarding the parameters.
- Both MLE and MAP return only single and specific values for the parameter Θ.
- Bayesian estimation, by contrast, calculates fully the posterior distribution $prob(\Theta|X)$.

Learning Graph Structure

17

- Structure Learning:
  - inferring *G* and θ from data



Parents	P(W|Pa)	P(~W|Pa)
~L,~R	$\theta_1 = ?$	$1 - \theta_1$
~L,R	$\theta_2 = ?$	$1 - \theta_2$
L,~R	$\theta_3 = ?$	$1 - \theta_3$
L,R	$\theta_4 = ?$	$1 - \theta_4$

Structural Learning Methods

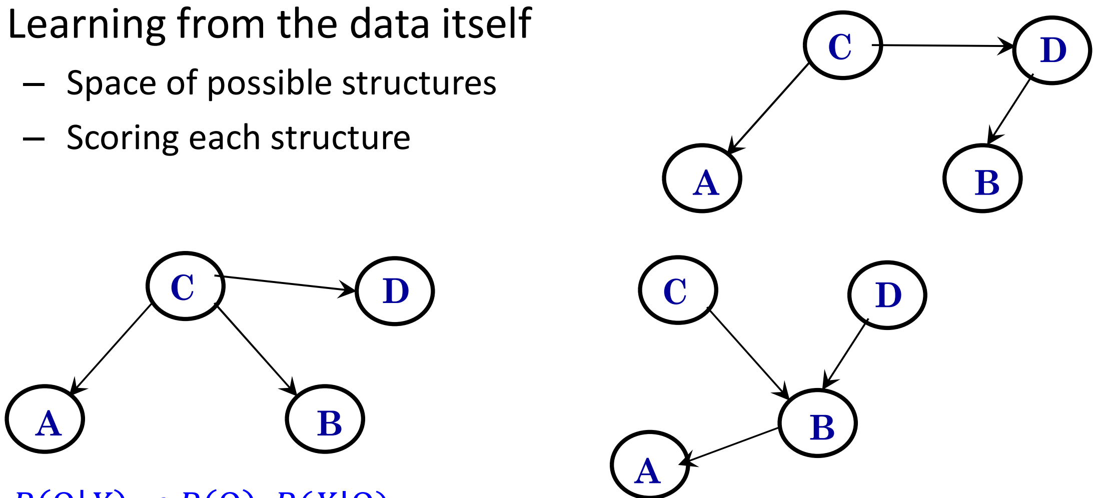
18

- Constraint Based
  - Test independencies
  - Add edges according to the tests
- Search and Score
  - Define a selection criterion that measures goodness of a model
  - Search in the space of all models (or orders)
- Mix models (recent)
  - Test for almost all independencies
  - Search and scoring

Search and Score

19

- Learning from the data itself
  - Space of possible structures
  - Scoring each structure



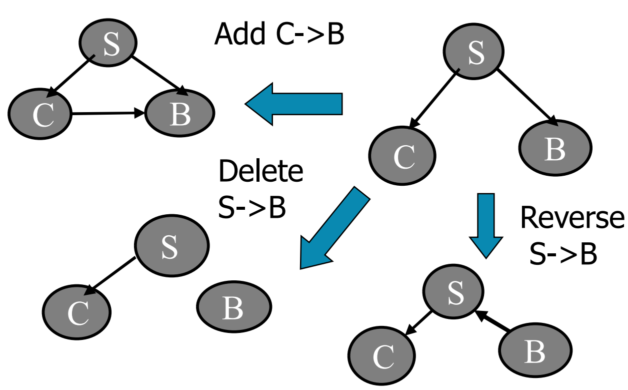
$P(\Theta|X) \propto P(\Theta).P(X|\Theta)$

Search and Score

20

Find $\hat{\mathbf{G}} = \arg\max_{G} \mathbf{Score(G)}$

- Heuristic search:
  - Greedy local search
  - Best-first search
  - Simulated annealing



Bayesian Score

21

- Main principle of the Bayesian approach

Marginal likelihood

Prior over structures

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)}$$

Marginal probability of Data

P(D) does not depend on the network

Bayesian Score: $Score_B(G : D) = \log P(D \mid G) + \log P(G)$

22

Questions?