CS5056
DATA ANALYTICS
BAYESIAN NETWORKS

HÉCTOR G. CEBALLOS, FRANCISCO J. CANTÚ

CEBALLOS@TEC.MX, FCANTU@TEC.MX

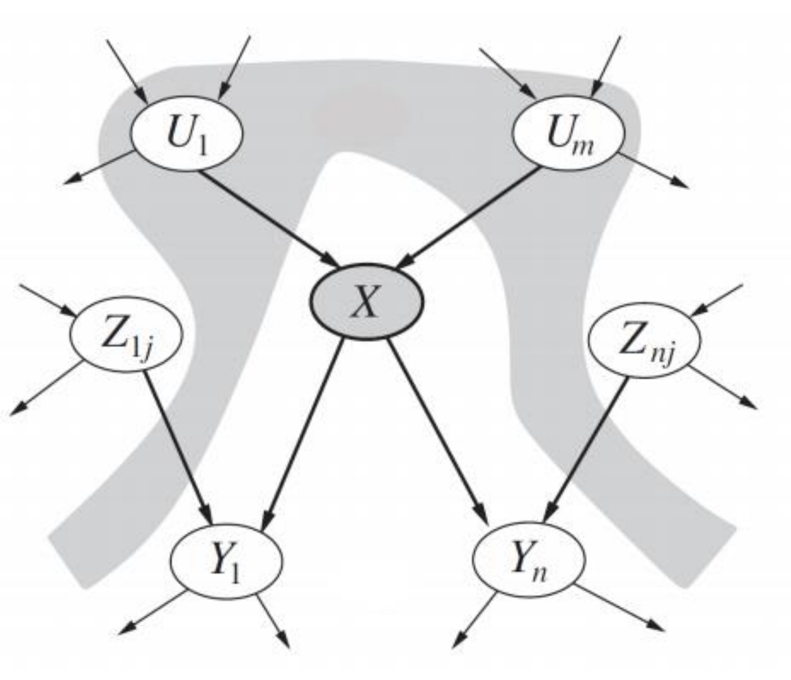


Image: https://towardsdatascience.com/graph-theory-on-to-network-theory-379b390fb19b

AGENDA

- Introduction to Bayesian Networks
- Python BNLearn
- Case Study: Discovering causal relations in BPDs
- HWA08 – Bayesian Networks

BAYESIAN NETWORKS

- A **Bayesian Network (BN)** is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- BNs are ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor.
  - A BN could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

GRAPHICAL MODEL

- BNs are directed acyclic graphs (DAGs) whose **nodes** represent variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses.
- **Edges** represent conditional dependencies; nodes that are not connected (no path connects one node to another) represent variables that are conditionally independent of each other.
- Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability of the variable represented by the node.

THE SPRINKLER EXAMPLE

Rain is independent in this model

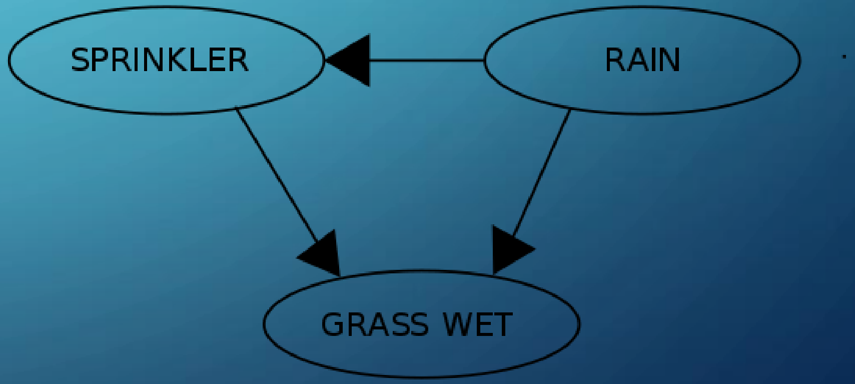
Sprinkler depends on ________

Grass wet depends on _________

Which (unobserved) variables could be used to predict rain?

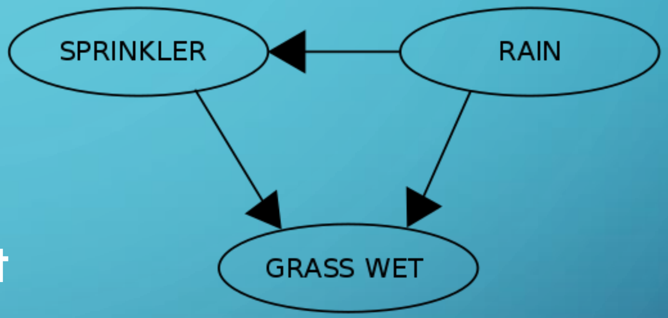
- ______________
- ______________
- ______________

Two events can cause grass to be wet: an active sprinkler or rain. Rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler usually is not active).



CONDITIONAL PROBABILITY TABLES (CPTS)

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

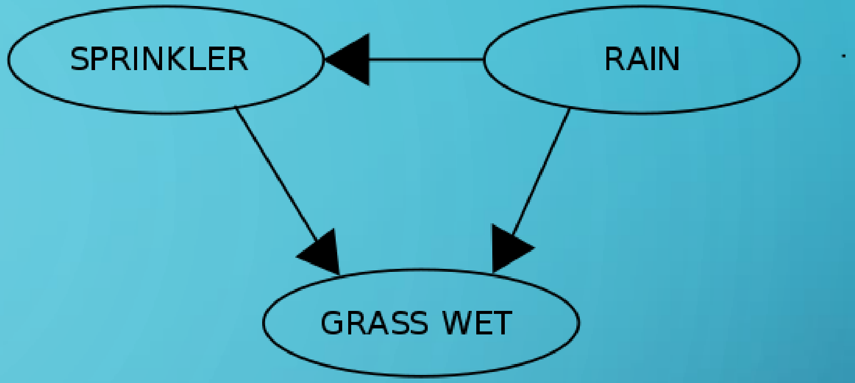
How likely is that the sprinkler is on given it is raining?

P(SPRINKLER=T|RAIN=T)

How likely is the grass is wet as long as it is raining and the sprinkler is on?

P(GRASS_WET=T|RAIN=T, SPRINKLER=T)

JOINT PROBABILITY



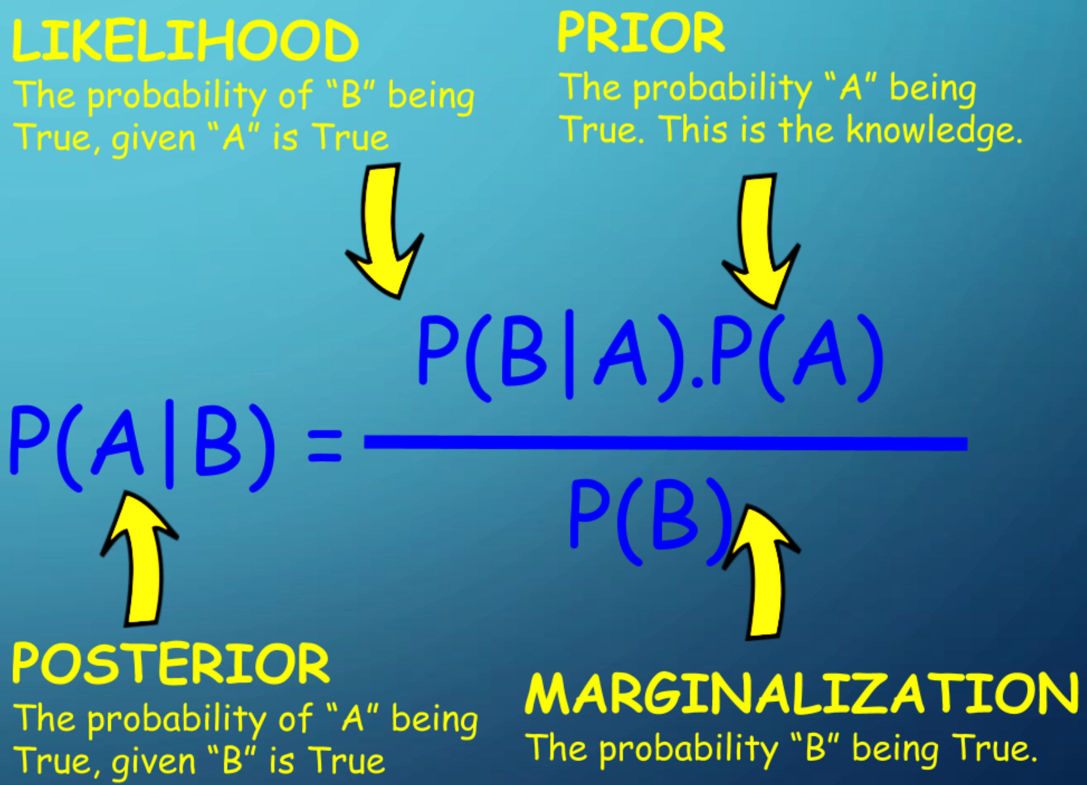
- As long as the network encodes conditional independence, the joint probability function is:

  P(G,S,R) = P(G|S,R) P(S|R) P(R)

- where G = "Grass wet (true/false)", S = "Sprinkler turned on (true/false)", and R = "Raining (true/false)".

POSTERIOR PROBABILITY

- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- For example, if a disease is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have the disease, compared to the assessment of the probability of disease made without knowledge of the person's age.



LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

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

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

https://towardsdatascience.com/bayes-rule-with-a-simple-and-practical-example-2bce3d0f4ad0

MARGINALIZATION

- When we know the value of all of the joint probabilities for a set of random variables, then we can calculate a **marginal probabilities** for one of those random variables. This can be done through the rule of **marginalization**, which states:
  - For mutually exclusive states (a_i, b_1), …, (a_i, b_n)
  - $P(a_i) = \sum_j P(a_i\,, b_j) = P(a_i, b_1) + \ldots + P(a_i, b_n)$

http://www.cse.unsw.edu.au/~cs9417ml/Bayes/Pages/Joint_Probability.html

INVERSE PROBABILITY

Bayes theorem

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

- The model can answer questions about the presence of a cause given the presence of an effect (so-called **inverse probability**).
- For instance, what is the probability that it is raining, given the grass is wet?
    - P(R=T|G=T)
- By using the conditional probability formula and summing over all nuisance variables:

$$\Pr(R=T \mid G=T) = \frac{\Pr(G=T,R=T)}{\Pr(G=T)} = \frac{\sum_{x\in\{T,F\}}\Pr(G=T,S=x,R=T)}{\sum_{x,y\in\{T,F\}}\Pr(G=T,S=x,R=y)}$$

INVERSE PROBABILITY

- Using the expansion for the joint probability function P(G,S,R) and the conditional probabilities from the CPTs, one can evaluate each term in the sums in the numerator and denominator.
- For example,

$$\begin{aligned}\Pr(G=T,S=T,R=T) &= \Pr(G=T\mid S=T,R=T)\Pr(S=T\mid R=T)\Pr(R=T)\\ &= 0.99\times 0.01\times 0.2\\ &= 0.00198.\end{aligned}$$

- Then the numerical results (subscripted by the associated variable values) are

$$\Pr(R=T\mid G=T)=\frac{0.00198_{TTT}+0.1584_{TFT}}{0.00198_{TTT}+0.288_{TTF}+0.1584_{TFT}+0.0_{TFF}}=\frac{891}{2491}\approx 35.77\%.$$

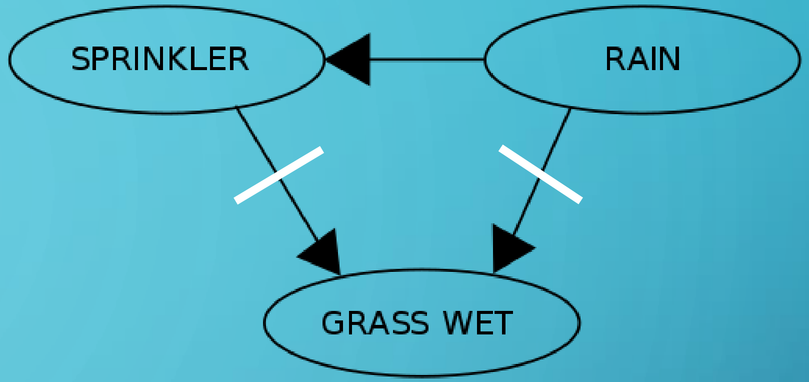
INFERENCE BY VARIABLE ELIMINATION

- The basic concept of variable elimination is same as doing marginalization over Joint Distribution.
- But variable elimination avoids computing the Joint Distribution by doing marginalization over much smaller factors.
- So basically if we want to eliminate **X** from our distribution, then we compute the product of all the factors involving **X** and marginalize over them, thus allowing us to work on much smaller factors.
- The algorithm has exponential time complexity, but could be efficient in practice for the low-treewidth graphs, if the proper elimination order is used.

INFERENCE ALGORITHMS

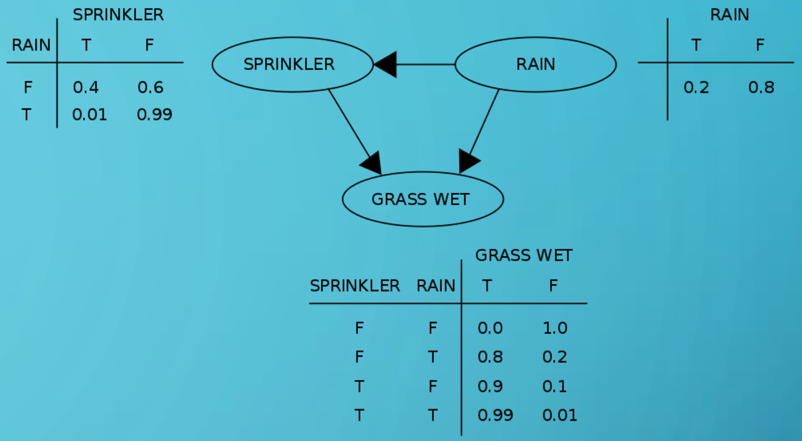
- The main categories for inference algorithms:
  - Exact Inference: These algorithms find the exact probability values for our queries.
  - Approximate Inference: These algorithms try to find approximate values by saving on computation.
- Two common Inference algorithms with variable Elimination
  - **Clique Tree Belief Propagation**. it entails performing belief propagation on a modified graph called a junction tree.
  - **Variable Elimination**. It is a simple and general exact inference algorithm.

INTERVENTION (DO)



- To answer an interventional question, such as "What is the probability that it would rain, given that we wet the grass?" the answer is governed by the post-intervention joint distribution function.
  - P(S,R|do(G=T)) = P(S|R) P(R)
- The do operator forces the value of G to be true. The probability of rain is unaffected by the action :
  - P(R|do(G=T)) = P(R)

PARAMETRIC LEARNING



	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Parameter learning is the task to estimate the values of the conditional probability distributions (CPDs).
- To make sense of the given data, we can start by counting how often each state of the variable occurs.
- If the variable is dependent on the parents, the counts are done conditionally on the parents states, i.e. for separately for each parent configuration.

PARAMETRIC LEARNING FOR DISCRETE VARIABLES

- **Maximum Likelihood Estimation** ('ml'). It uses the *relative frequencies*, with which the variable states have occurred. MLE has the problem of *overfitting* to the data.
- **Bayesian Parameter Estimation** ('bayes'). It starts with already existing prior CPDs, that express our beliefs about the variables *before* the data was observed. Those "priors" are then updated, using the state counts from the observed data.

STRUCTURAL LEARNING

- With structure learning we can estimate a DAG that captures the dependencies between the variables in the data set.
- However, realize that the search space of DAGs is super-exponential in the number of variables and you may end up in finding a local maxima.
- Commonly used scoring functions to measure the fit between model and data are Bayesian Dirichlet scores such as BDeu (‘dbeu’) or K2 (‘k2’) and the Bayesian Information Criterion (‘bic’).

STRUCTURAL LEARNING METHODS

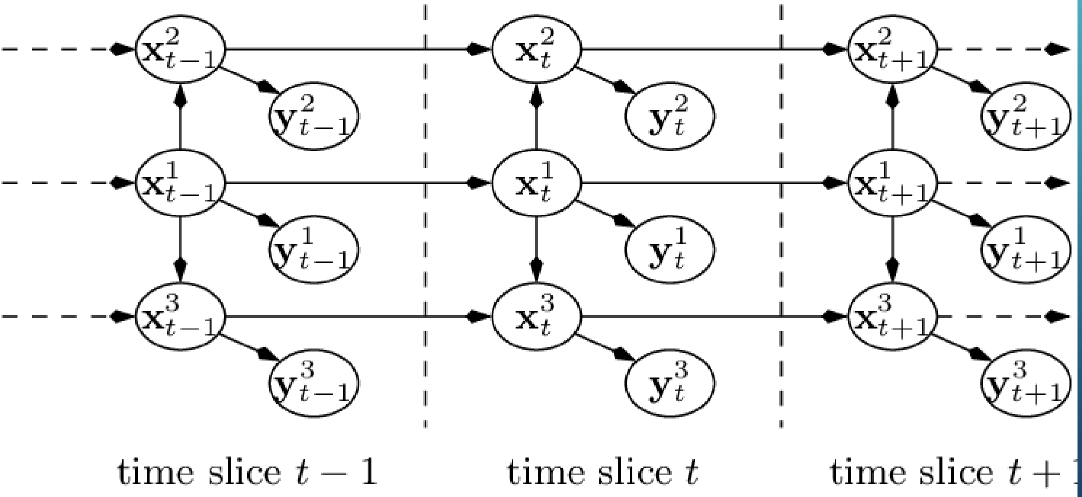
- **ExhaustiveSearch** (‘ex’) can be used to compute the score for every DAG and returns the best-scoring one. This search approach is only tractable for very small networks, and prohibits efficient local optimization algorithms to always find the optimal structure.
- **HillClimbSearch** (‘hc’) implements a greedy local search that starts from the DAG “start” (default: disconnected DAG) and proceeds by iteratively performing single-edge manipulations that maximally increase the score. The search terminates once a local maximum is found.

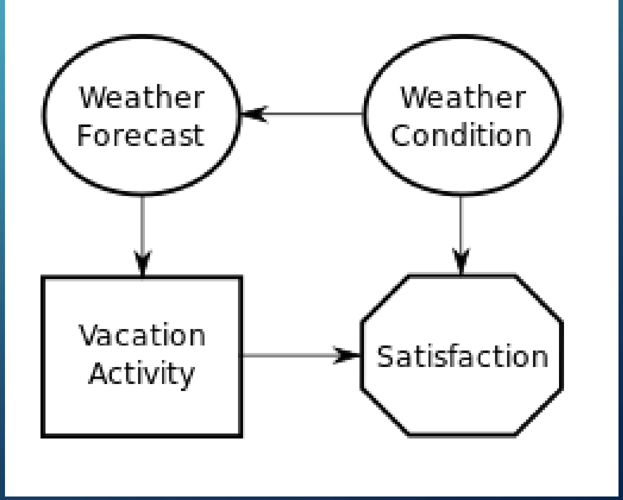
STRUCTURAL LEARNING METHODS

- The **Chow-Liu** ('cl') Algorithm is a specific type of score based approach. The Chow-Liu algorithm finds the maximum-likelihood tree structure where each node has at most one parent. Note that here our score is simply the maximum likelihood, we do not need to penalize the complexity since we are already limiting complexity by restricting ourselves to tree structures.
- **Constraint-based.** It identifies independencies in the data set using hypothesis tests, such as chi2 test statistic. The p_value of the test, and a heuristig flag that indicates if the sample size was sufficient.

EXTENSIONS OF BAYESIAN NETWORKS

- **Dynamic Bayesian Networks (DBNs):** Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences).
- **Influence Diagrams (IDs):** Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty.





HWA08 – ASIA EXAMPLE

- Look at the demo of the Asia Bayesian Network:

  https://www.bayesserver.com/examples/networks/asia

- Follow the instructions of HWA08 at Canvas.