IS 7033: Artificial Intelligence and Machine Learning

Dr. Peyman Najafirad (Paul Rad)

Associate Professor

Cyber Analytics and AI

210.872.7259

https://github.com/paulNrad/ProbabilisticGraphModels

Bayes Networks

Reading

- Kevin Murphy, Machine Learning: A probabilistic Perspective, Chapter 10
- Chris Bishop, Pattern Recognition and Machine Learning, Chapter 8
- Jordan, M. I. (2007). An introduction to probabilistic graphical models. In preparation (Chapter 2) –Also review article entitled.

Bayes Theorem

θ — parameters

X — observations

Posterior Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)}$$

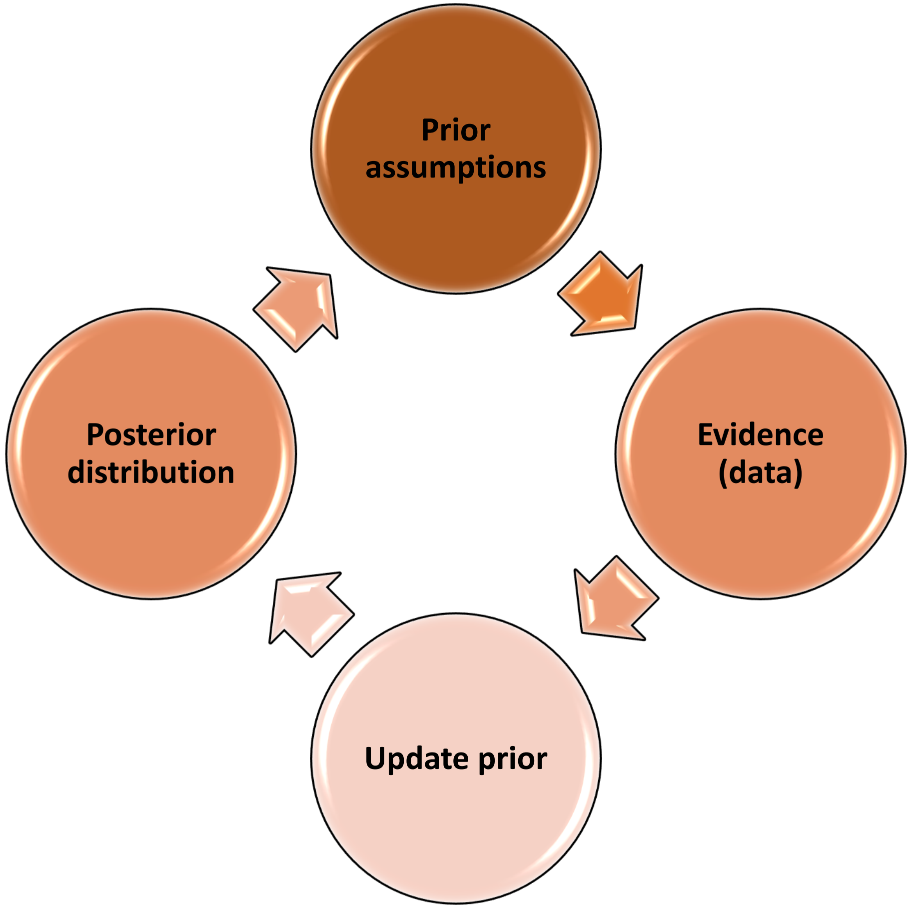
Evidence

Bayesian Inference

- Example from Bayesian Methods for Hackers:
  - You know the code you write is likely to have some bugs somewhere. So you start testing it against a really simple case; it passes. Continue to increase the complexity of the cases you test against. The more complex cases it passes, the more you're sure that the code is bug-free. You are already thinning Bayesian!

- Updating beliefs based on evidence
  - Never 100% sure, same as in software testing, but we can get pretty close

Bayesian Inference



Bayesian Inference via probabilistic Programming

- Solving Bayes' theorem in practice requires taking integrals
- If you don't want to do that, we need to use numerical solution methods
- Lots of development in terms of new methods of sampling
  - Markov Chain Monte Carlo and Hamiltonian Monte Carlo
  - NUTS, No Turn Sampler such as Gelman
  - Variational inference
    - Make distributions similar to each other
    - Optimization, not sampling, so more appropriate for big data

Bayesian Inference

- Use to be called “inverse probability”, now called the posterior distribution
  - What is the most likely value of our parameter of interest, conditioning on the data we observe?
  - Reason from effects (observations) to causes (parameters)

- Outputs differ from traditional Frequentist statistics
  - Frequentist: point estimates of parameters and confidence intervals
  - Bayesian: posterior probability distributions on parameters

Introduction

- Key problems in probabilistic modeling of complex high-dimensional problems include the following?
  - Once we observe multiple correlated variables, how can we compactly represent the joint distribution $p(x|\theta)$
  - How can we lean the parameters θ of this distribution with limited data?
  - How can we reduce computational complexity

Why graphical models?

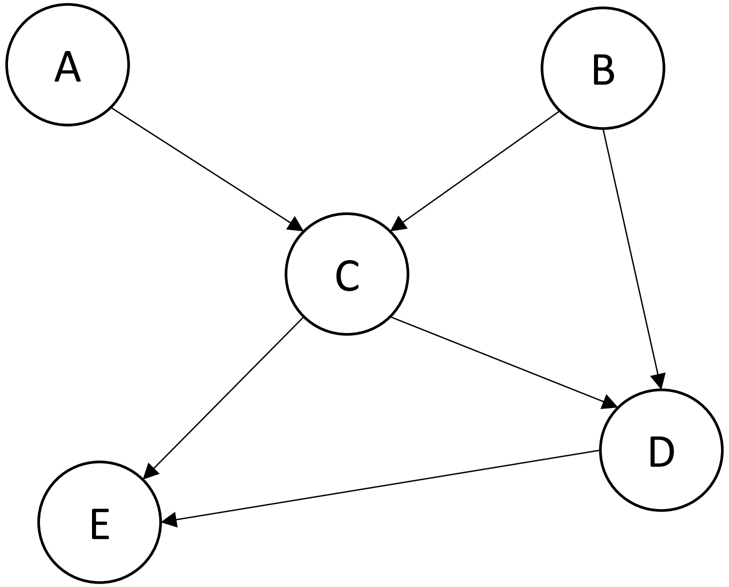
- Graphs are an intuitive way of representing and visualizing the relationships between many variables.

- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can answer questions like:
  - "Is A dependent on B given that we know the value of C?"

- Graphical models allow us to define general message-passing algorithms that implement probabilistic inference efficiently. Thus we can answer queries like:
  - "What is $p(A|C=c)$?" without enumerating all settings of all variables in the model.

Representing knowledge through graphical models

- Nodes correspond to random variables
- Edges represent statistical dependencies between the variables



Graphical models = statistics x graph theory x computer science

Joint Probability Distributions

- Consider a set of random variables (X_1, X_2, …X_v)
- Joint Probability Distribution using Chain Rule:

P(X_1, X_2, …X_v) = $\prod p(x_i \mid x1, \ldots, x_{i-1})$

$$= p(x_1)\, p(x_2|x_1)\, p(x_3|x_2,x_1) \,..\, p(x_N|x_{N-1},.., x_2,x_1)$$

Applications of PGMs

- Bio-informatics
- Medical Diagnostics
- Computer Vision
- Speech Recognition
- Most area of Machine Learning and Computational Statistics
- Natural Language Processing
- Many more

Terminology

A graph G = ($\mathcal{V}$,$\mathcal{E}$) comprises nodes or vertices connected by links or edges.

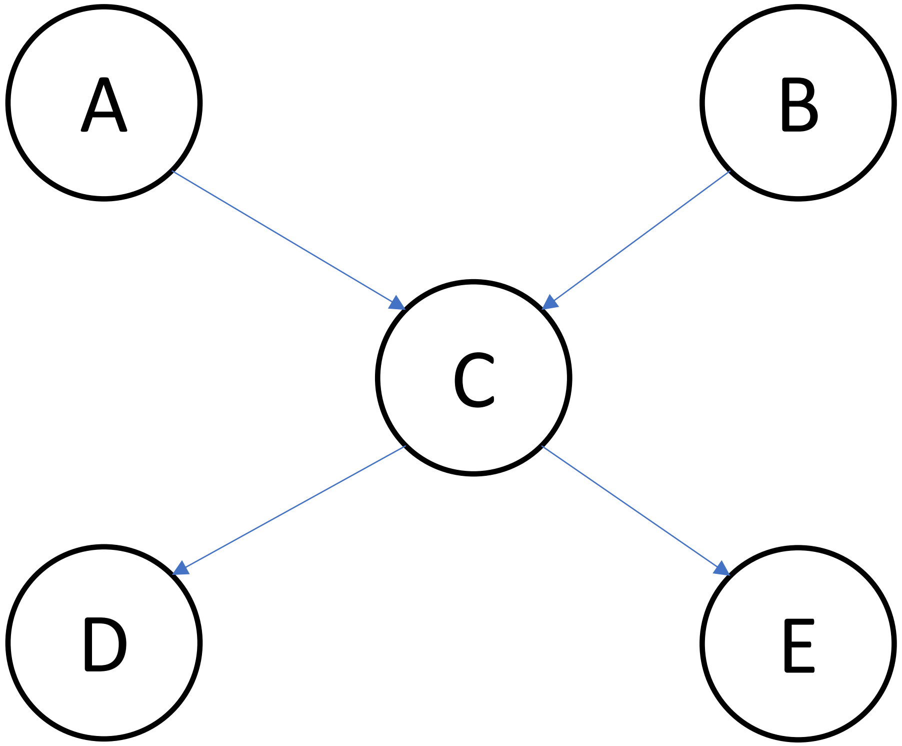
Node represents random variables. $\mathcal{V} = \{v_1,\ldots,v_n\}$

Edge expresses probabilistic relationship or dependencies between these variables. $\mathcal{E} = \{(vi, vj): vi, vj \subseteq \mathcal{V}\}$

- The graph captures the way in which the joint distribution over all of the random variables can be decomposed into product of factors each depending only on a subset of the variables.
- A graph representation abstracts out the conditional independence relations between the variables from the actual probabilistic distributions.

Bayes Networks

- Define probability distribution over graphs of random variables



P(A), P(B)
P(C|A,B)
P(D|C) P(E|C)

$2^5 - 1 = 31$

P(A,B,C,D,E) =

$$\underbrace{P(A) * P(B) * P(C|A,B) * P(D|C) * P(E|C)}_{\text{Only 10}}$$

Conditional Probability Tables

Consider the chain rule of probability, using any ordering of N variables, we can write a joint distribution as:

$$p\,(x_1, x_2, .., x_N) = \prod_{i=1}^{N} p(x_i \mid x1, \ldots, x_{i-1})$$

$$= p(x_1)\, p(x_2|x_1)\, p(x_3|x_2,x_1) \,..\, p(x_N|x_{N-1},.., x_2,x_1)$$

It becomes expensive to represent $p(x_N|x_{N-1},.., x_2,x_1)$. For discrete random variables each with K states we need K^{N-1} parameters. Computational Complexity = $O(K^N)$

Addressing the Curse of Dimensionality

We need a lot of data to learn $O(K^N)$ parameters

$$p\ (x_1, x_2, .. , x_N) = \prod p(x_i \mid x1, \ldots, x_{i-1})$$

$$= p(x_1)\, p(x_2|x_1)\, p(x_3|x_2,x_1)\ ..\ p(x_N|x_{N-1},.., x_2,x_1)$$

Marginal Independence

Marginal independence

$x \perp y = y \perp x$ means $p(x , y) = p(x) * p(y)$

Conditional Independence

X independent of Y given Z if for all values of Z,

$X \perp Y \mid Z$,if

$p(X|Y, Z) = p(X|Z)$, when $p(Y,Z) > 0$

Also we can write:

$p(X, Y \mid Z) = p(X|Y,Z) * p(Y|Z) = p(X|Z) * p(Y|Z)$

Conditional independence

- To efficiently representing large joint distribution we make conditional independence (CI) assumptions. X, Y are conditional independence given Z, denoted $X \perp Y \mid Z$, iff

$$p(X, Y \mid Z) = p(X|Z) * p(Y|Z)$$

Let us see how conditional independence can address the curse of dimensionality. Consider a Markov Chain with $X_{t+1} \perp X_{t-1} \mid X_t$ the future is independent of the past given the present. **We call this the first order Markov assumption**.

Conditional independence

Then during the chain rule for a Markov Chain

$$p\,(x_1, x_2, .., x_N) = p(x_1) \prod_{i=2}^{N} p(x_i \mid x_{i-1})$$

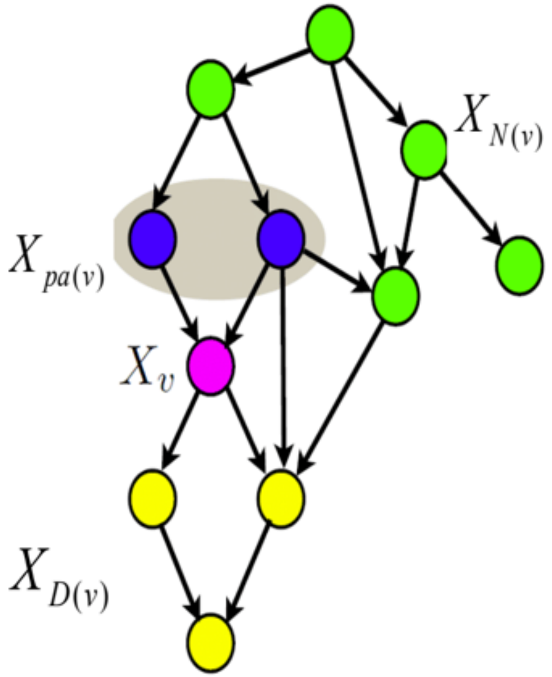
to characterize this 1st order Markov chain, we need to initial distribution, $p(x_1)$ and a state transition matrix $p(x_i \mid x_{i-1})$

Local Markov Property in DAGs

- DAGs are known as Bayesian Networks or Belief Networks
- Nodes are random variables and **edges represent causation**. No directed cycles allowed. The Graph is a DAG (Directed Acyclic Graph)

- **Local Markov property**: node is conditionality independent of its non-descendants given its parents

$$\{ X_v \perp X_{N(v)} \} | \ X_{parent(v)}$$

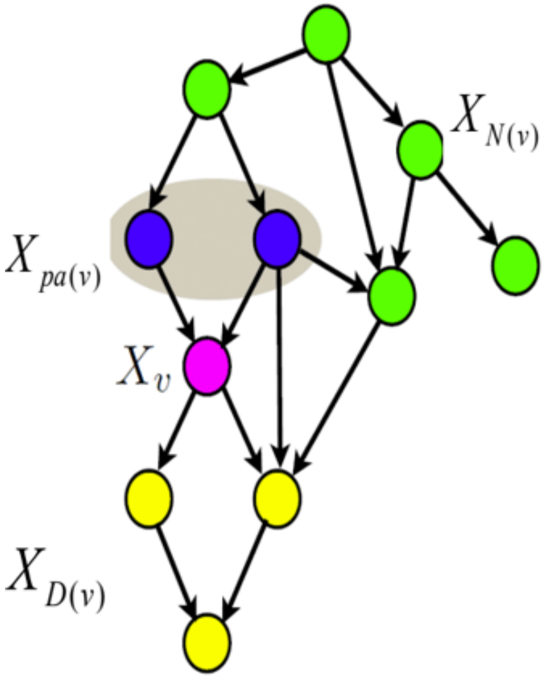


Local Markov Property

- In DAGs the nodes can be ordered such that parents come before children. This is called a **topological ordering.**
- I: ordering of the nodes in graph G is topological if for every node X_v the parents of the node appear before V_x in I.

- **Ordered Markov property** in DAGs: A node only depends on its immediate parents, not on all **predecessors** in the ordering

$$\{ X_v \mid X_{predecessors(v)} \} \mid X_{parent(v)}$$

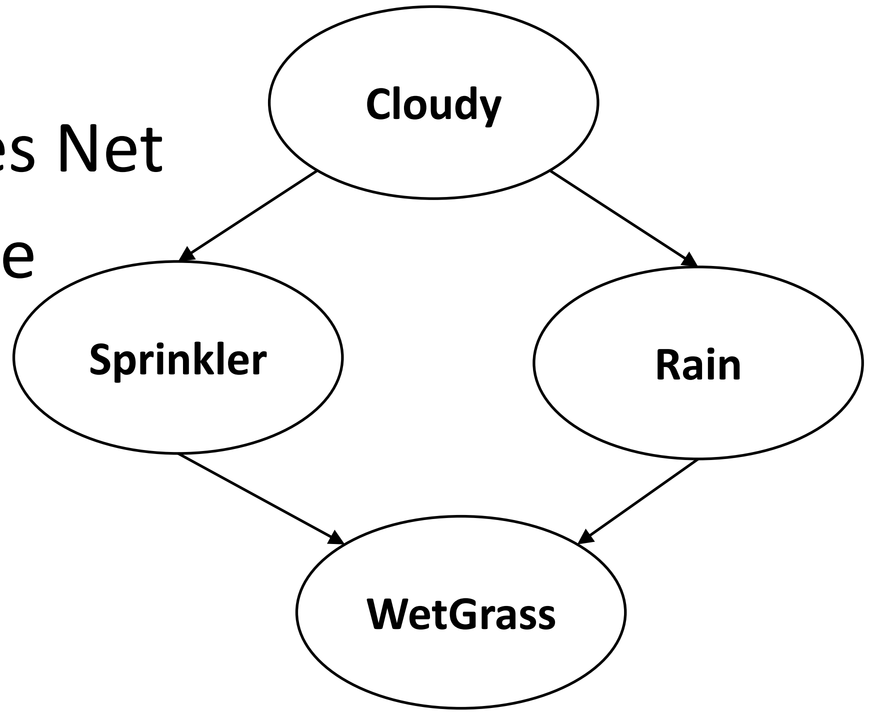


$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n}\, p(X_i \mid Parents(X_i))$$

Bayesian Network Example

- The joint probability distribution for the Bayes Net

Using Chain rule and Conditional Independence



$(R \perp S \mid C)$ && $(W \perp C \mid S, R)$

Then

$P(C, S, R, W)$

$= P(C) * P(R|C) * P(S|R,C) * P(W|S,R,C)$

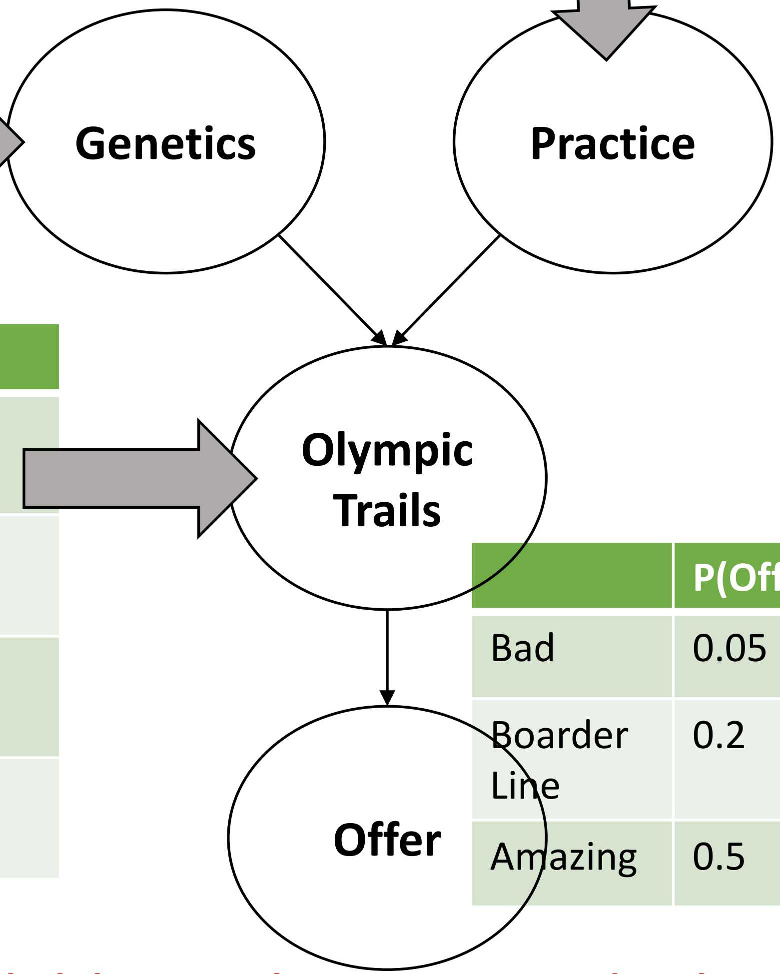
$= P(C) * P(R|C) * P(S|C) * P(W|S,R)$

Bayesian Network Example

Conditional Probability Distributions

Practice	P(Practice)
Yes	0.7
No	0.3

Genetics	P(Genetics)
Good	0.2
Bad	0.8



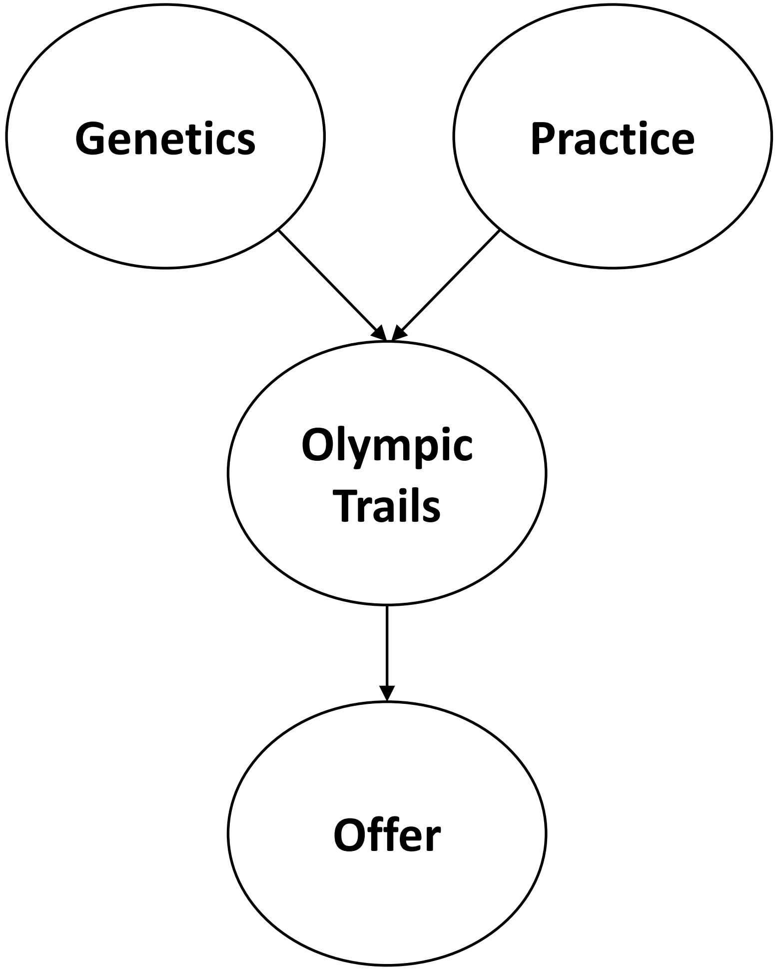
	Bad	Border Line	Amazing
Good Genes, Did Practice	0.5	0.3	0.2
Good Genes, Didn't Practice	0.8	0.15	0.05
Bad Genes, Did Practice	0.8	0.1	0.1
Bad Genes, Didn't Practice	0.9	0.08	0.02

	P(Offer)	P(-Offer)
Bad	0.05	0.95
Boarder Line	0.2	0.8
Amazing	0.5	0.5

- **Each node (random variable) in our Bayes Net has a Conditional Probability Distribution associated with it.**
- **If a node has parents, the associated Conditional Probability Distribution represent P(value | parents value)**

Questions

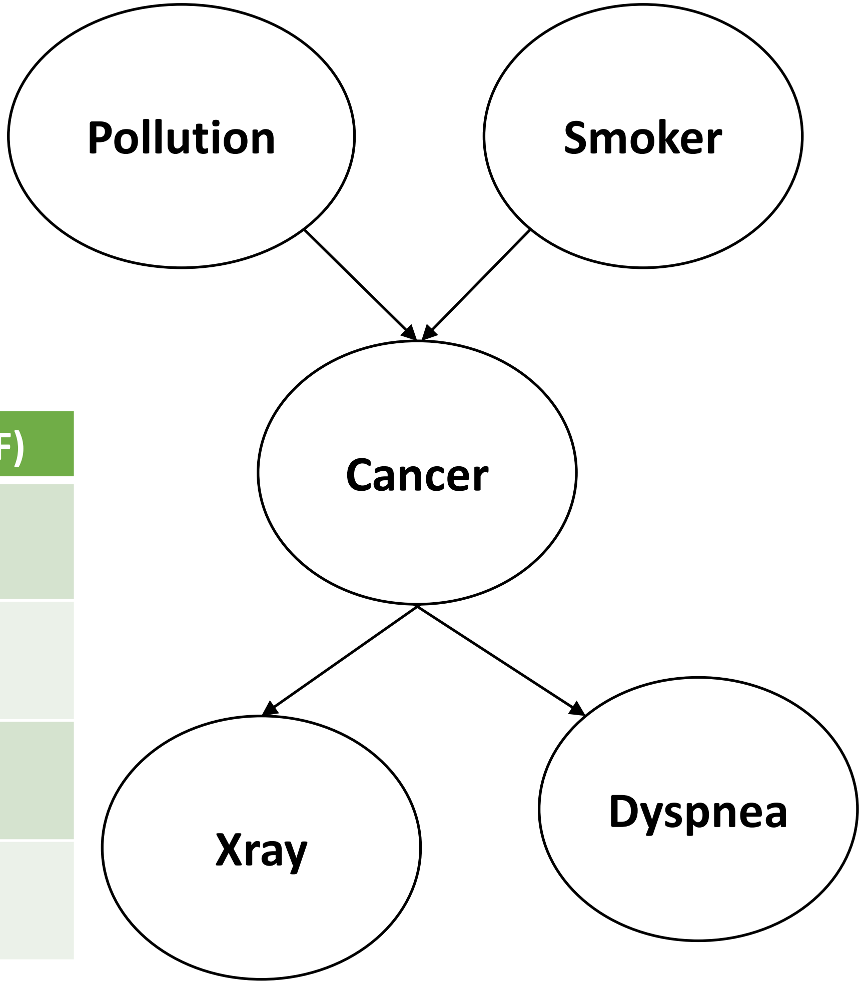
- Does an Offer depend on Genetics?
- Does an Offer depend on Genetics if you know Practice?
- Does an Offer depend on Genetics if you know Olympic Trails performance?



Bayesian Network Example

Smoker	P(Smoker)
True	0.3
False	0.7

Pollution	P(Pollution)
Low	0.9
High	0.1



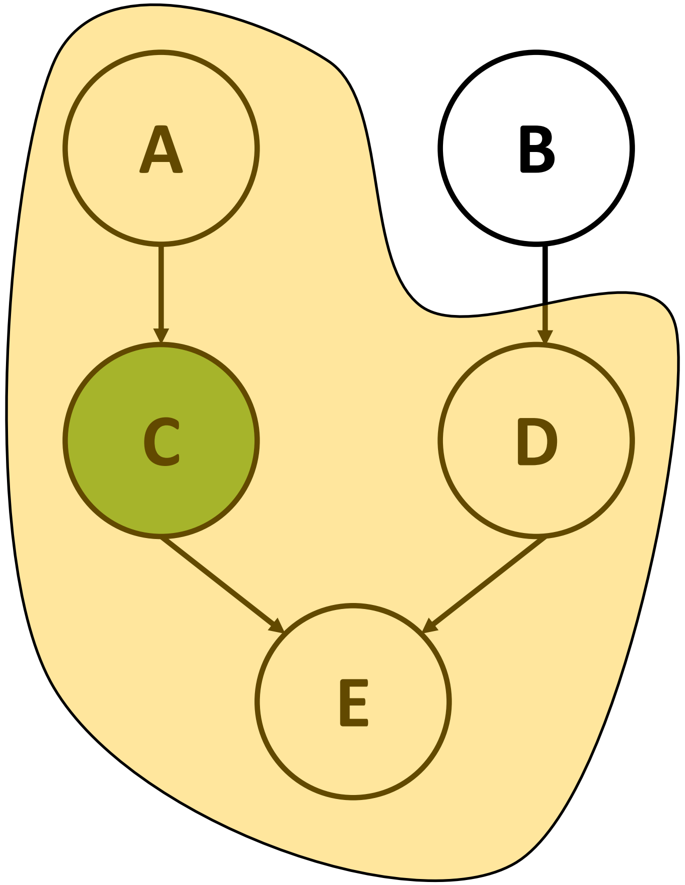
Smoker	Pollution	P (Cancer = T)	P (Cancer = F)
False	Low	0.001	0.999
True	Low	0.03	0.97
False	High	0.02	0.98
True	High	0.05	0.95

Independence in Bayes Nets

- Each variable is conditionally independent of its non-descendants given its parents
- Each variable is conditionally independent of any other variable given its **Markov blanket**
  - Parents, children, and children's parents

C's Markov blanket:

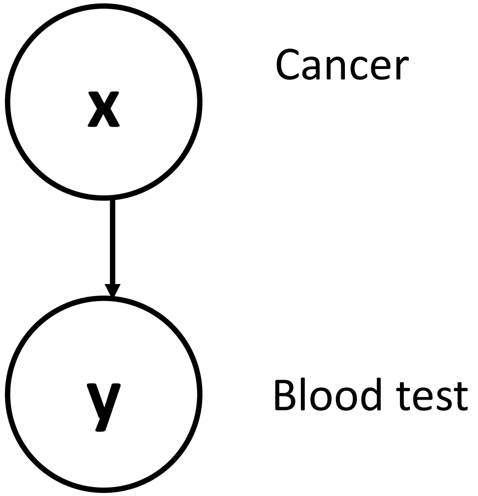
$C \perp B \mid A, D, E$



Causality in DAGs

- Directed graphs can express causality

- By observing child variables, we can infer the posterior distribution of parent variables



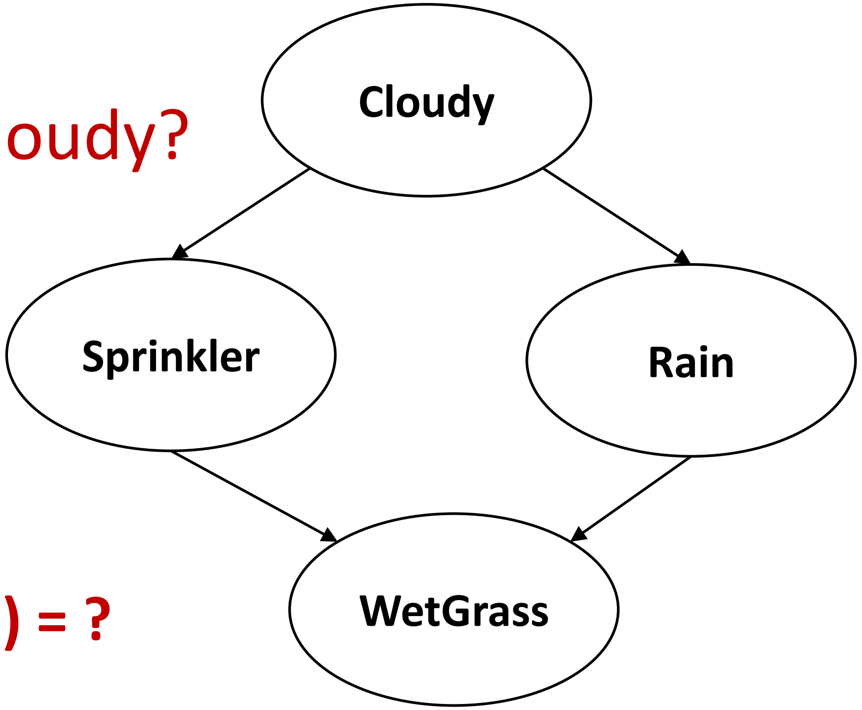
P(x|y) = P(x). P(y/x) / $\sum_{x'} P(x')p(y|x')$

Causal Reasoning – Prediction

- Given a set of observed variables, we want to estimate the values of hidden variables – this is an inference problem

<u>From causes to effects</u>

How likely is for the grass to be wet of it is cloudy?

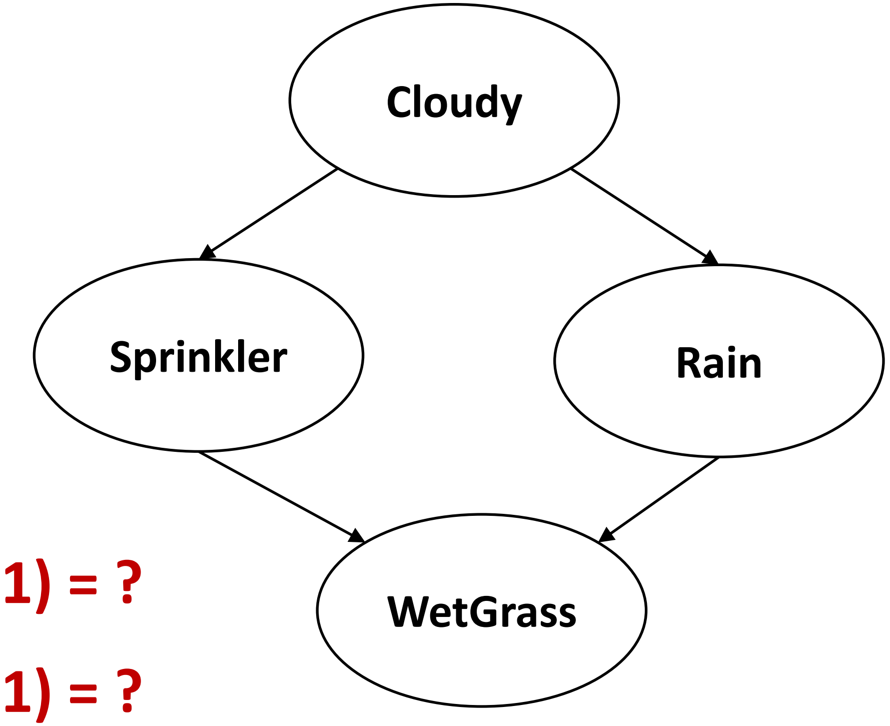


$P(W=1|c=1) = ?$

Diagnostic or Evidential Reasoning

- Given a set of observed variables, we want to estimate the values of hidden variables – this is an inference problem

From effects to causes



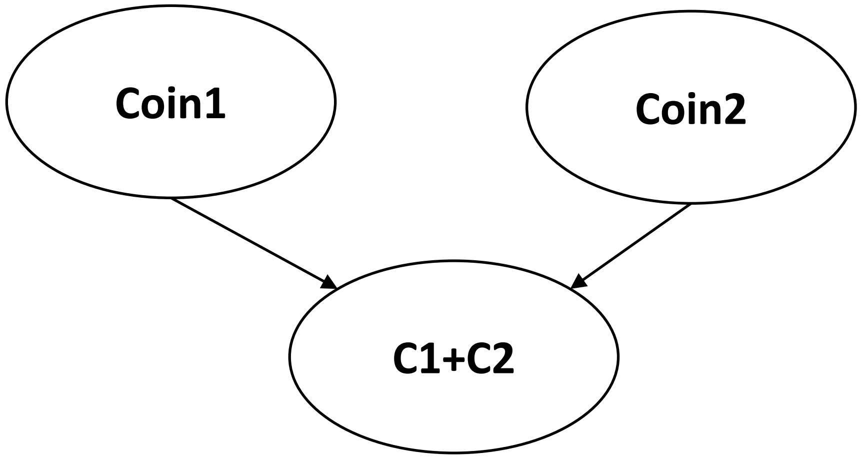
$P(R=1|W=1) = ?$

$P(S=1|W=1) = ?$

Explaining Away

- Suppose we toss coins representing the binary numbers 0 and 1, and we observe the sum of their values.

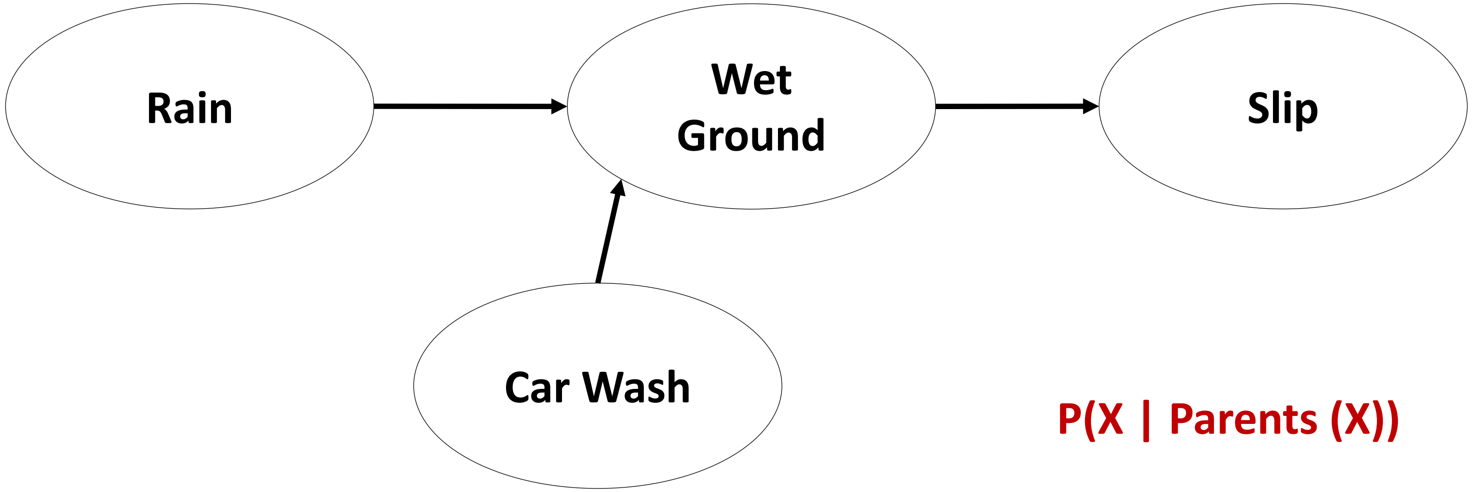
- A prior, the coins are independent, but once we observe their sum, they become coupled
  - e.g. if the sum is 1,
    and the first coin is 0,
    then we know the second coint is 1



Inference

- Given a Bayesian Network describing $P(X,Y,Z)$, what is $P(Y)$
  - First approach: **enumeration**
  - Second approach: **Variable Elimination**

Bayesian Networks



P(X | Parents (X))

P(R, W, S, C)

= P(R) P(C) P(W|R,C) P(S|W)

Enumeration approach

P(R, W, S, C)

= P(R) P(C) P(W|R,C) P(S|W)

P(R = r|S = s) = P(R, S)/P(S) =$\sum_w \sum_c \; P(R, W, S, C)/P(S)$

P(R = r|S = s) ∝ $\sum_w \sum_c \; P(R, W, S, C)$

= $\sum_w \sum_c$ P(R) P(C) P(W|R,C) P(S|W)

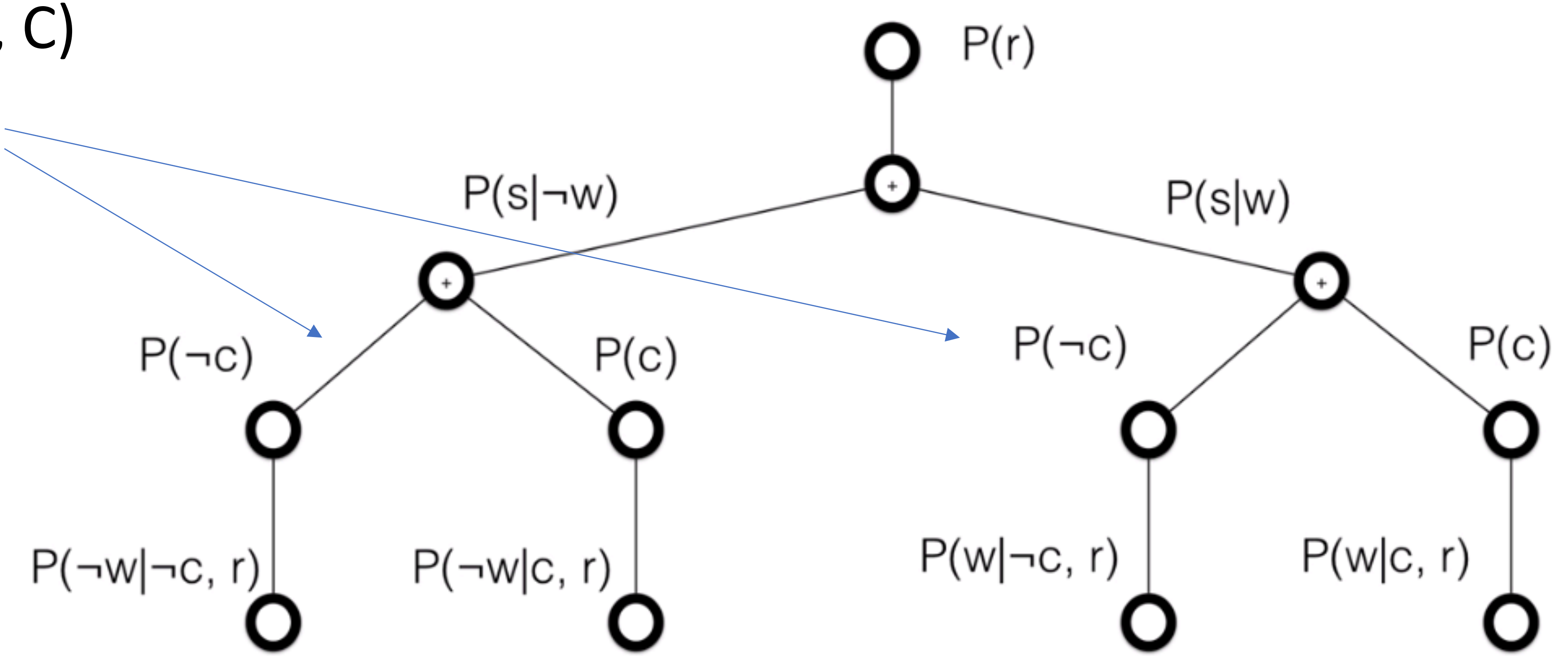
= P(R) $\sum_w$ P(S|W) $\sum_c$ P(C) P(W|R,C)

Enumeration approach

$$P(R = r \mid S = s) \propto P(R) \sum_{w} P(S \mid W) \sum_{c} P(C)\, P(W \mid R, C)$$

P(R, W, S, C)

$O(2^n)$



Variable Elimination

$$P(R = r|S = s) \propto P(R) \sum_{w} P(S|W) \sum_{c} P(C)\, P(W|R,C)$$

$f_c(w) = \sum_c P(C)\, P(W|R,C)$

$$P(R = r|S = s) \propto P(R) \sum_{w} P(S|W) f_c(w)$$

$$W \rightarrow X \rightarrow Y \rightarrow Z$$

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

P(Y)?

$$P(Y) = \sum_w \sum_x \sum_z P(W)P(X|W)P(Y|X)P(Z|Y)$$

$$f_w(x) = \sum_w P(W)P(X|W)$$

$$P(Y) = \sum_x \sum_z f_w(X)\, P(Y|X)P(Z|Y)$$

$$f_x(Y) = \sum_x f_w(X)\, P(Y|X)$$

$$P(Y) = \sum_z f_x(Y)\, P(Z|Y)$$

Variable Elimination

- Every variable that is not an ancestor of a query variable or evidence variable is irrelevant to the query

Loop

Choose variable to eliminate

Sum terms relevant to variable, generate new factor

While no more variable to eliminate

In tree structure BNs are linear time.