Learning Analytics

Dr. Bowen Hui
Computer Science
University of British Columbia Okanagan

Last Class

- Review of probability
  - Basic terminology: random variables, joint distribution
  - Conditional probability, sum-out rule, product rule
  - A few calculation examples

- All in the context of multiagent interaction
  - Inference to model our world
  - Estimate values of hidden variables using observations

Introduced Bayes Rule

- Real world problems typically requires us to compute $Pr(H|e)$
  - Recall Asian flu example: given $Pr(A)$, $Pr(F)$, $Pr(F|A)$

- Bayes rule rewrites $Pr(H|e) \propto Pr(e|H)Pr(H)$

Posterior probability	$\propto$	Likelihood	X	Prior probability

Belief Perseverance

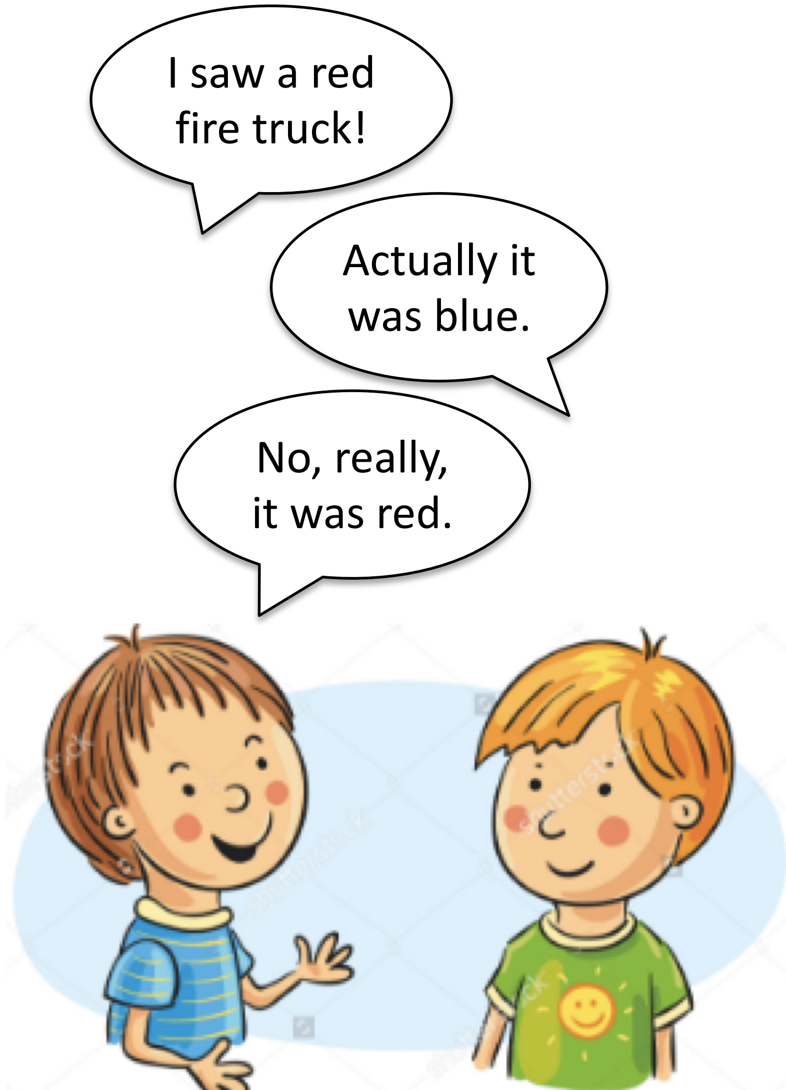


Image taken from www.fotosearch.com

Image taken from www.shutterstock.com

Changes in Representation

- Propositional logic
  - E.g. P = John sees Mary.
  - E.g. If P is true then Q is also true

Changes in Representation

- Propositional logic
  - E.g. P = John sees Mary.
  - E.g. If P is true then Q is also true

- Predicate logic
  - E.g. sees(John, Mary)
  - E.g. $\forall x\ \exists y\ s.t.\ loves(\ x, y\ )$

Changes in Representation

- Propositional logic
  - E.g. P = John sees Mary.
  - E.g. If P is true then Q is also true
- Predicate logic
  - E.g. sees(John, Mary)
  - E.g. $\forall x\ \exists y\ s.t.\ loves(\ x, y\ )$
- Bayesian inference: Reasoning under uncertainty
  - E.g. Pr(JohnSeesMary) = p_1
  - E.g. Pr(JohnSeesMary $\wedge$ MarySeesJohn) = p_2

Probabilistic Inference

- Formally:
  - Given a prior distribution *Pr* over some variables (represents degrees of belief over variables)
  - Given new evidence $E = e$ for some variable E
  - Revise your degrees of belief to get the posterior distribution, Pr_e

Probabilistic Inference

- Formally:
  - Given a prior distribution *Pr* over some variables (represents degrees of belief over variables)
  - Given new evidence $E = e$ for some variable E
  - Revise your degrees of belief to get the posterior distribution, Pr_e

- Intuition:
  - How do your degrees of belief change as a result of learning $E = e$?
    (or more generally, $\mathbf{E} = \mathbf{e}$, for set $\mathbf{E}$)

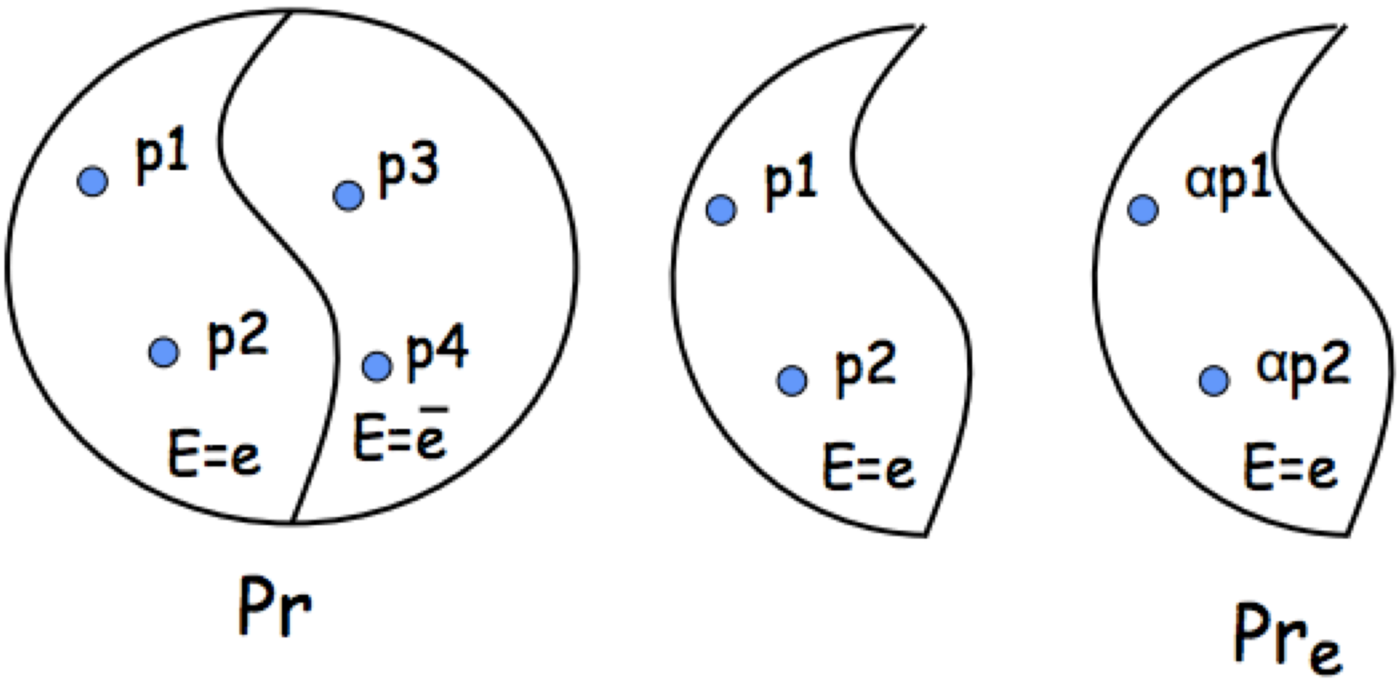
Conditioning

- We define:

$$Pr_e(a) = Pr(a|e)$$

- That is, we produce Pr_e by conditioning the prior distribution on the observed evidence e

Semantics of Conditioning



$\alpha = 1/(p1+p2)$
normalizing constant

Computational Bottleneck

- How do we specify the full joint distribution over a set of RVs $X_1, \ldots, X_n$?
- Inference in this representation is frightfully slow

Computational Bottleneck

- How do we specify the full joint distribution over a set of RVs $X_1, \ldots, X_n$?
  - Exponential number of possible worlds
  - These numbers are not robust/stable
  - These numbers are not natural to assess

- Inference in this representation is frightfully slow

Computational Bottleneck

- How do we specify the full joint distribution over a set of RVs $X_1, \ldots, X_n$?

- Inference in this representation is frightfully slow
  - Must sum over exponential number of worlds to answer query *Pr(a)* or to condition on evidence *e* to determine $Pr_e(a)$

Recall Headache Example

sunny	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

$$\Pr(headache) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

$$\Pr(headache \wedge cold | sunny) = \Pr(headache \wedge cold \wedge sunny) \,/\, \Pr(sunny)$$
$$= 0.108/(0.108 + 0.012 + 0.016 + 0.064) = 0.54$$

$$\Pr(headache \wedge cold | {\sim}sunny) = \Pr(headache \wedge cold \wedge {\sim}sunny) \,/\, \Pr({\sim}sunny)$$
$$= 0.072/(0.072 + 0.008 + 0.144 + 0.576) = 0.09$$

Practical Solution

- How to avoid these two bottlenecks?
  - No solution in general
  - In practice, we will exploit structure

- Use independence assumptions

Independence

- Two variables *A* and *B* are independent if knowledge of *A* does not the change uncertainty of *B* (and vice versa)
  - $\Pr(A|B) = \Pr(A)$
  - $\Pr(B|A) = \Pr(B)$
  - $\Pr(AB) = \Pr(A)\Pr(B)$

Independence Example

- Consider: Bennett smiles and squint eyes

Image taken from iemoji.com

- If Pr(Smile|Squint) = Pr(Smile)
  - Chance of him smiling when he squints
  - Chance of him smiling in anyway

- And Pr(Squint|Smile) = Pr(Squint)
  - Chance of him squinting when he smiles
  - Chance of him squinting no matter what else he's doing

- Then Smile and Squint are independent

What does Independence Buy Us?

- Product rule changes:

$\Pr(ab) = \Pr(a|b)\Pr(b)$

$\Pr(ab) = \Pr(a)\Pr(b)$

- Chain rule changes:

$\Pr(abcd) = \Pr(a|bcd)\Pr(b|cd)\Pr(c|d)\Pr(d)$

$\Pr(abcd) = \Pr(a)\Pr(b)\Pr(c)\Pr(d)$

Conditional Independence

- To loosen the independence assumption, we can use conditional independence

- Two variables *A* and *B* are conditionally independent given *C* if:
  - $\Pr(a|b,c) = \Pr(a|c) \qquad \forall\ a, b, c$

- Knowing the value of B does not change the prediction of A given the presence of C

Conditional Independence Example

Image taken from pinterest.com

- Consider: Want tea, pink cup, and rainy
- If Pr(Tea|Pink,Rainy) = Pr(Tea|Rainy)
  - Chance of wanting tea on rainy days in pink cup is the same as chance of wanting tea on rainy days in any cup
- And Pr(Tea|Pink,~Rainy) = Pr(Tea|~Rainy)
  And Pr(Tea|~Pink,Rainy) = Pr(Tea|Rainy)
  And Pr(Tea|~Pink,~Rainy) = Pr(Tea|~Rainy)
  And Pr(~Tea|Pink,Rainy) = Pr(~Tea|Rainy)
  And …
  - Check equivalence for all other combinations
- Then Tea is independent of Pink given Rainy

Formal Definitions

- *x* and *y* are independent iff:
  $\Pr(x) = \Pr(x|y) \Leftrightarrow \Pr(y) = \Pr(y|x) \Leftrightarrow \Pr(xy) = \Pr(x)\Pr(y)$
  - Intuitively, learning *y* doesn't influence beliefs about *x*

- *x* and *y* are conditionally independent given *z* iff:
  $\Pr(x|z) = \Pr(x|yz) \Leftrightarrow \Pr(y|z) = \Pr(y|xz) \Leftrightarrow$
  $\Pr(xy|z) = \Pr(x|z)\Pr(y|z) \Leftrightarrow \ldots$
  - Intuitively, learning *y* doesn't influence beliefs about *x* if you already know *z*

What Good is Independence?

- Given (say, Boolean) variables $X_1,\ldots,X_n$ are mutually independent

- How to specify the full joint distribution $Pr(X_1,\ldots,X_n)$?

What Good is Independence?

- Given (say, Boolean) variables $X_1,\ldots,X_n$ are mutually independent

- How to specify the full joint distribution $Pr(X_1,\ldots,X_n)$?
  - Joint is simplified as: $\prod_{i=1}^{n} \Pr(Xi)$
  - Can specify the full joint using only n parameters (linear) instead of $2^n - 1$ (exponential)

Example

- Given 4 mut. Indep. Boolean RVs: X_1, X_2, X_3, X_4

  $Pr(x_1) = 0.4$, $Pr(x_2) = 0.2$, $Pr(x_3) = 0.5$, $Pr(x_4) = 0.8$

- $Pr(x_1, \sim x_2, x_3, x_4) = ?$

- $Pr(x_1, x_2, x_3 | x_4) = ?$

The Value of Independence

- Complete independence reduces both representation of joint distribution and inference from $O(2^n)$ to $O(n)$
- Unfortunately, complete independence is very rare
  - Most realistic domains don't exhibit this property
- Fortunately, most domains exhibit a fair amount of conditional independence
  - Can exploit conditional independence for representation and inference too
  - Bayesian networks do just this

An Aside on Notation

- *Pr(X)* for variable *X* (or set of variables) refers to the (marginal) distribution over *X*
  - Distinguish from *Pr(x)* or *Pr(~x)* (or *Pr*(x_i) for non-Boolean vars) which are numbers
  - Think of Pr(X) as a function that accepts any $x_i \ni Dom(X)$ as an argument and returns Pr(x_i)

An Aside on Notation

- *Pr(X)* for variable *X* (or set of variables) refers to the (marginal) distribution over *X*
  - Distinguish from *Pr(x)* or *Pr(~x)* (or *Pr(x_i)* for non-Boolean vars) which are numbers
  - Think of Pr(X) as a function that accepts any $x_i \ni Dom(X)$ as an argument and returns Pr(x_i)

- *Pr(X|Y)* refers to family of conditional distributions over *X*, one for each $y \ni Dom(Y)$
  - Think of *Pr(X|Y)* as a function that accepts any x_i and y_k and returns *Pr(x_i|y_k)*

Think truth tables

Exploiting Conditional Independence

- Consider the following story:
  - If Bowen woke up too early (E), she needs caffeine (C)
  - If Bowen needs caffeine, she's likely to be grumpy (G)
  - If she is grumpy, then her lecture won't be as good (L)
  - If lecture doesn't go smoothly, then students will be disappointed (S)

E = Woke up too early
G = gets grumpy
S = students disappointed
C = need caffeine
L = lecture not smooth

Conditional Independence

- If you learned any of E,C,G,L, would your assessment of Pr(S) change?

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- If you learned any of E,C,G,L, would your assessment of Pr(S) change?
  - If any of E,C,G,L are true, you would increase Pr(s) and decrease Pr(~s)
  - Therefore, S is not independent of E,C,G,L

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- If you knew the value of L (true or false), would learning the value of E,C, or G influence your assessment of Pr(S)?

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- If you knew the value of L (true or false), would learning the value of E,C, or G influence your assessment of Pr(S)?
  - Influence that E,C,G has on S is mediated by L
  - E.g. Students aren't disappointed because Bowen is grumpy, it's because the lecture wasn't smooth
  - So S is independent of E,C,G, <u>given</u> L

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- We have: S is independent of E,C,G, <u>given</u> L
- Similarly:
  - L is independent of E,C <u>given</u> G
  - G is independent of E <u>given</u> C
- This translates to:
  - Pr(S|L,G,C,E) = Pr(S|L)
  - Pr(L|G,C,E) = Pr(L|G)
  - Pr(G|C,E) = Pr(G|C)
  - Pr(C|E) % doesn't simplify further
  - Pr(E) % doesn't simplify further

E = Woke up too early

C = need caffeine

G = gets grumpy

L = lecture not smooth

S = students disappointed

Conditional Independence

- Specifying the full joint distribution Pr(S,L,G,C,E)?

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- Specifying the full joint distribution Pr(S,L,G,C,E)?

- By the chain rule:
  $\Pr(S,L,G,C,E) =$
  $\Pr(S|L,G,C,E)\Pr(L|G,C,E)\Pr(G|C,E)\Pr(C|E)\Pr(E)$

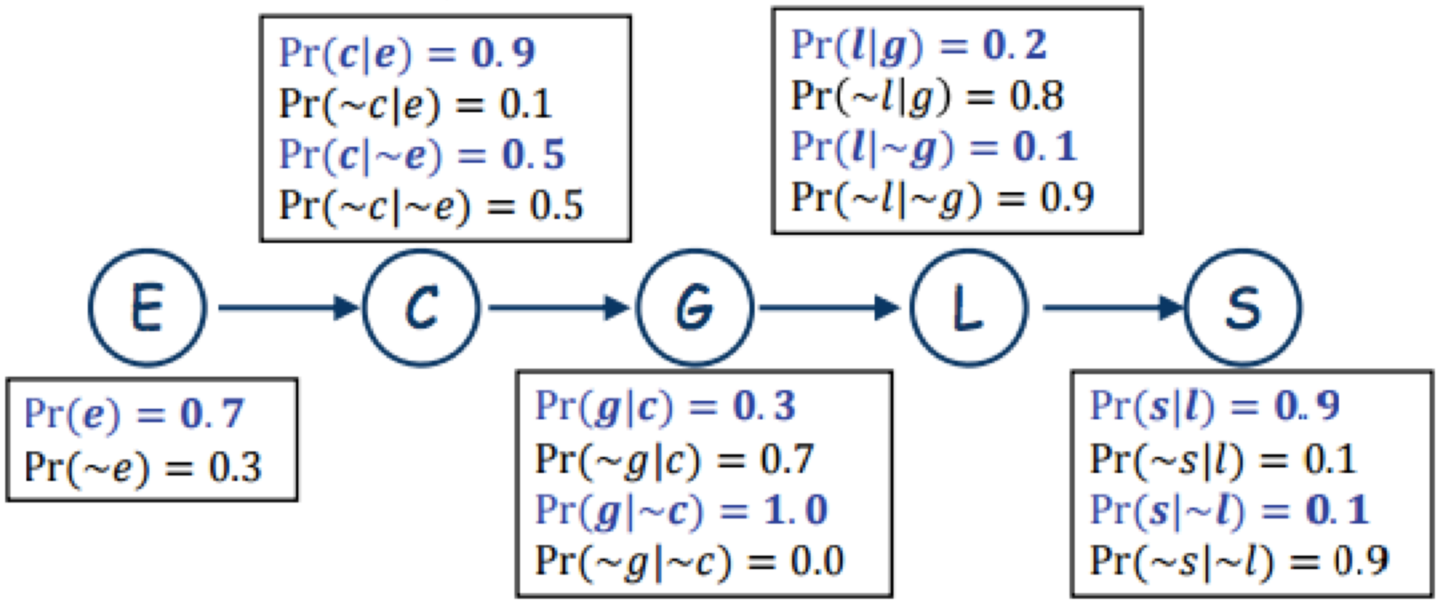
E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Conditional Independence

- Specifying the full joint distribution Pr(S,L,G,C,E)?
- By the chain rule:
  $\Pr(S,L,G,C,E) =$
  $\Pr(S|L,G,C,E)\Pr(L|G,C,E)\Pr(G|C,E)\Pr(C|E)\Pr(E)$
- By our independence assumptions:
  $\Pr(S,L,G,C,E) = \Pr(S|L)\Pr(L|G)\Pr(G|C)\Pr(C|E)\Pr(E)$
- The full joint is specified by 5 local conditional distributions!

E = Woke up too early
G = gets grumpy
S = students disappointed
C = need caffeine
L = lecture not smooth

Example Quantification



- Specifying the joint requires only 9 parameters!
  - Instead of 31 (= $2^5 - 1$) for explicit representation

Inference is Easy

- How to compute Pr(g)?

E = Woke up too early

C = need caffeine

G = gets grumpy

L = lecture not smooth

S = students disappointed

Inference is Easy

- How to compute Pr(g)?
  - Apply the sum-out rule
- $$P(g) = \sum_{c_i \in Dom(C)} \Pr(g \mid c_i)\Pr(c_i)$$

$$= \sum_{c_i \in Dom(C)} \Pr(g \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)$$

Terms available in our local distributions!

E = Woke up too early
C = need caffeine
G = gets grumpy
L = lecture not smooth
S = students disappointed

Inference is Easy

- Concrete example to compute Pr(g):

E = Woke up too early

C = need caffeine

G = gets grumpy

L = lecture not smooth

S = students disappointed

Modeling Example

- Suppose you have a simple world with 3 variables: weather, sprinkler, and grass condition
  - If it’s rainy, the grass is wet.
  - If the sprinkler is on, the grass is wet.
  - If it’s cloudy, the sprinkler should be off.

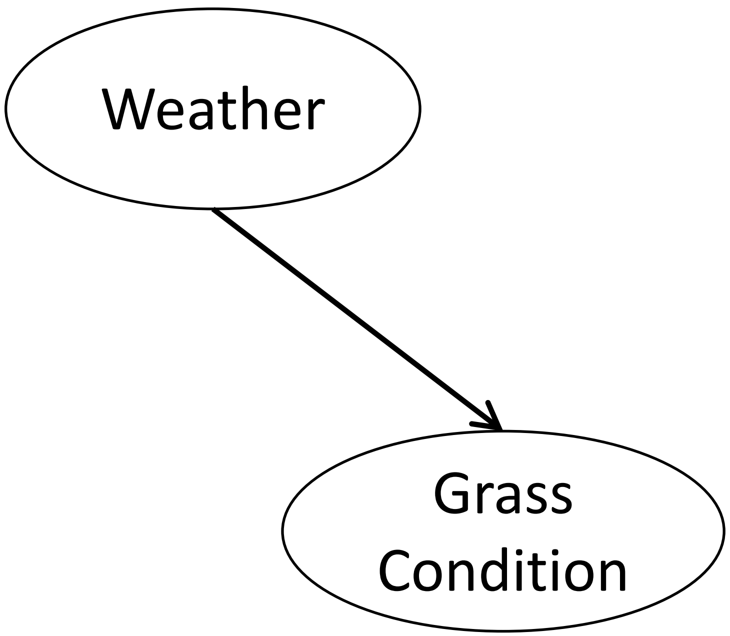
- How to model these interactions?

Modeling with a Bayes Net

Weather = {sunny, cloudy, rainy}
Sprinkler = {on, off}
GrassCondition = {wet, dry}

Modeling with a Bayes Net

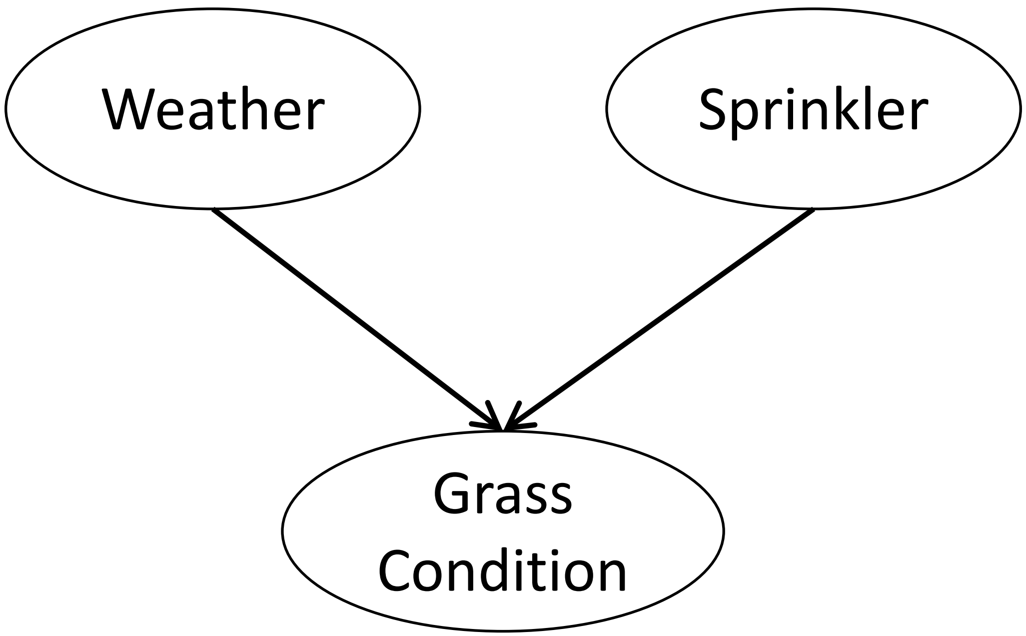
If it's rainy, the grass is wet.



Weather = {sunny, cloudy, rainy}
Sprinkler = {on, off}
GrassCondition = {wet, dry}

Modeling with a Bayes Net

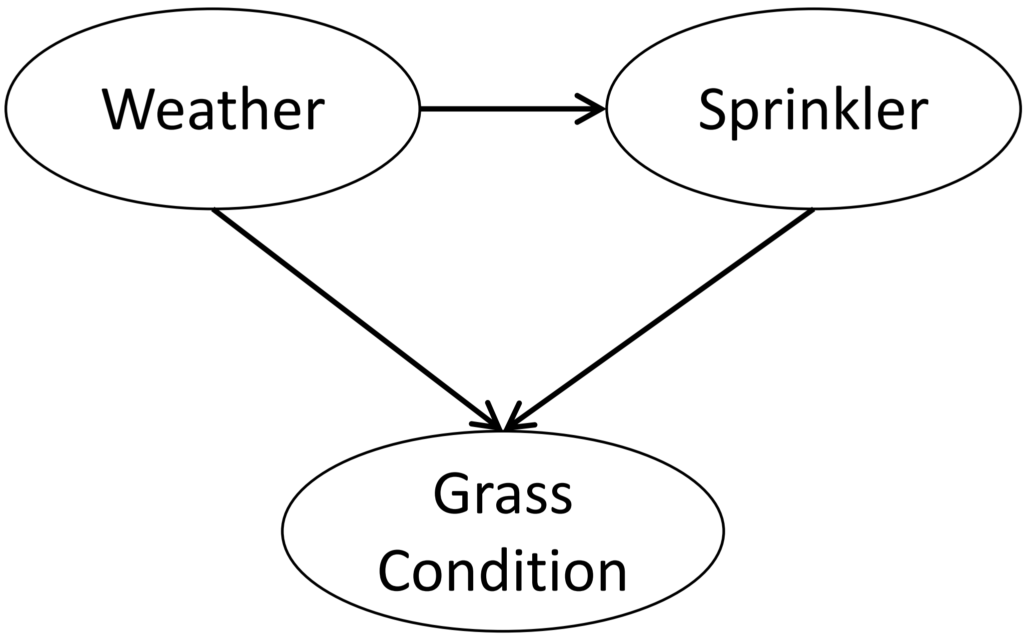
If the sprinkler is on, the grass is wet.



Weather = {sunny, cloudy, rainy}
Sprinkler = {on, off}
GrassCondition = {wet, dry}

Modeling with a Bayes Net

If it's cloudy, the sprinkler should be off.

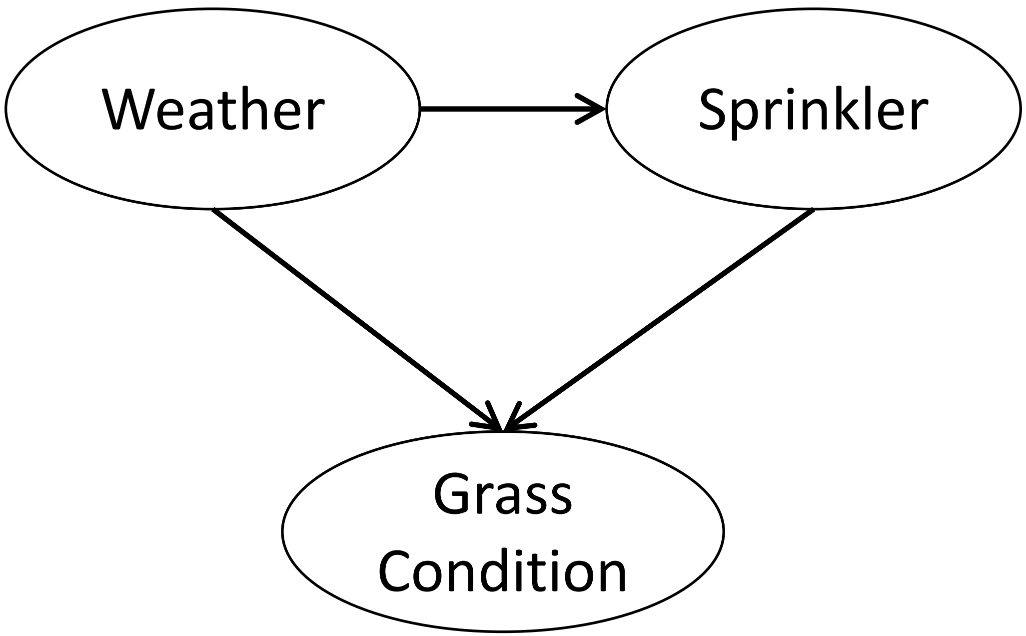


Weather = {sunny, cloudy, rainy}

Sprinkler = {on, off}

GrassCondition = {wet, dry}

Most Popular Bayes Net Example



Weather = {sunny, cloudy, rainy}
Sprinkler = {on, off}
GrassCondition = {wet, dry}

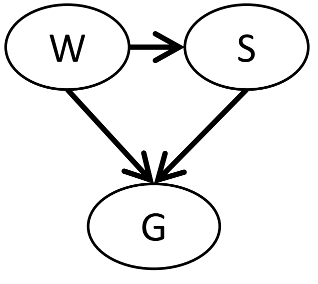
What is a Bayes Net (BN)

- Also called Bayesian network, belief network
- A graphical representation of the direct dependencies over a set of variables
- Directed dependencies express the causality between the variables
- Each variable has an associated conditional probability tables (CPTs) quantifying the strength of those influences

BN Definition

- A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:
  - A directed acyclic graph whose nodes are variables
  - A set of CPTs $Pr(X_i|Parents(X_i))$ for each X_i

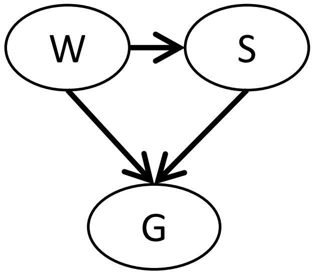
Pr(W=sunny)	Pr(W=cloudy)	Pr(W=rainy)
0.6	0.3	0.1



BN Definition

- A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:
  - A directed acyclic graph whose nodes are variables
  - A set of CPTs $Pr(X_i | Parents(X_i))$ for each X_i

Pr(W=sunny)	Pr(W=cloudy)	Pr(W=rainy)
0.6	0.3	0.1

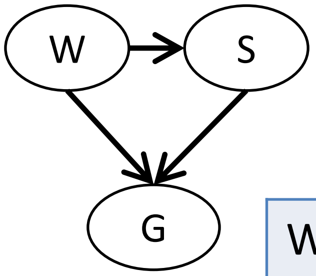


	Pr(S=on|W)	Pr(S=off|W)
W=sunny	0.1	0.9
W=cloudy	0.8	0.2
W=rainy	0.001	0.999

BN Definition

- A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:
  - A directed acyclic graph whose nodes are variables
  - A set of CPTs $Pr(X_i|Parents(X_i))$ for each X_i

Pr(W=sunny)	Pr(W=cloudy)	Pr(W=rainy)
0.6	0.3	0.1

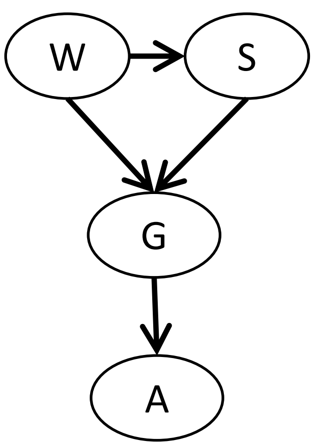


	Pr(S=on|W)	Pr(S=off|W)
W=sunny	0.1	0.9
W=cloudy	0.8	0.2
W=rainy	0.001	0.999

		Pr(G=wet|W,S)	Pr(G=dry|W,S)
W=sunny	S=on	0.9	0.1
W=sunny	S=off	0.001	0.999
W=cloudy	S=on	0.99	0.01
W=cloudy	S=off	0.2	0.8
W=rainy	S=on	1	0
W=rainy	S=off	0.9	0.1

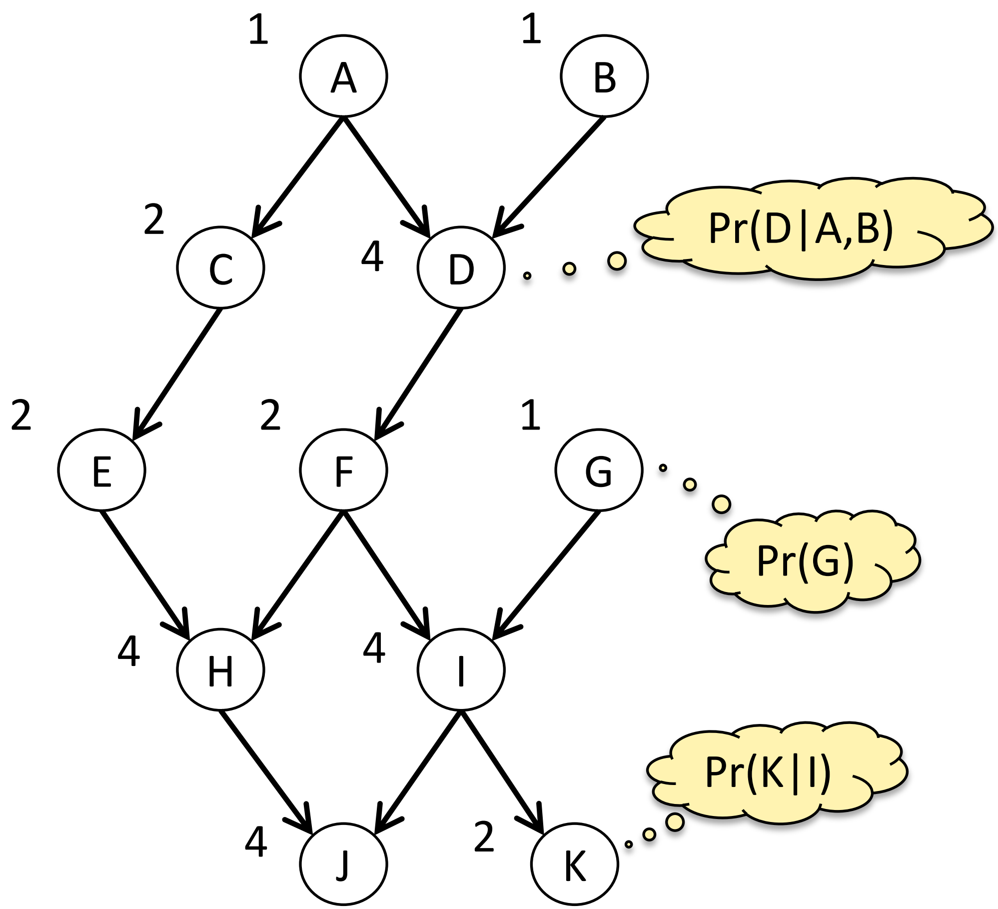
Key Terminology

- Parents of a node: *Parents(X_i)*
- Children of a node
- Descendants of a node
- Ancestors of a node
- Family: set of nodes consisting of Xi and its parents
  - CPTs are defined over families in the BN



Parents(W) = ?
Children(S) = ?
Descendants(W) = ?
Ancestors(A) = ?
Family(G) = ?

An Example Bayes Net



- A few CPTs "shown"
- Explicit joint requires $2^{11} - 1 = 2047$ params
- BN requires only 27 params (the number of entries for each CPT is written)

Semantics of Bayes Nets

- The structure of the BN means:
  - Every X_i is conditionally independent of all its non-descendants given its parents
  - Intuition: your parents is the only ones who has influence on you

- Formally:

$$\boxed{\begin{array}{c} Pr(X_i \mid S \cup Par(X_i)) = Pr(X_i \mid Par(X_i)) \\ \text{for any subset } S \subseteq NonDescendants(X_i) \end{array}}$$

Semantics of Bayes Nets

- If we ask for *$Pr(x_1,\ldots,x_n)$*
  - Assuming an ordering consistent with the network

- By the chain rule, we have:
  $Pr(x_1,\ldots,x_n)$
  $= Pr(x_n | x_{n-1},\ldots,x_1)Pr(x_{n-1} | x_{n-2},\ldots,x_1)\ldots Pr(x_1)$
  $= Pr(x_n | Par(x_n))Pr(x_{n-1} | Par(x_{n-1}))\ldots Pr(x_1)$

- Thus, the joint is recoverable using the parameters (CPTs) specified in an arbitrary BN

Key Ideas

- Main concept
  - Computational bottlenecks in computing joint probability distributions appear in representation and inference
  - Exploit independence and conditional independence
  - Computation is linear rather than exponential
- Representation:
  - Bayes net is a directed acyclic graph whose nodes are random variables with associated CPTs
  - Expresses the joint probability distribution using the product of local distributions, i.e. $Pr(X_i | Par(X_i))$