Advanced Predictive Modeling

Dr. Bowen Hui
Computer Science
University of British Columbia Okanagan

Introduction

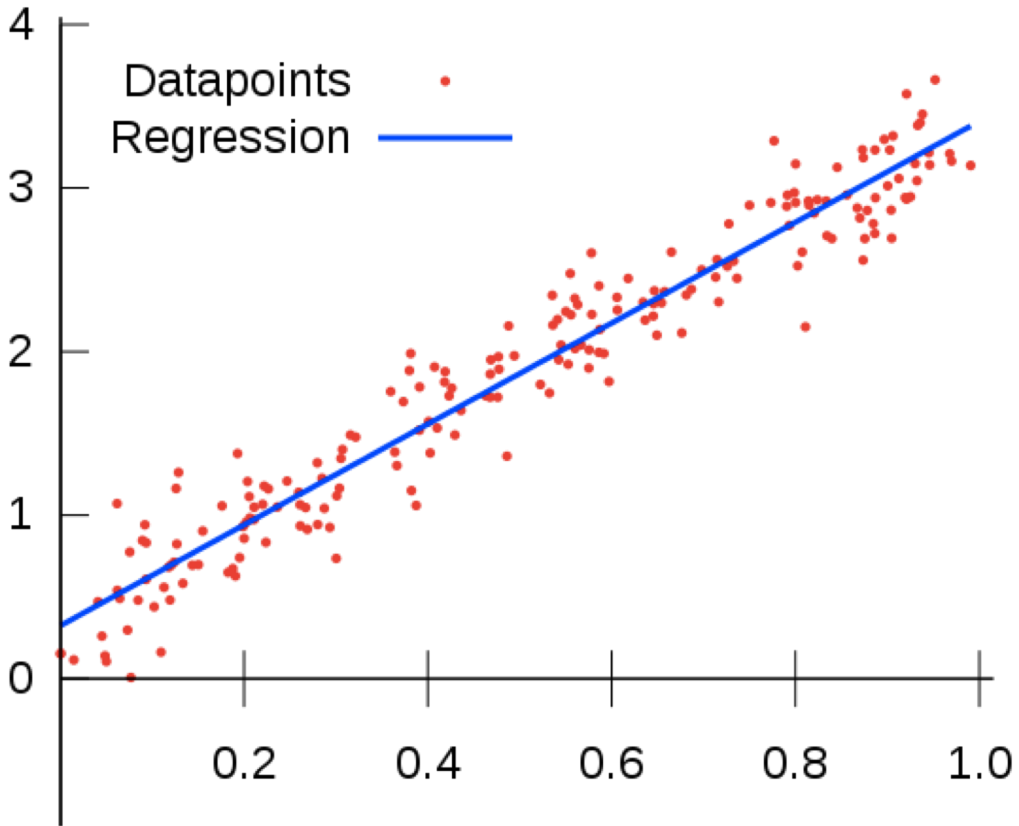
- Computer science instructor @ UBCO since 2012
- PhD in probabilistic modeling (machine learning), applied to human-computer interaction, tested with real datasets
  - Partially observable Markov decision processes
  - Closely related to Bayesian networks
- Research interests: learning analytics, educational games, language modeling

Plan for Upcoming Lectures

- Lecture 1 (this class):
  - Regression -> classification -> Bayesian networks
- Lecture 2:
  - Rational decision making
- Lecture 3:
  - Sequential decision making

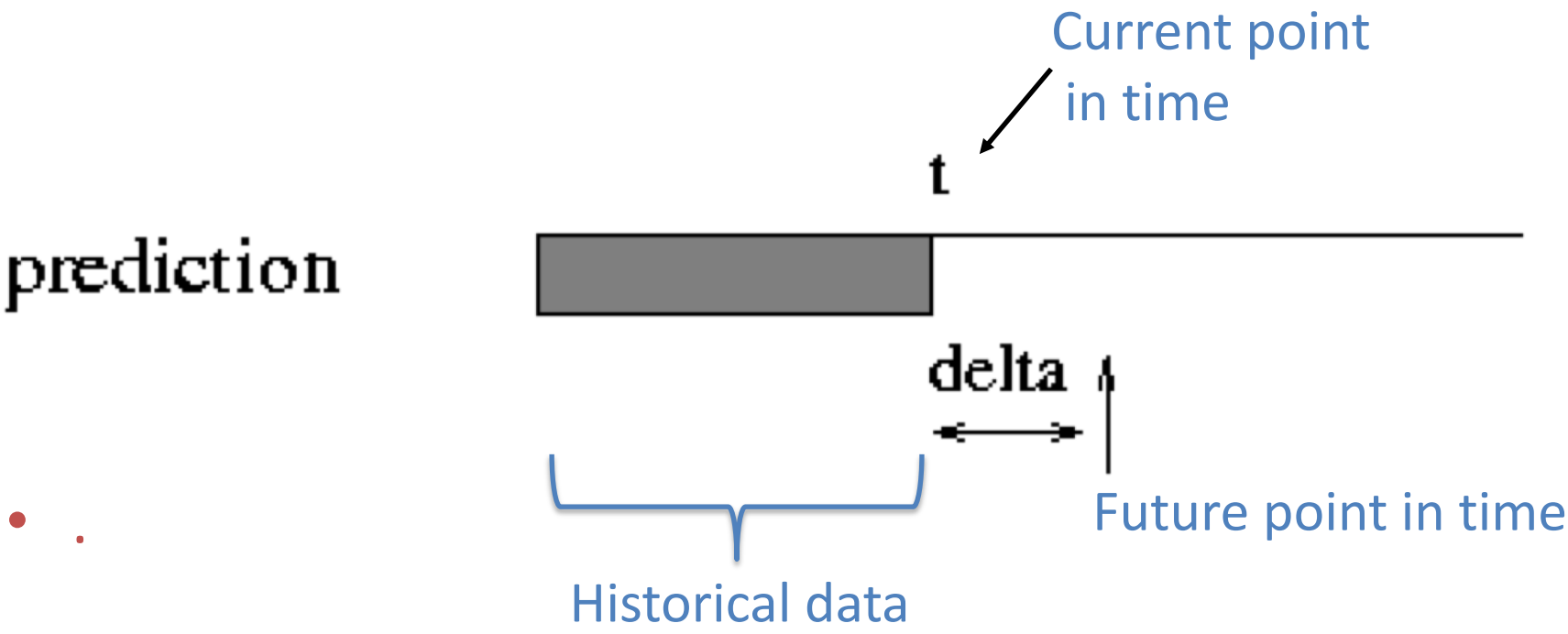
A Simple View of Regression

- Statistical processes for estimating relationships among variables
- Quantify the relative impact of predictor/independent variables on an outcome/dependent variable
- Simple example:



Power of Regression

- Helps understand relationships between data points
- Task of interest: Data driven method for forecasting (or prediction)

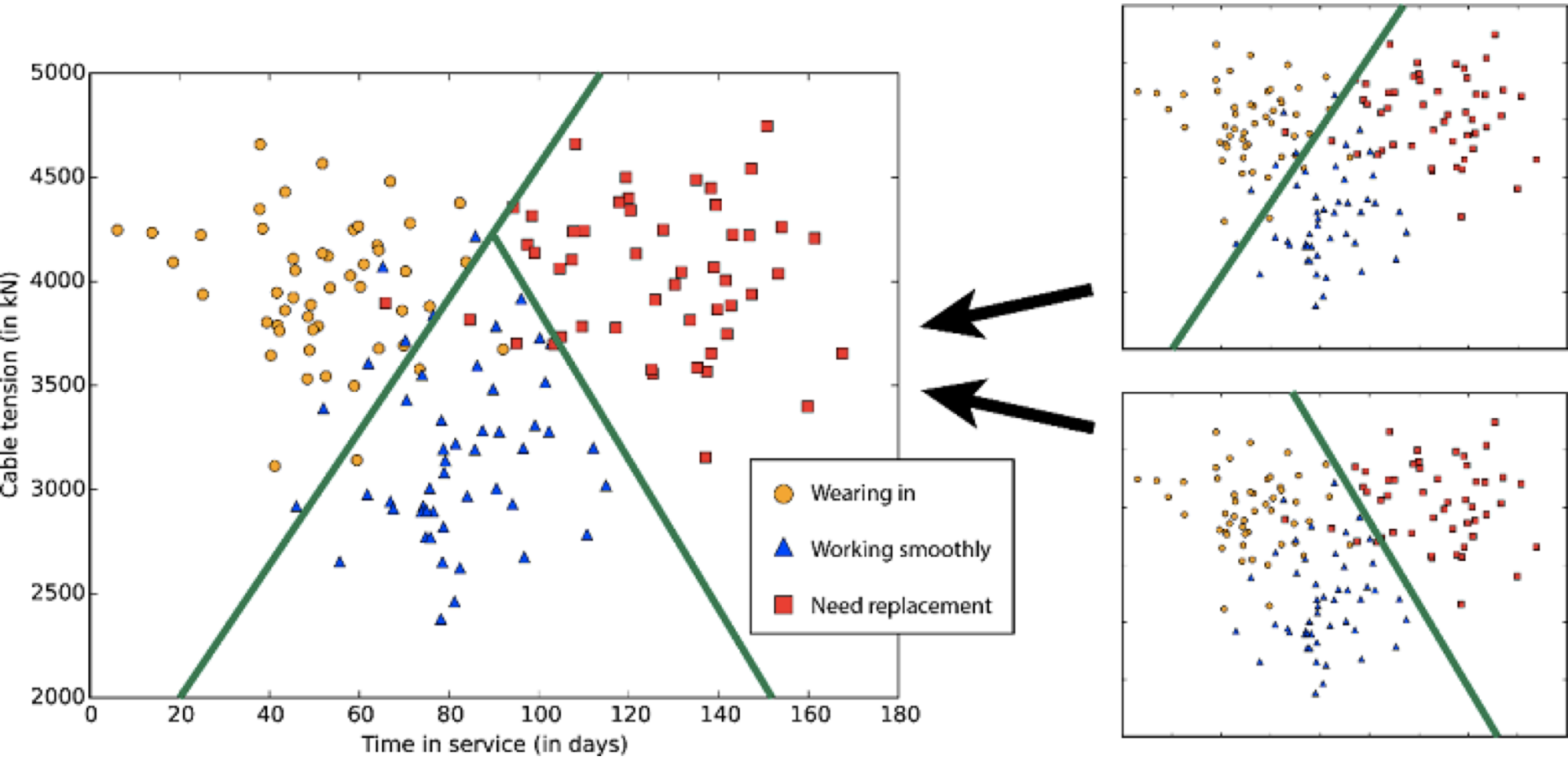


- .

Beyond Regression

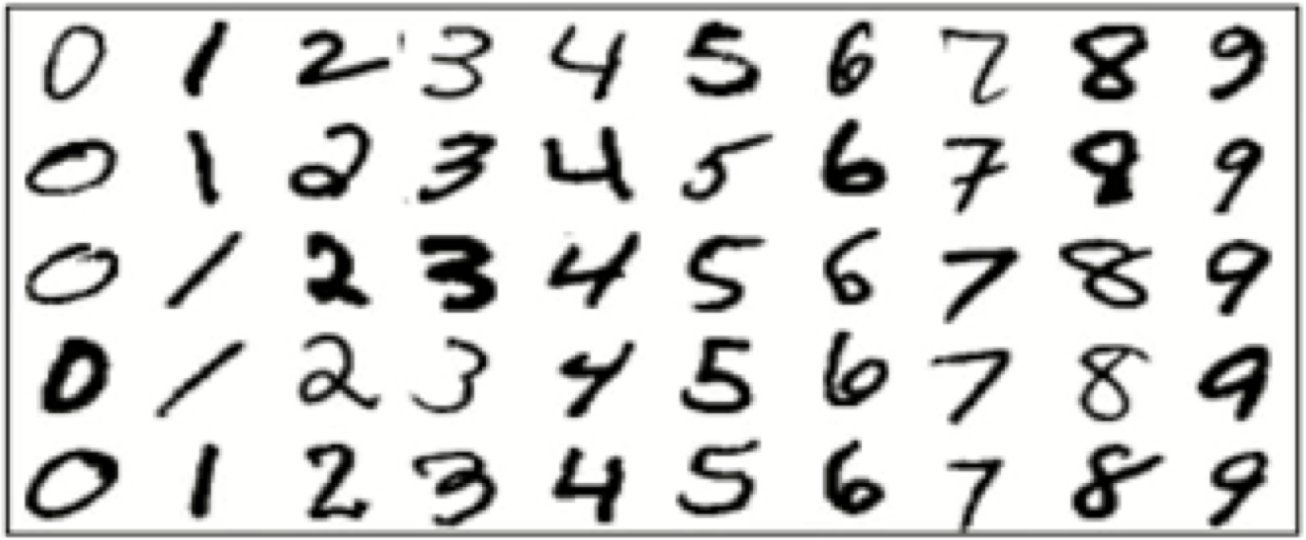
- Regression uses training data to predict an outcome value
  - Outcome variables are usually continuous variable
- Classification uses training data to group an outcome into a class
  - Class variables are usually categorical and unordered
- Later: decision making tasks

A Visual Example of Classification



Example Classification Tasks

- Predict handwritten digit (0,..,9)



Example Classification Tasks

- Predict handwritten digit (0,..,9)
- Classify credit card transaction as fraudulent or not
- Predict if an email is spam or not
- Classify webpages into topics
- Classify students activities/performance into letter grades

Classification Process

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

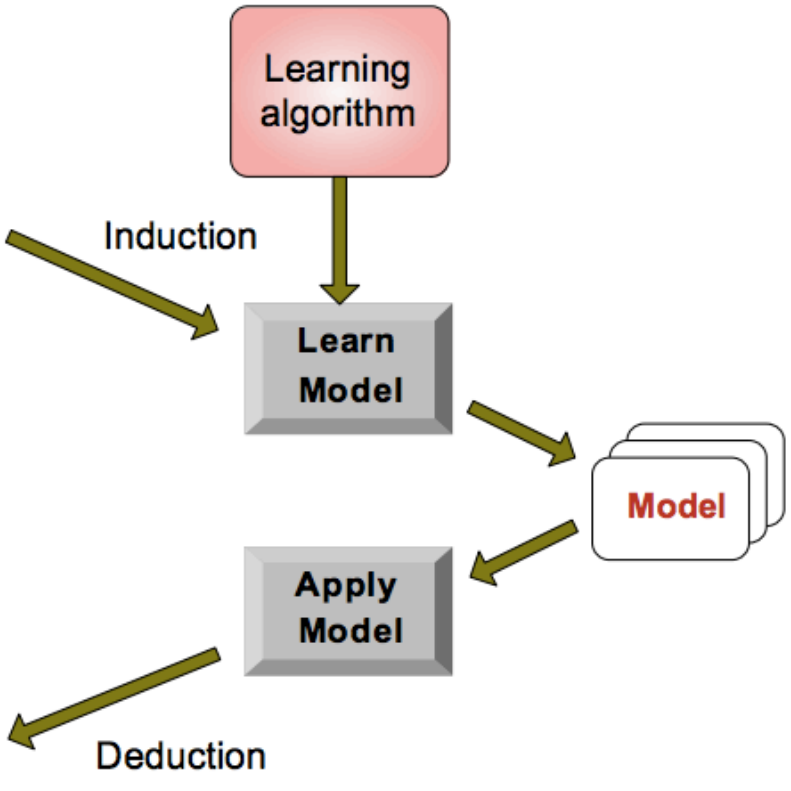


Image taken from https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf

Aside: Machine Learning Terminology

- Categorization just means putting things into different groups somehow

- Classification
  - Supervised learning task
  - All data requires labels of what the "right" answer is
  - Based on labeled dataset, learn the model underlying the data and predict labels for unseen data

- Clustering
  - Unsupervised learning task
  - Data does not have labels
  - Based on unlabeled dataset, discover the model underlying the data to find labels for each group

Different Types of Classifiers

- A classifier is a model that performs a classification task
  - Not necessarily deterministic
  - What is the probability that X belongs in class A, Pr(X=A)? What about Pr(X=B), Pr(X=C), etc.?

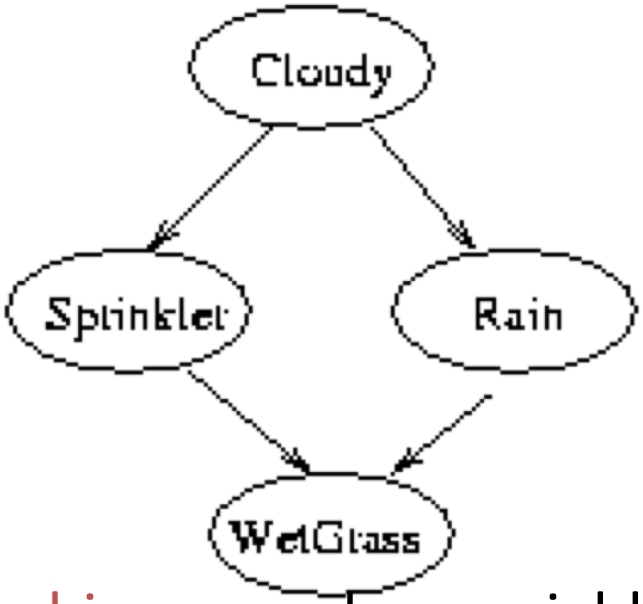
- Popular classifiers:
  - Naïve Bayes (basic but powerful)
  - Logistic regression
  - Decision tree
  - k-Nearest neighbour
  - Support vector machine
  - Neural network

Bayesian Network (BN)

- Graphical models provide visual explanatory power in modeling real world problems
- BN is a directed acyclic graph that represents causality among a set of random variables
  - Nodes: random variables, each with an associated probability function
  - Edges: represents conditional dependencies between nodes
- Specifies a complete joint probability distribution over all the variables in the model
- Answers possible inference queries by marginalization

A Simple Example of BN

- What makes the grass wet?

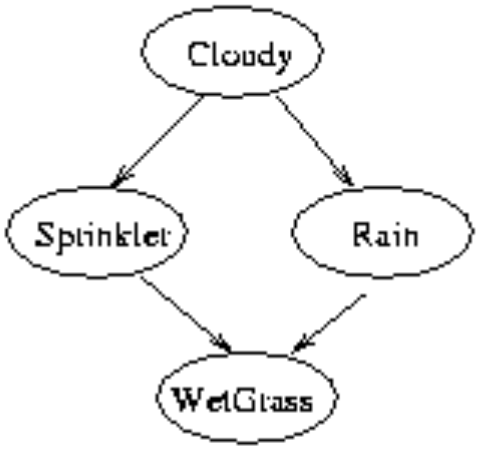


- All the variables are binary random variables (true/false)
- Model:
  - "grass is wet" has two possible causes
  - Whether it's cloudy or not influences if we turn on the sprinkler or if it rains

A Simple Example of BN (cont.)

Pr(C):

P(C=F)	P(C=T)
0.5	0.5



Pr(S|C):

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

Pr(R|C):

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

Each variable has an associated conditional probability distribution (CPD), or table (CPT)

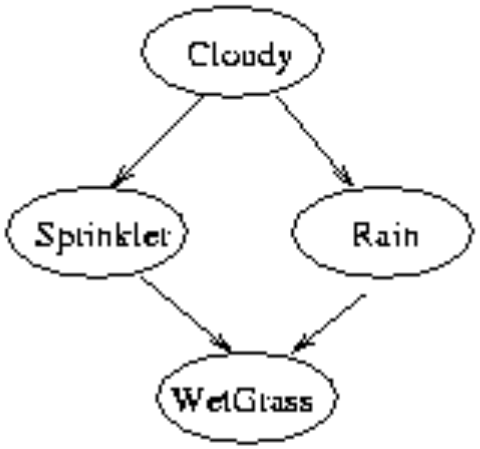
Pr(W|S,R):

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Complete joint distr. specified by this BN is: Pr(C,S,R,W)

A Simple Example of BN (cont.)

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

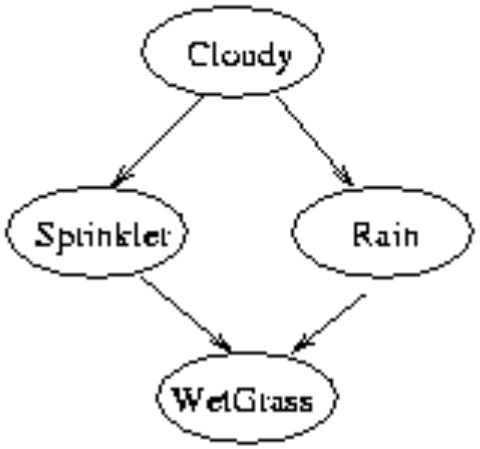
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

Grass is wet (W=true) occurs either when sprinkler is on(S=true) or it's raining (R=true)

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

A Simple Example of BN (cont.)

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

CPTs define the strength of relationships

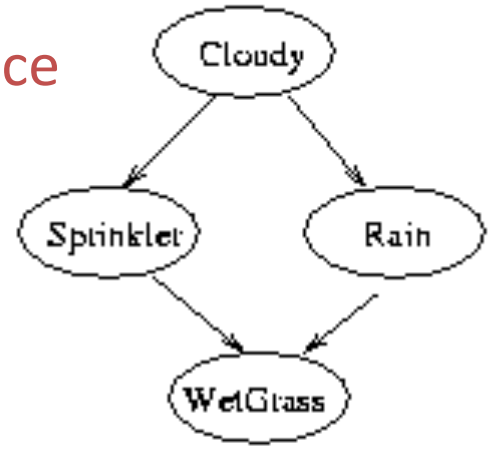
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

e.g. Pr(W=T | S=T, R=T) = 0.99

A Simple Example of BN (cont.)

P(C=F)	P(C=T)
0.5	0.5

BNs model exploit
conditional independence



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

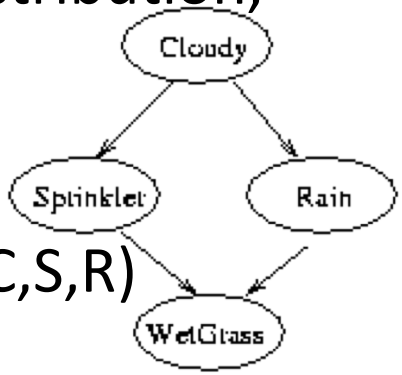
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

e.g. W is independent of C, given S and R

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

A Simple Example of BN (cont.)

- This model is specified by the joint probability distribution, Pr(C,S,R,W)



- By the chain rule of probability, we can write:
  Pr(C,S,R,W) = Pr(C) * Pr(S|C) * Pr(R|C,S) * Pr(W|C,S,R)
- Using conditional independence relationships in the model, we can further simplify:
  Pr(C,S,R,W) = Pr(C) * Pr(S|C) * Pr(R|C) * Pr(W|S,R)
- Simpler model means fewer parameters means easier to learn

Need for Probability

- Assumes knowledge of statistical inference
- Our focus:
  - How to model the real world problem
  - How to compute probability estimations
  - Apply probabilistic inference algorithms for complex models
- Shift to machine learning/AI terminology
  - Bring in stats as needed

AI Modeling

- At current time, predict likelihood of a future event
  - Inference task: Estimate the probability that an event will occur, given evidence of set of past observations
  - What do we know already?
  - What influences an event occurrence?
  - How to model our knowledge of the world?

Modeling the State in a Chess Game

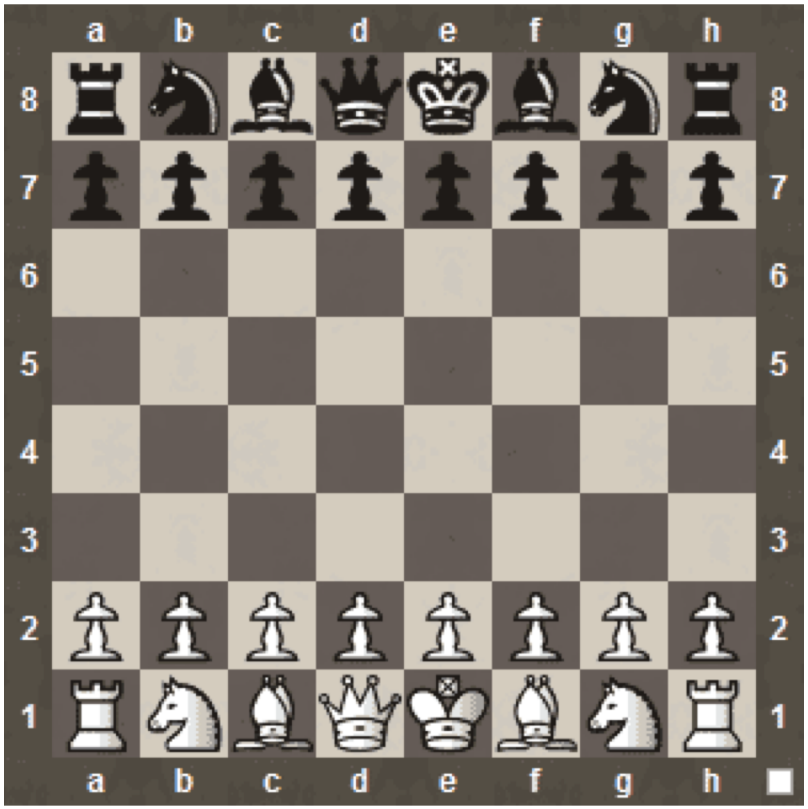


Image taken from computerchessonline.net

State of game: black horse at b7, etc.

Modeling Tic Tac Toe State

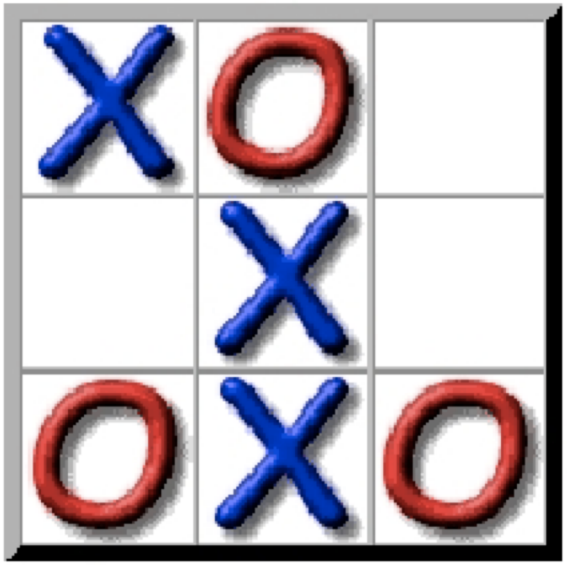
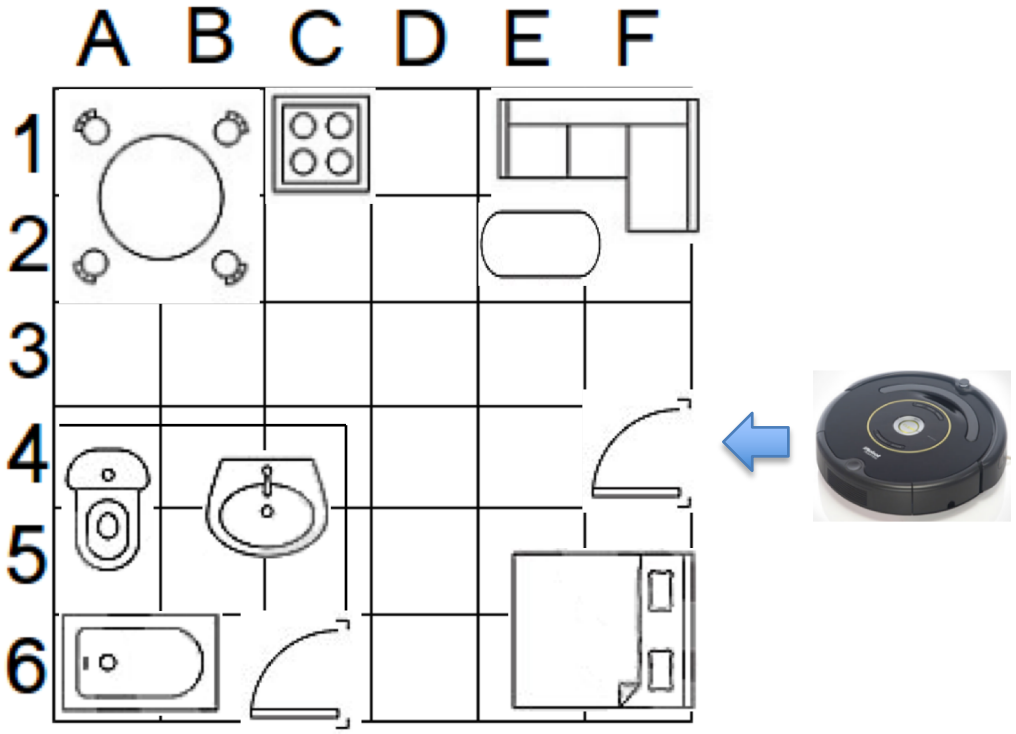


Image taken from www.mathematik.uni-ulm.de

State of game: x at {a1, b2, b3}, o at {a3, b1, c3}

Modeling an Apartment

Will furniture move?



State of apartment: Stove at C1, coffee table at E2, etc.

State of the World

- In games and simple worlds, we can assume or expect:
  - Everything is fully observable
    (that means, there are no hidden variables)
  - The world has no uncertainty

- Is this always going to be feasible?

Modeling Person's Emotional State

Image taken from www.pinterest.com

State of user: currently 'happy'

… Are you sure?

Beliefs over State

- Uncertainty arises when:
  - You can't observe the value of a state
  - When the state changes
- When there's uncertainty:
  - We have beliefs of the world
  - Need to quantify level of uncertainty
- Then we can make decisions properly
- Use probability theory to model our beliefs

Example

- Simple world:
  stove is on or off,
  time is 9:00am, 9:01am, …

- Variables:
  Stove, Time

- Sample state:
  Stove=on and Time=9:00am

- Sample estimations:
  Pr(Stove=on)
  Pr(Stove=on|Time=noon)

Random Variables

- Assume set **V** of random variables: X, Y, etc.
  - Each RV X has a domain of values Dom(X)
  - X can take on any value from Dom(X)
  - Assume **V** and Dom(X) are finite

- Examples:
  - Dom(X) = $\{x_1, x_2, x_3\}$
  - Dom(Weather) = {sunny, cloudy, rainy}
  - Dom(IamHappy) = {true, false}

Modeling Example

- Student asks you (TA) a programming question
- You consider how to answer the question
  - What are RVs of the student?
  - What are RVs about you?

Which variables are observable, which are hidden?

State

- A formula is a logical combination of variable assignments
  - E.g. $(X = x_2 \vee X = x_3) \wedge Y = y_2$

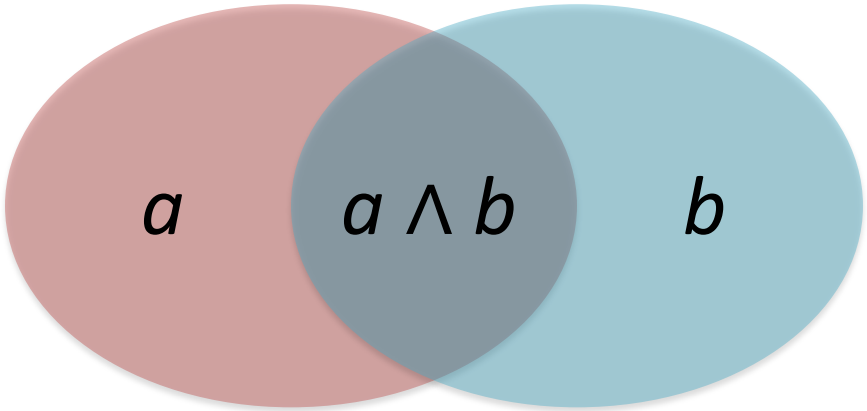
- A state is an assignment of values to each variable
  - One state represents one possible world
  - The set of states denote the set of possible worlds
  - Note: Think truth tables for discrete RVs

Modeling Example cont.

- Student asks you (TA) a programming question
- You consider how to answer the question
  - What are RVs of the student?
  - What are RVs about you?
- Draw out the set of states in truth table format

Probability Distributions

- A probability distribution Pr: $\mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq \Pr(a) \leq 1$
  - *Pr(a) = Pr(b)* if *a* is logically equivalent to *b*
  - *Pr(a) = 1* if *a* is a tautology
  - $\Pr(a \vee b) = \Pr(a) + \Pr(b) - \Pr(a \wedge b)$



Probability Distributions

- A probability distribution Pr: $\mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq \Pr(a) \leq 1$
  - *Pr(a) = Pr(b)* if *a* is logically equivalent to *b*
  - *Pr(a) = 1* if *a* is a tautology
  - $\Pr(a \vee b) = \Pr(a) + \Pr(b) - \Pr(a \wedge b)$

Caution: Probability axioms not always followed in theories that use probability!

Probability Distributions

- A probability distribution Pr: $\mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq \Pr(a) \leq 1$
  - *Pr(a) = Pr(b)* if *a* is logically equivalent to *b*
  - *Pr(a) = 1* if *a* is a tautology
  - $\Pr(a \vee b) = \Pr(a) + \Pr(b) - \Pr(a \wedge b)$

- *Pr(a)* denotes our degree of belief in *a*
  - *Pr(a) = 0* if you consider it to be impossible

- The sum of the distribution must be 1.0

Examples

- $\Pr(X = x_1) = 0.9$
  - *Pr(Stove = on) = 0.9*
  - *Pr(Stove = off) = 1 – Pr(Stove = on) = 0.1*
  - *Pr(Time = noon) = 0.001*

Visualizing Probability Distribution

- X-axis: set of possible outcomes
- Y-axis: probability

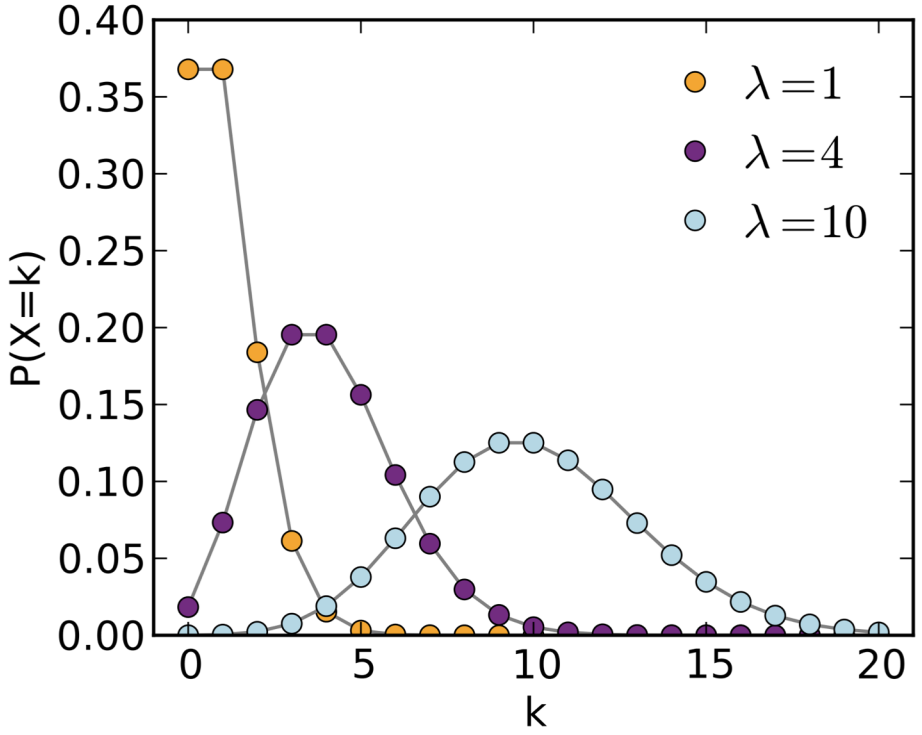


Image taken from wikipedia.org

Commonly used distributions:

- Normal (Gaussian)
- Uniform

Joint Probability Distribution

- Probability distribution
  - Involves one RV to describe state space
  - Probabilities must sum up to 1.0

- Joint probability distribution
  - When the state is described by two or more RVs
  - Specify probabilities for all combinations of events
  - Probabilities must sum up to 1.0

Examples

- *Pr($X = x_1 \wedge Y = y_2$) = 0.6*
  - $\Pr(Stove = on \wedge Time = noon) = 0.6$
  - $\Pr(Stove = on \wedge Time = 12{:}01pm) = 0.2$
  - $\Pr(Stove = on \wedge Time = 12{:}02pm) = 0.1$
  - …

Modeling Example cont.

- Student asks you (TA) a programming question
- You consider how to answer the question
  - What are RVs of the student?
  - What are RVs about you?
- Draw out the set of states in truth table format
- Assign probabilities to each state in the table

The Summing Out Property

- $$\Pr(x) = \sum_{y \in Dom(Y)} \Pr(x \wedge y)$$
- Also called marginalization

- Example:
  Pr(Stove=on) = Pr(Stove = on ΛTime = 9:00am)
  + Pr(Stove = on Λ Time = 9:01am)
  + Pr(Stove = on Λ Time = 9:02am)
  + Pr(Stove = on Λ Time = 9:03am)
  …
  for all values in Dom(Y)

The Inference Task

- General structure:
  - You have general knowledge about the world
  - You observe an event (or series of events)
  - You want to estimate the probability of an event (or several events) that you cannot observe

Example

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

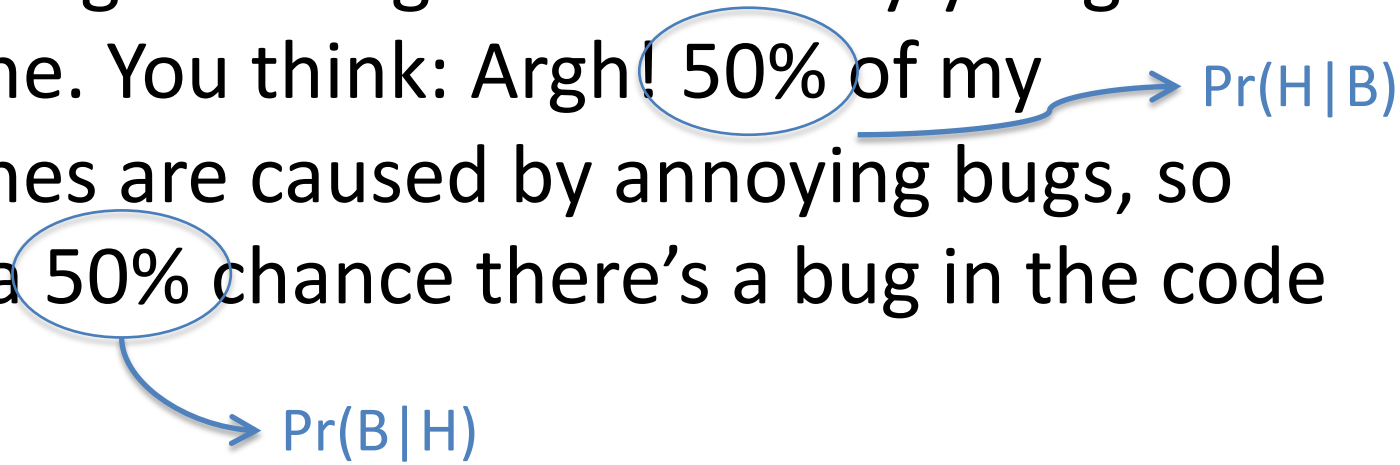
Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

- H = have headache
- B = have bugs in code

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹



- H = have headache
- B = have bugs in code

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

- Given:
  - Pr(H) = 1/10
  - Pr(B) = 1/40
  - Pr(H|B) = 1/2

We observe H.
Task is to compute Pr(B|H).
(conditional probability)

Recall: Conditional Probability

- $\Pr(b \mid a) = \dfrac{\Pr(b \wedge a)}{\Pr(a)}$
- If *Pr(a) = 0*, we set *Pr(b|a) = 1* by convention

- Intuition:
  - Numerator: What is the probability both events occur together?
  - Denominator: What is the probability a occurs at all (regardless of what other events that are happening)?
  - *Pr(b|a)* gives relative weight of *b*-worlds among *a*-worlds

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

$\Pr(H) = 1/10$
$\Pr(B) = 1/40$
$\Pr(H|B) = 1/2$

- **Want:** $\Pr(B|H) = ?$

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

  Pr(H) = 1/10
  Pr(B) = 1/40
  Pr(H|B) = 1/2

- **Want:** Pr(B|H) = ?
  By definition: Pr(B|H) = Pr(B ∧ H) / Pr(H)

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

Pr(H) = 1/10
Pr(B) = 1/40
Pr(H|B) = 1/2

- **Want:** $Pr(B|H) = Pr(B \wedge H) / Pr(H)$
- $Pr(B \wedge H) = Pr(B)Pr(H|B) = (1/40)(1/2) = 1/80$
- Pr(H) is given

Example (cont.)

- You're programming and suddenly you get a headache. You think: Argh! 50% of my headaches are caused by annoying bugs, so there's a 50% chance there's a bug in the code ☹

Pr(H) = 1/10
Pr(B) = 1/40
Pr(H|B) = 1/2
Pr(B ∧ H) = 1/80

- **Want:** Pr(B|H) = Pr(B ∧ H) / Pr(H)
- So Pr(B|H)
  = Pr(B ∧ H)/Pr(H)
  = (1/80) / (1/10) = 1/8 ☺

Bayes Rule

- Note: $Pr(ab) = Pr(ba)$
- We have: $Pr(ab) = Pr(a|b)Pr(b)$
- So: $Pr(a|b)Pr(b) = Pr(ab) = Pr(ba) = Pr(b|a)Pr(a)$

- Bayes rule states:

$$Pr(b|a) = \frac{Pr(a|b)Pr(b)}{Pr(a)}$$

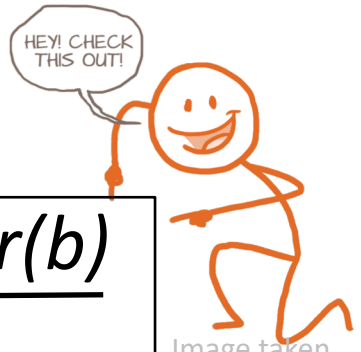


Image taken from giphy.com

- Why is this so important?

Using Bayes Rule for Inference

- We may want to form a hypothesis (H) about the world based on the evidence (e) we observe

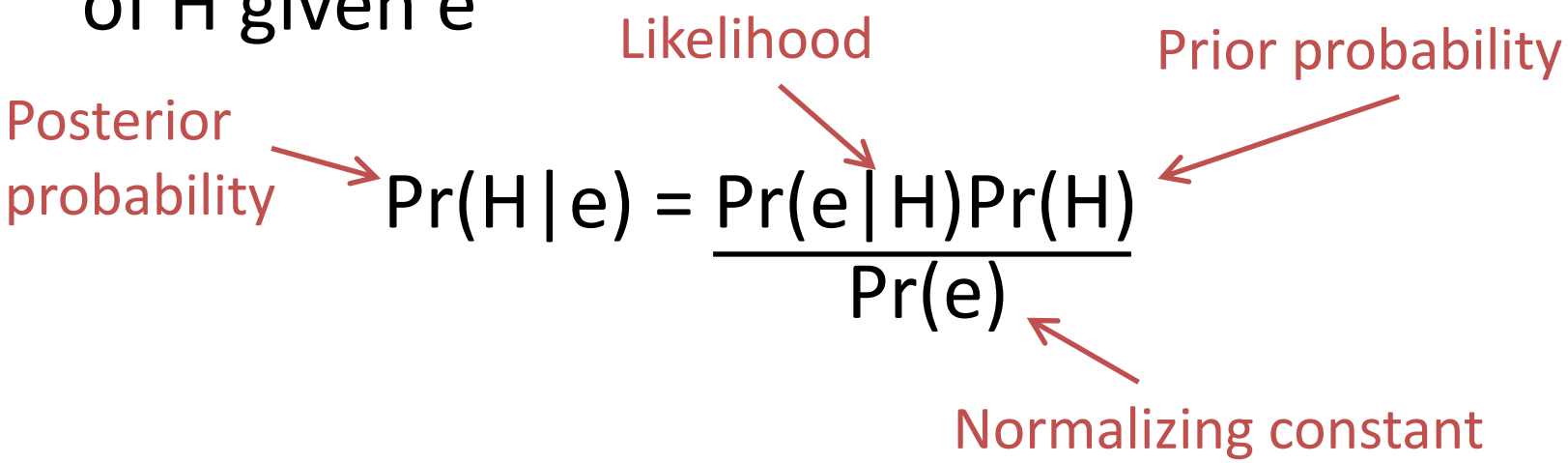
- Bayes rule expresses this notion as the belief of H given e

$$Pr(H|e) = \frac{Pr(e|H)Pr(H)}{Pr(e)}$$

Using Bayes Rule for Inference

- We may want to form a hypothesis (H) about the world based on the evidence (e) we observe

- Bayes rule expresses this notion as the belief of H given e



Example

- Doctor X knows that Asian flu causes fever 95% of the time.
- X knows that a random person has a 10^{-7} chance of having Asian flu.
- X knows that 1 in 100 people suffer from a fever.
- Joe has a fever: what are the chances that Asian flu is the cause of the fever?

Evidence is symptom (F)

Hypothesis is illness causing symptom (A)

- A = Asian flu
- F = fever
- Pr(A|F) = ?

Same as before – but no need to use joint distr.

Example

- Doctor X knows that Asian flu causes fever 95% of the time.
- X knows that a random person has a 10^{-7} chance of having Asian flu.
- X knows that 1 in 100 people suffer from a fever.
- Joe has a fever: what are the chances that Asian flu is the cause of the fever?

Evidence is symptom (F)

Hypothesis is illness causing symptom (A)

- A = Asian flu
- F = fever
- $Pr(A|F) = \frac{Pr(F|A)Pr(A)}{Pr(F)} = \frac{0.95 \times 10^{-7}}{0.01} = 0.95 \times 10^{-5}$

Need for Simplifying Assumptions

- Previously: compute posterior distribution
- More often: compute posterior joint distribution

- Problem: joint distribution is usually too big
  - Exponential in # variables
- Solution: use independence
  - To simplify computational needs
  - To simplify model

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)
  - In general:

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

Only need n numbers to specify the joint!

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:
  - $\Pr(a|b,c) = \Pr(a|c) \qquad \forall\ a, b, c$

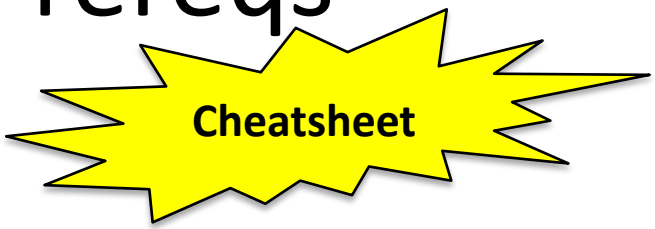
- Knowing the value of B does not change the prediction of A given the presence of C

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

Review: Statistics Prereqs



- Probability distribution
  - All values must sum up to 1.0

- Conditional probability: $\Pr(b|a) = \frac{\Pr(b,a)}{\Pr(a)}$
- Product rule: $\Pr(a,b) = \Pr(a|b)\Pr(b)$
- Sum-out rule (marginalization): $\Pr(a) = \sum_b \Pr(a,b) = \sum_b \Pr(a|b)\Pr(b)$
- Chain rule: $\Pr(abcd) = \Pr(a|bcd)\Pr(b|cd)\Pr(c|d)\Pr(d)$
  - Applies to any number of variables

- Bayes rule: $\Pr(b|a) = \frac{\Pr(a|b)\Pr(b)}{\Pr(a)}$

Developing a Bayes Net

- 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.

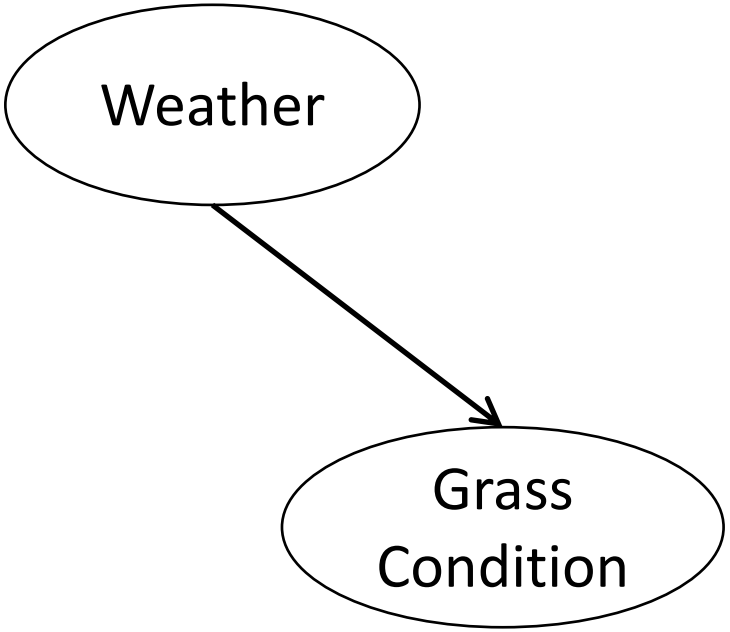
- How to model these interactions?

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.



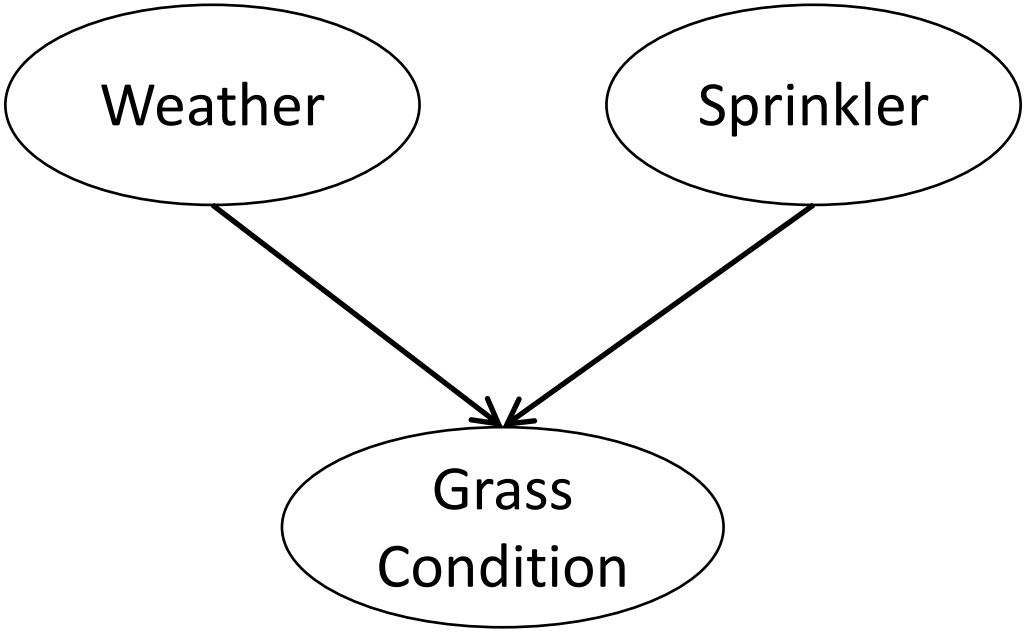
Weather = {sunny, cloudy, rainy}

Sprinkler = {on, off}

GrassCondition = {wet, dry}

Modeling with a Bayes Net

If the sprinkler is on, the grass is wet.



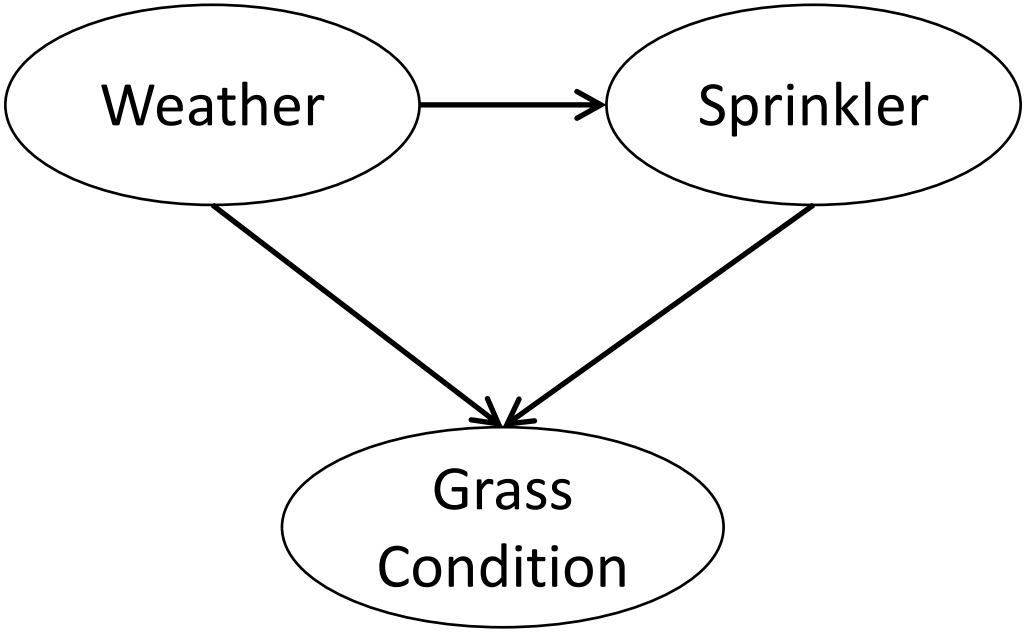
Weather = {sunny, cloudy, rainy}

Sprinkler = {on, off}

GrassCondition = {wet, dry}

Modeling with a Bayes Net

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

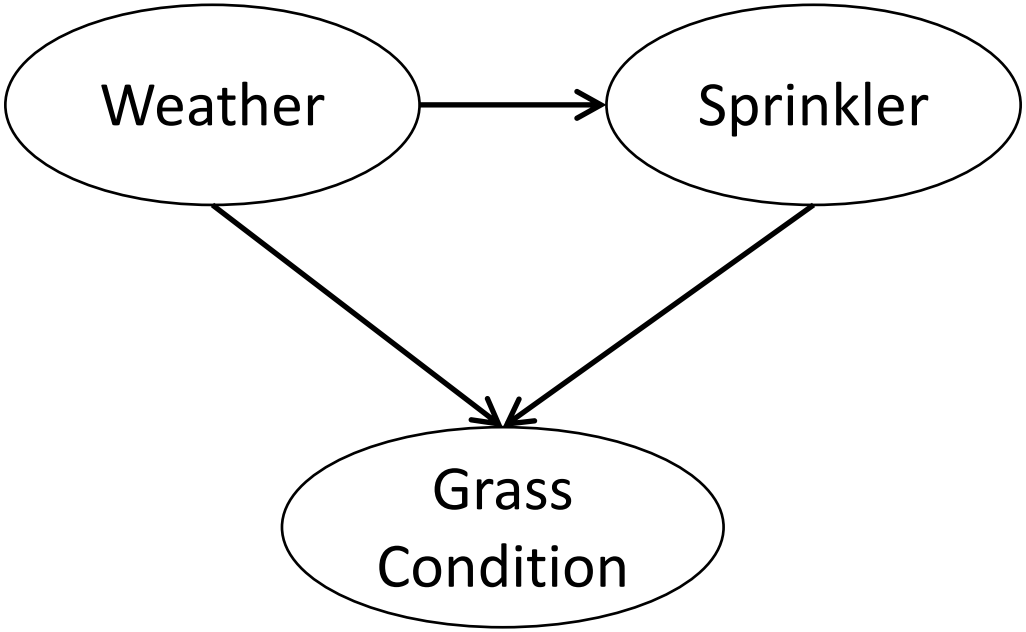


Weather = {sunny, cloudy, rainy}

Sprinkler = {on, off}

GrassCondition = {wet, dry}

Most Popular Bayes Net Example



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

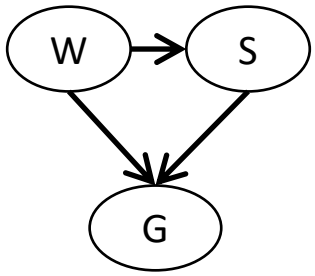
What is a Bayes Net (BN)

- 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

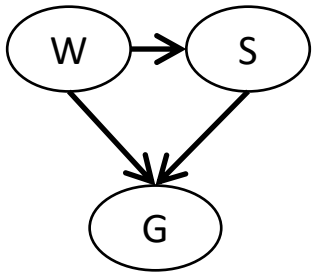
Pr(W=sunny)	Pr(W=cloudy)	Pr(W=rainy)
0.6	0.3	0.1



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

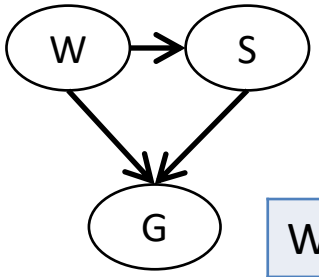


	Pr(S=on|W)	Pr(S=off|W)
W=sunny	0.1	0.9
W=cloudy	0.8	0.2
W=rainy	0.001	0.999

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

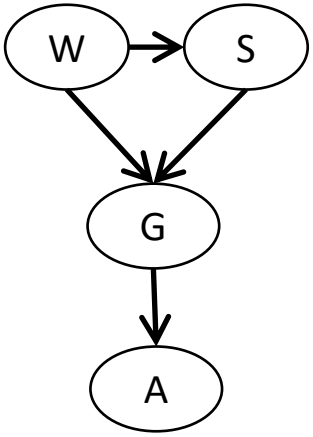


	Pr(S=on|W)	Pr(S=off|W)
W=sunny	0.1	0.9
W=cloudy	0.8	0.2
W=rainy	0.001	0.999

		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

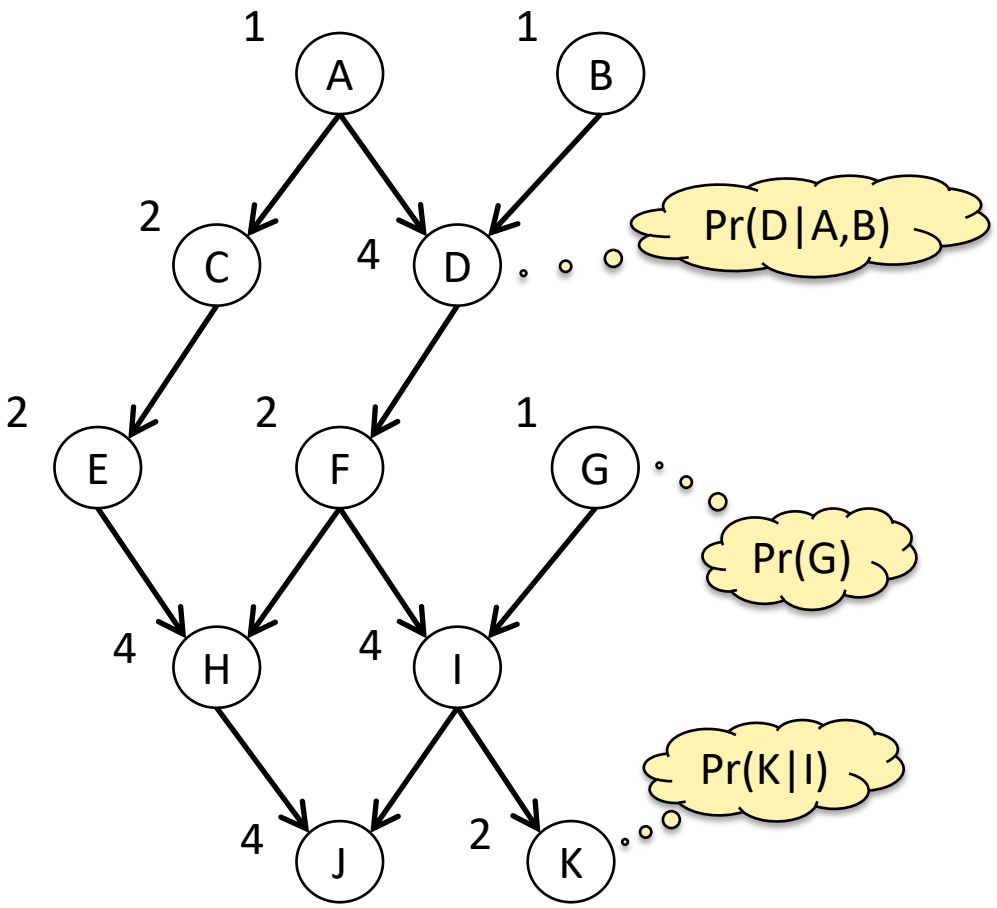
Key Terminology

- 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:

$$Pr(X_i | S \cup Par(X_i)) = Pr(X_i | Par(X_i))$$
for any subset $S \subseteq NonDescendants(X_i)$

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:
  $Pr(x_1,\ldots,x_n)$
  $= Pr(x_n \mid x_{n-1},\ldots,x_1)Pr(x_{n-1} \mid x_{n-2},\ldots,x_1)\ldots Pr(x_1)$
  $= Pr(x_n \mid Par(x_n))Pr(x_{n-1} \mid Par(x_{n-1}))\ldots Pr(x_1)$

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

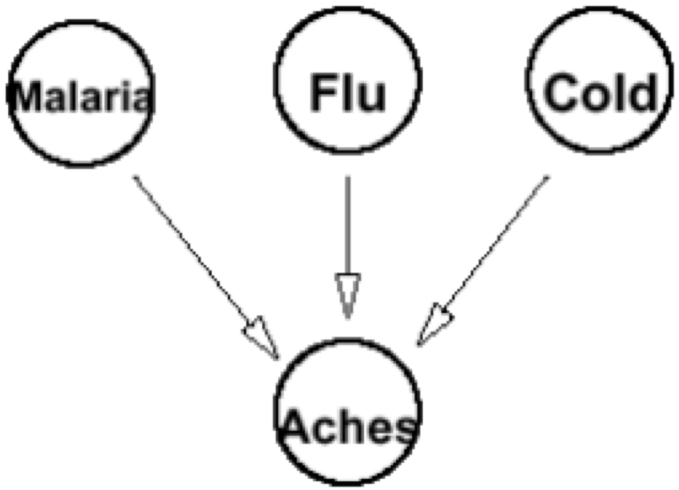
Constructing a Bayes Net

- Given any distribution over variables $X_1,\ldots,X_n$, we can construct a BN that faithfully represents that distribution

- Procedure is simple
  - Works with arbitrary orderings of variable set
  - But some orderings are much better than others!
  - Generally, if ordering/dependence structure reflects causal intuitions, a more natural, compact BN results

Causal Intuitions

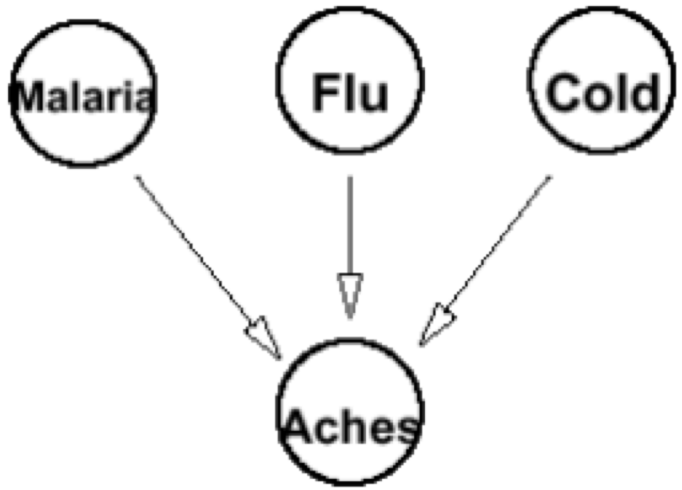
- In this BN, we've used the ordering:
  Malaria, Cold, Flu, Aches
  to build the BN for distribution *P* for Aches
  - Variables can only have parents that come earlier in the ordering



How many parameters needed?

Causal Intuitions

- In this BN, we've used the ordering:
  Malaria, Cold, Flu, Aches
  to build the BN for distribution *P* for Aches
  - Variables can only have parents that come earlier in the ordering

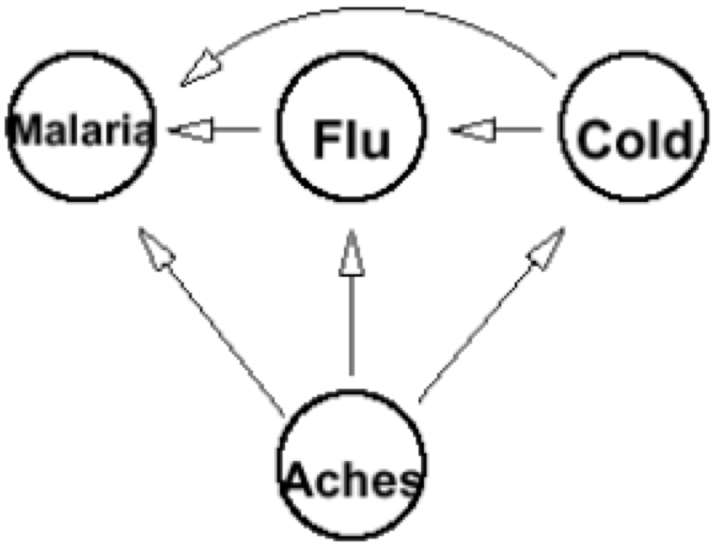


How many parameters needed?

- Top CPTs has $1 = 2^0$ numbers each
- Aches CPT has $8 = 2^3$ numbers

Causal Intuitions

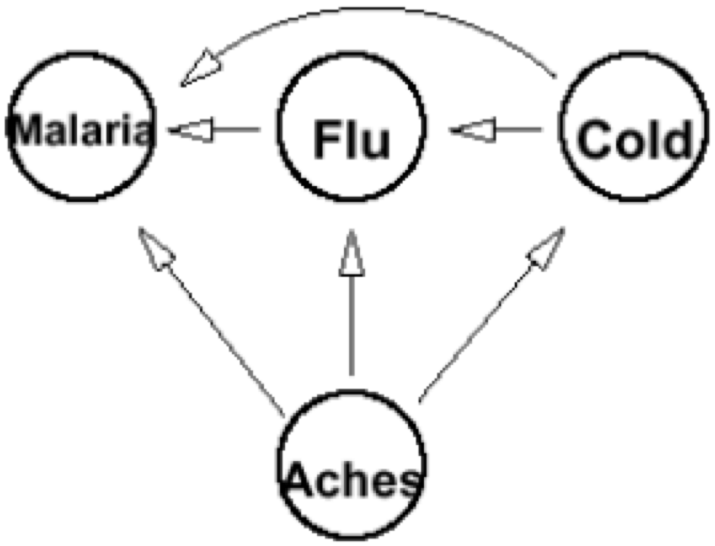
- Suppose we build the BN for distribution *P* using the opposite ordering:
  Aches, Cold, Flu, Malaria
  resulting network is more complicated!



- Malaria depends on Aches;
  Malaria depends on
  Flu, Cold given Aches
- Flu depends on Aches;
  Flu depends on Cold given Aches
- Cold depends on Aches

Causal Intuitions

- Suppose we build the BN for distribution *P* using the opposite ordering:
  Aches, Cold, Flu, Malaria
  resulting network is more complicated!



How many parameters needed?

- Malaria CPT has $8 = 2^3$ numbers
- Flu CPT has $4 = 2^2$ numbers
- Cold CPT has $2 = 2^1$ numbers
- Aches CPT has $1 = 2^0$ number

General Guidelines for Building BNs

- We usually do not build the model based on knowledge about the joint probability distribution
- Typically, we have some vague idea of the dependencies in the world, then we define it precisely into a graphical model
- Steps to follow:
  - Formulate the problem
  - Define the RVs involved
  - Choose independence relations
  - Assign probabilities in the CPTs

Guidelines for Choosing RVs

- Variables must be precise
  - What are the values?
  - How to define them?
  - How to measure them?
  - E.g. weather: difference between the values cold vs. bitter-cold?
- Our discussion: discrete variables
- Different kinds of variables:
  - Observable
  - Hidden – may or may not be useful to include, depending on other independencies they generate

Guidelines for Building the Graph

- When we have information about causality, use causal connections to simplify graph

- Consider tradeoffs between precision of the model and size/sparsity of the graph

Guidelines for Defining Numerical Parameters

- Where do the probabilities come from?
  - Expert
  - Approximate analysis
  - Guessing
  - Learning from data

Guidelines for Defining Numerical Parameters

- Where do the probabilities come from?
  - Expert
  - Approximate analysis
  - Guessing
  - Learning from data

- **Bad news:**
  - In all these cases, the numbers are approximate!

Guidelines for Defining Numerical Parameters

- Where do the probabilities come from?
  - Expert
  - Approximate analysis
  - Guessing
  - Learning from data

- **Bad news:**
  - In all these cases, the numbers are approximate!
- **Good news:**
  - The numbers usually do not matter all that much
  - Sensitivity analysis can help decide if certain numbers are critical or not for the conclusions

Guidelines for Defining Numerical Parameters

- Avoid assigning zero probability to any event
- The relative values of conditional probabilities for $Pr(X_i|Par(X_i))$ given different values of $Par(X_i)$ is important
- Having probabilities that are orders of magnitude different can cause problems in the network

Key Ideas

- Main concept
  - Regression vs. classification
  - Using probability to model uncertainty in the world
- Representation:
  - States as an assignment of values to each RV
  - Beliefs over states as probability distributions
  - 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. $Pr(X_i|Par(X_i))$
- Computational issues:
  - Joint distributions are often too large to compute
  - Computational bottlenecks in computing joint probability distributions appear in representation and inference
  - Exploit independence and conditional independence
  - Computation is linear rather than exponential

Further Readings

- A. Darwiche. Bayesian Networks. Handbook of Knowledge Representation. Volume 3, pages 467—509. Elsevier. 2008.

- S. Russell and P. Norvig. AIMA: Chapter 15.