15-780: Graduate Artificial Intelligence

Probabilistic Reasoning and Inference

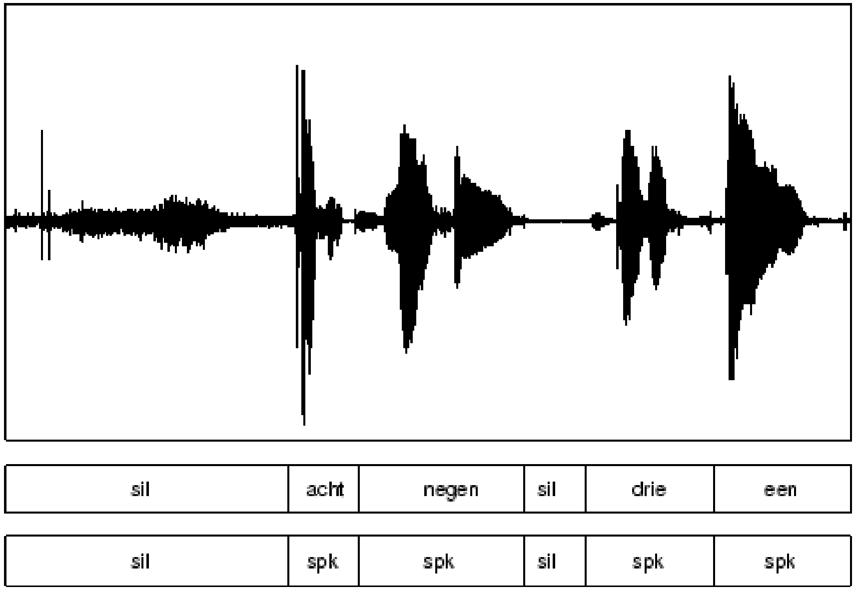
Advantages of probabilistic reasoning

- Appropriate for complex, uncertain, environments
  - Will it rain tomorrow?
- Applies naturally to many domains
  - Robot predicting the direction of road, biology, Word paper clip
- Allows to generalize acquired knowledge and incorporate prior belief
  - Medical diagnosis
- Easy to integrate different information sources
  - Robot's sensors

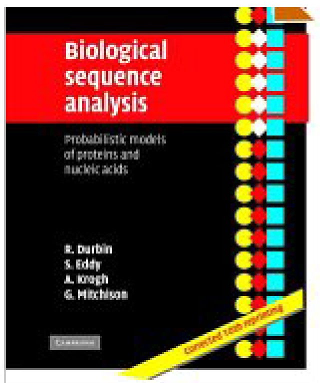
Examples

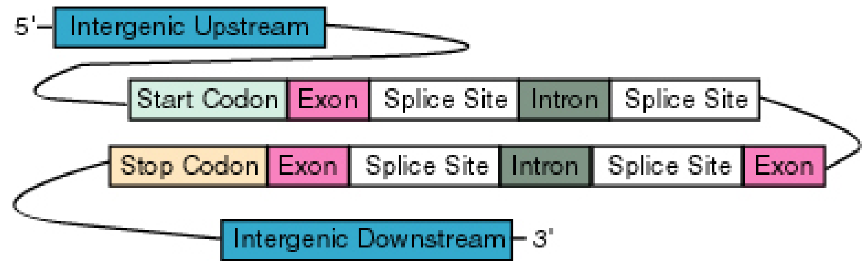
- Unmanned vehicles

Examples: Speech processing



Example: Biological data





ATGAAGCTACTGTCTTCTATCGAACAAGCATGCG
ATATTTGCCGACTTAAAAAGCTCAAG
TGCTCCAAAGAAAAACCGAAGTGCGCCAAGTGT
CTGAAGAACAACTGGGAGTGTCGCTAC
TCTCCCAAAACCAAAAGGTCTCCGCTGACTAGG
GCACATCTGACAGAAGTGGAATCAAGG
CTAGAAAGACTGGAACAGCTATTTCTACTGATTT
TTCCTCGAGAAGACCTTGACATGATT

Basic notations

- Random variable

  - referring to an element / event whose status is unknown:

    A = “it will rain tomorrow”
- Domain

  - The set of values a random variable can take:

    - “A = The stock market will go up this year”: Binary
    - “A = Number of Steelers wins in 2006”: Discrete
    - “A = % change in Google stock in 2006”: Continuous

Priors

Degree of belief in an event in the absence of any other information

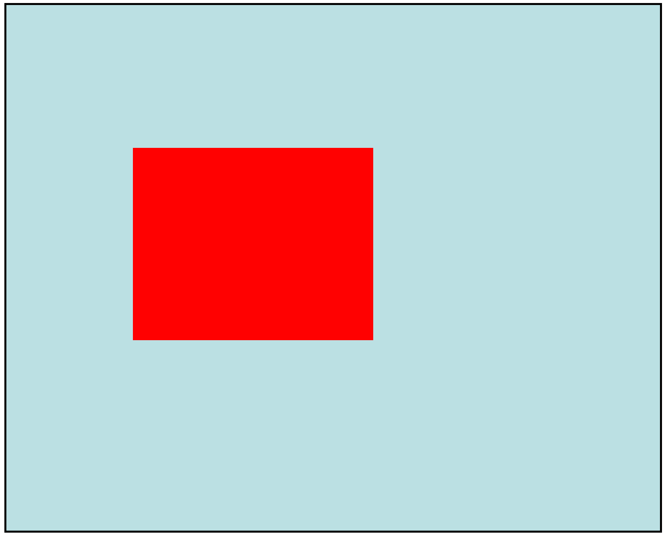
P(rain tomorrow) = 0.2

P(no rain tomorrow) = 0.8

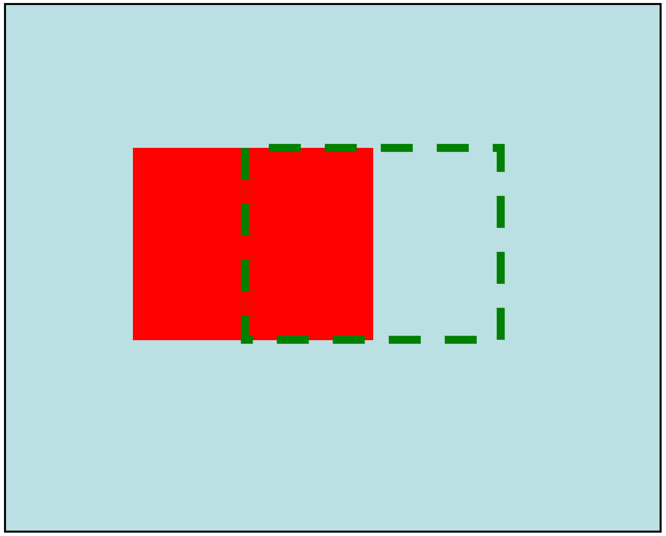
Conditional probability

- P(A = 1 | B = 1): The fraction of cases where A is true if B is true

P(A = 0.2)



P(A|B = 0.5)



Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- For example:

p(slept in movie) = 0.5

p(slept in movie | liked movie) = 1/3

p(didn't sleep in movie | liked movie) = 2/3

Liked movie	Slept	P
1	1	0.2
1	0	0.4
0	0	0.1
0	1	0.3

Joint distributions

- The probability that a set of random variables will take a specific value is their joint distribution.
- Notation: $P(A \wedge B)$ or $P(A,B)$
- Example: P(liked movie, slept)

Liked movie	Slept	P
1	1	0.2
1	0	0.4
0	0	0.1
0	1	0.3

Joint distribution (cont)

P(class size > 20) = 0.5

P(summer) = 1/3

P(class size > 20, summer) = 0

Evaluation of classes

Time (regular =2, summer =1)	Class size	Evaluation (1-3)
1	10	2
2	34	3
1	12	2
2	65	1
2	15	3
2	43	1
1	13	3
2	51	2

Joint distribution (cont)

P(class size > 20) = 0.5

P(eval = 1) = 2/9

P(class size > 20, eval = 1) = 2/9

Evaluation of classes

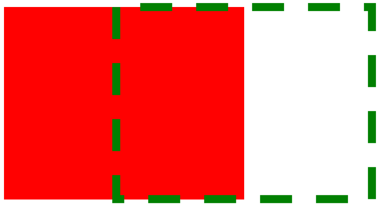
Time (regular =2, summer =1)	Class size	Evaluation (1-3)
1	10	2
2	34	3
1	12	2
2	65	1
2	15	3
2	43	1
1	13	3
2	51	2

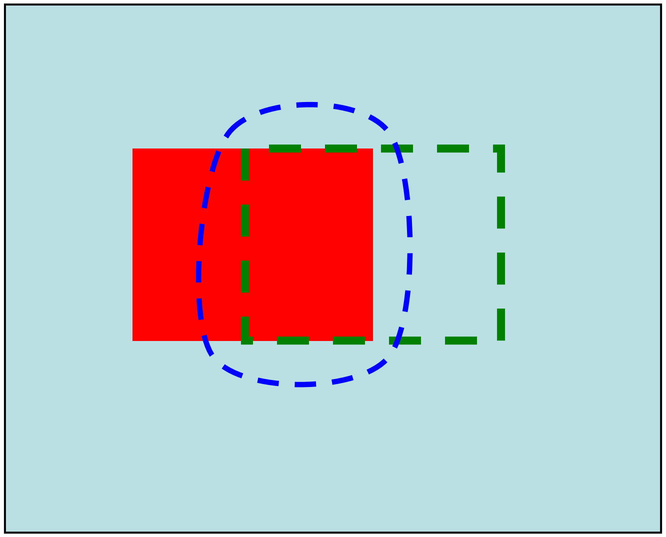
Chain rule

- The joint distribution can be specified in terms of conditional probability:

$$P(A,B) = P(A|B)*P(B)$$

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning





Axioms of probability (Kolmogorov's axioms)

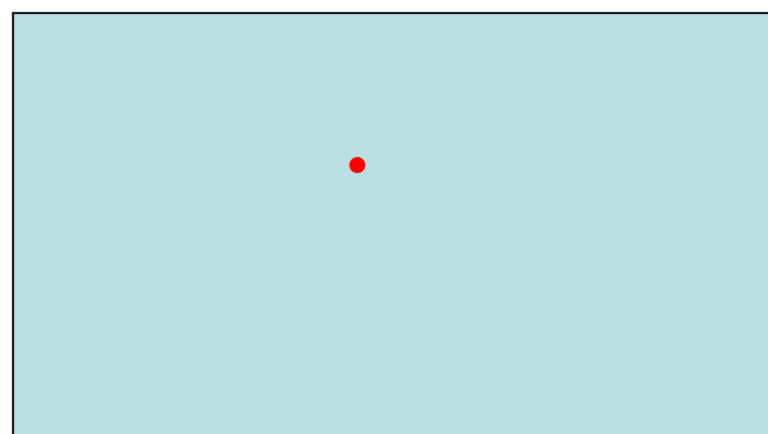
- A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\mathit{true}) = 1$, $P(\mathit{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Axioms of probability (Kolmogorov's axioms)

- A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\textit{true}) = 1$, $P(\textit{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Axioms of probability (Kolmogorov's axioms)

- A variety of useful facts can be derived from just three axioms:

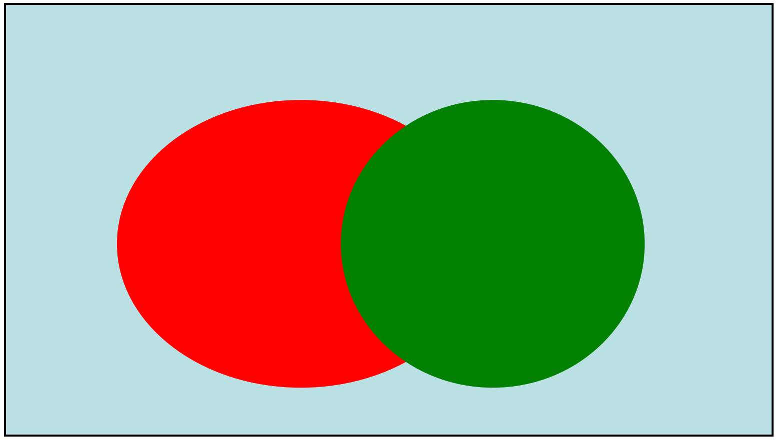
1. $0 \leq P(A) \leq 1$
2. $P(\textit{true}) = 1$, $P(\textit{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

P(Steelers win the 05-06 season) = 1

Axioms of probability (Kolmogorov's axioms)

- A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\mathit{true}) = 1$, $P(\mathit{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

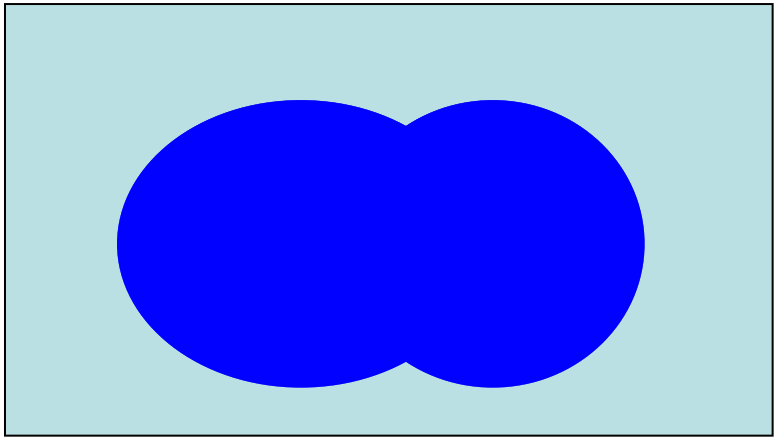


Axioms of probability (Kolmogorov's axioms)

- A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\textit{true}) = 1$, $P(\textit{false}) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

There have been several other attempts to provide a foundation for probability theory. Kolmogorov's axioms are the most widely used.



Using the axioms

- How can we use the axioms to prove that:
  $P(\neg A) = 1 - P(A)$
  ?

Bayes rule

- One of the most important rules for AI usage.
- Derived from the chain rule:

  P(A,B) = P(A | B)P(B) = P(B | A)P(A)
- Thus,

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

Thomas Bayes was an English clergyman who set out his theory of probability in 1764.

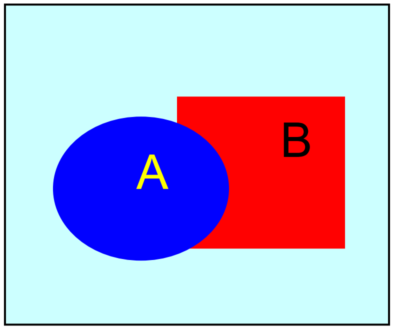
Bayes rule (cont)

Often it would be useful to derive the rule a bit further:

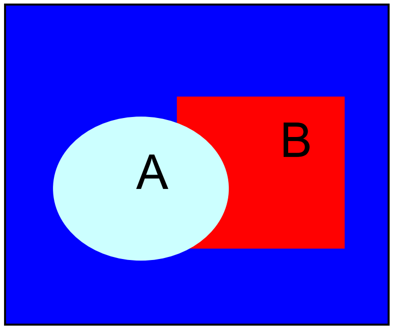
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_{A} P(B|A)P(A)}$$

This results from:
$P(B) = \sum_A P(B,A)$

P(B,A=1)



P(B,A=0)



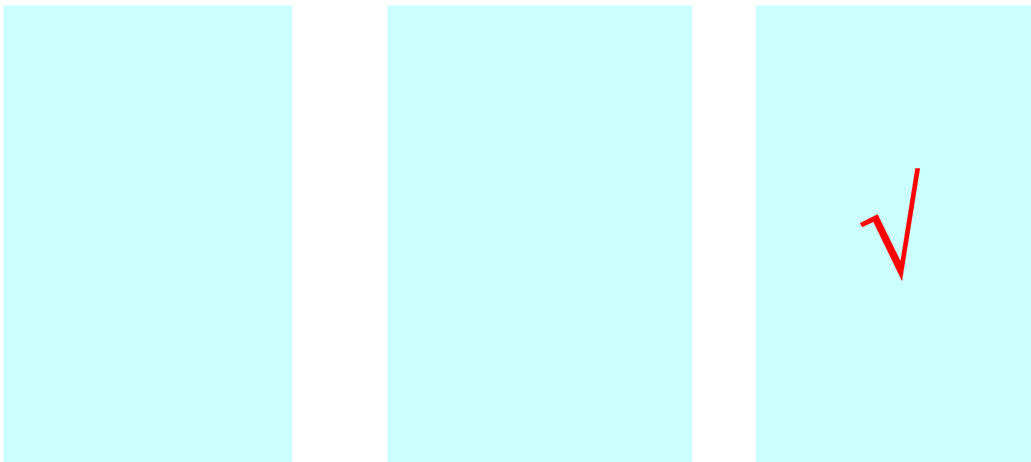
Using Bayes rule

- Cards game:

Place your bet on the location of the King!

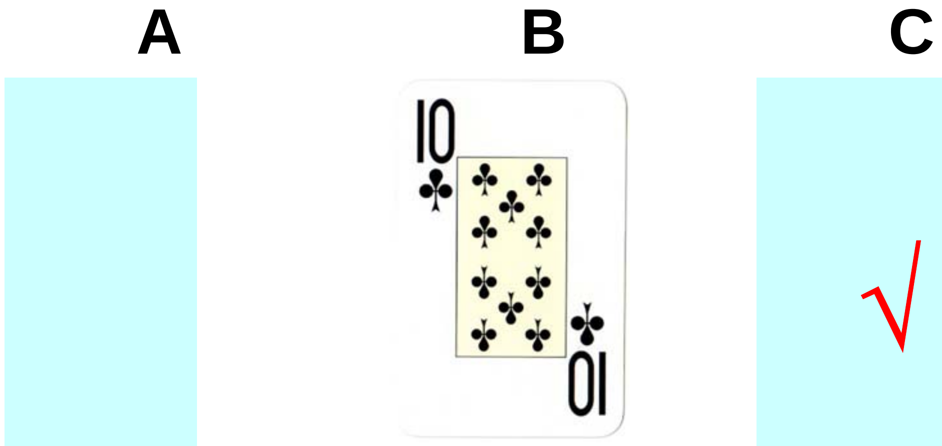
Using Bayes rule

- Cards game:



Do you want to change your bet?

Using Bayes rule



Bayes rule

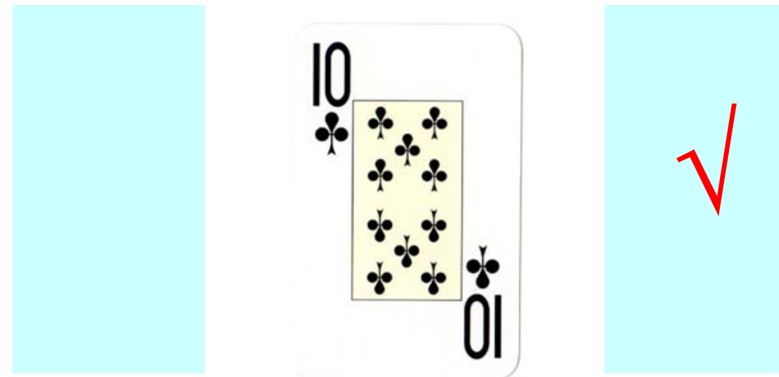
Computing the (posterior) probability: P(C = k | selB)

$$P(C = k \mid selB) = \frac{P(selB \mid C = k)P(C = k)}{P(selB)}$$

$$= \frac{P(selB \mid C = k)P(C = k)}{P(selB \mid C = k)P(C = k) + P(selB \mid C = 10)P(C = 10)}$$

Using Bayes rule

A **B** **C**



$P(C{=}k \mid selB) =$

$$= \frac{\overset{1/2}{P(selB \mid C=k)}\,\overset{1/3}{P(C=k)}}{\underset{1/2}{P(selB \mid C=k)}\,\underset{1/3}{P(C=k)} + \underset{1/2}{P(selB \mid C=10)}\,\underset{2/3}{P(C=10)}} = \mathbf{1/3}$$

Joint distributions

- The probability that a set of random variables will take a specific value is their joint distribution.
- Requires a joint probability table to specify the possible assignments
- The table can grow very rapidly …

Liked movie	Slept	P
1	1	0.2
1	0	0.4
0	0	0.1
0	1	0.3

How can we decrease the number of columns in the table?

Independence

- In some cases the additional information does not help

P(slept) = 0.5

P(slept | rain = 1) = 0.5

- In this case, the extra knowledge about rain does not change our prediction
- Slept and rain are independent!

Liked movie	Slept	raining	P
1	1	1	0.1
1	0	1	0.2
0	0	1	0.05
0	1	1	0.15
1	1	0	0.1
1	0	0	0.2
0	0	0	0.05
0	1	0	0.15

Independence (cont.)

- Notation: $P(S \mid R) = P(S)$
- Using this we can derive the following:
  - $P(\neg S \mid R) = P(\neg S)$
  - $P(S,R) = P(S)P(R)$
  - $P(R \mid S) = P(R)$

Independence

- Independence allows for easier models, learning and inference
- For our example:
  - P(raining, slept movie) = P(raining)P(slept movie)
  - Instead of 4 by 2 table (4 parameters), only 2 are required
  - The saving is even greater if we have many more parameters …
- In many cases it would be useful to assume independence, even if its not the case

Conditional independence

- Two dependent random variables may become independent when conditioned on a third variable:

  P(A,B | C) = P(A | C) P(B | C)

- Example

  P(liked movie) = 0.5

  P(slept) = 0.4

  P(liked movie, slept) = 0.1

  P(liked movie | long) = 0.4

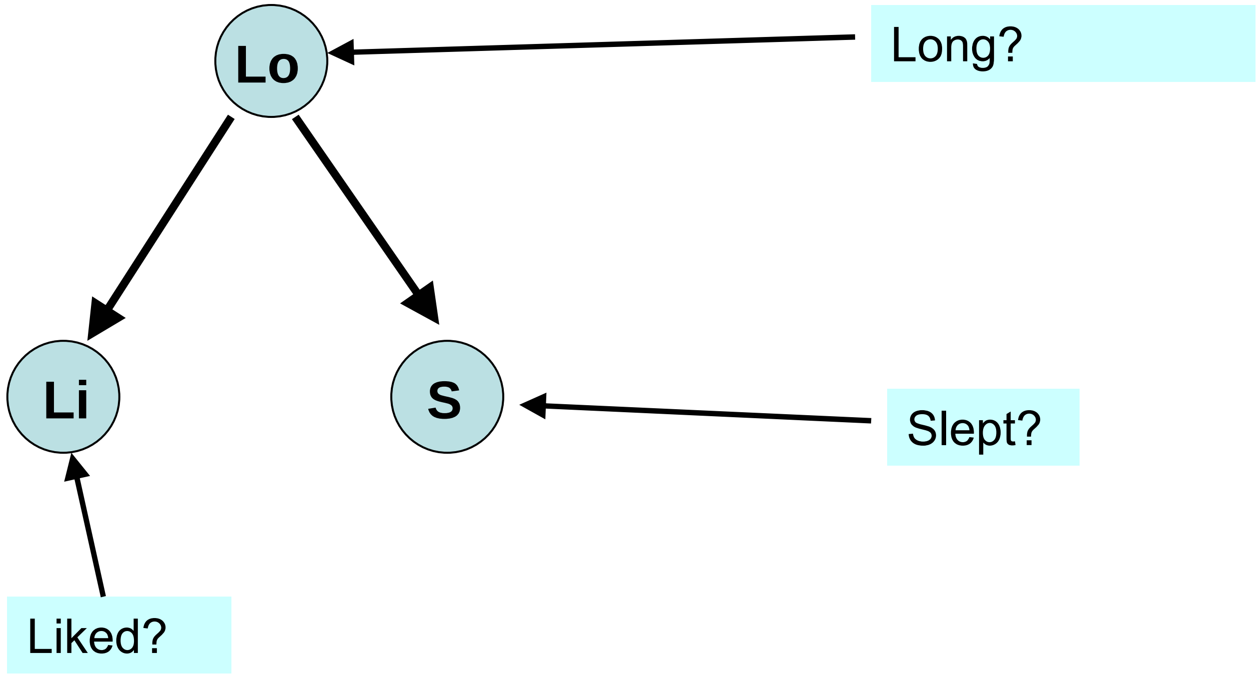
  P(slept | long) 0.6

  P(slept, like movie | long) = 0.24

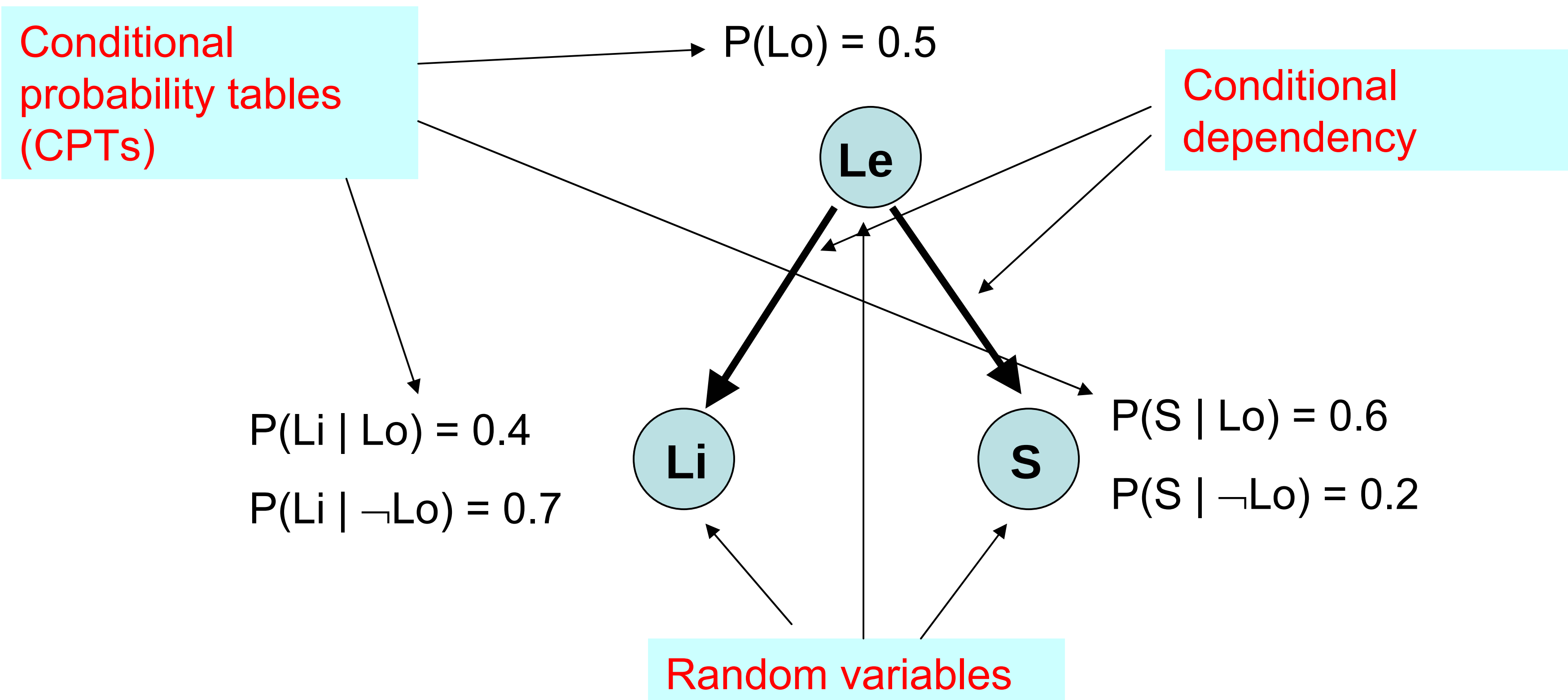
Given knowledge of length, the two other variables become independent

Bayesian networks

- Bayesian networks are *directed graphs* with nodes representing *random variables* and edges representing *dependency assumptions*



Bayesian networks: Notations



Constructing a Bayesian network

- How do we go about constructing a network for a specific problem?
- Step 1: Identify the random variables
- Step 2: Determine the conditional dependencies
- Step 3: Populate the CPTs

Can be learned from observation data!

A example problem

- An alarm system

  B – Did a burglary occur?

  E – Did an earthquake occur?

  A – Did the alarm sound off?

  M – Mary calls

  J – John calls

- How do we reconstruct the network for this problem?

Factoring joint distributions

- Using the chain rule we can always factor a joint distribution as follows:

  P(A,B,E,J,M) =

  P(A | B,E,J,M) P(B,E,J,M) =

  P(A | B,E,J,M) P(B | E,J,M) P(E,J,M) =

  P(A | B,E,J,M) P(B | E, J,M) P(E | J,M) P(J,M)

  P(A | B,E,J,M) P(B | E, J,M) P(E | J,M)P(J | M)P(M)
- This type of conditional dependencies can also be represented graphically.

A Bayesian network

Number of parameters:

A: 2^4

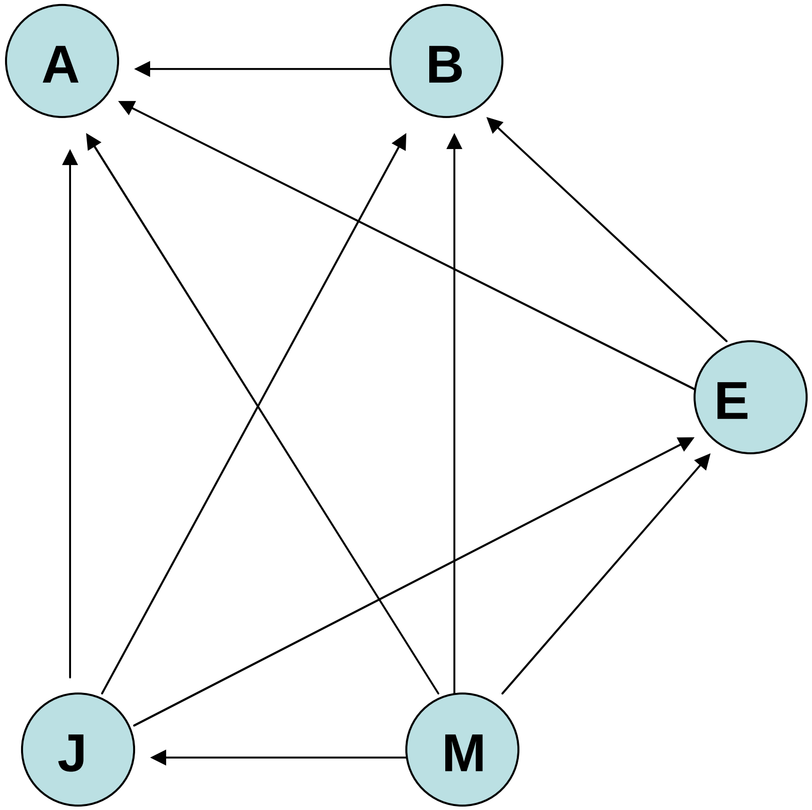
B: 2^3

E: 4

J: 2

M: 1

A total of 31 parameters



A better approach

- An alarm system

  B – Did a burglary occur?

  E – Did an earthquake occur?

  A – Did the alarm sound off?

  M – Mary calls

  J – John calls

- Lets use our knowledge of the domain!

Reconstructing a network

Number of parameters:

A: 4

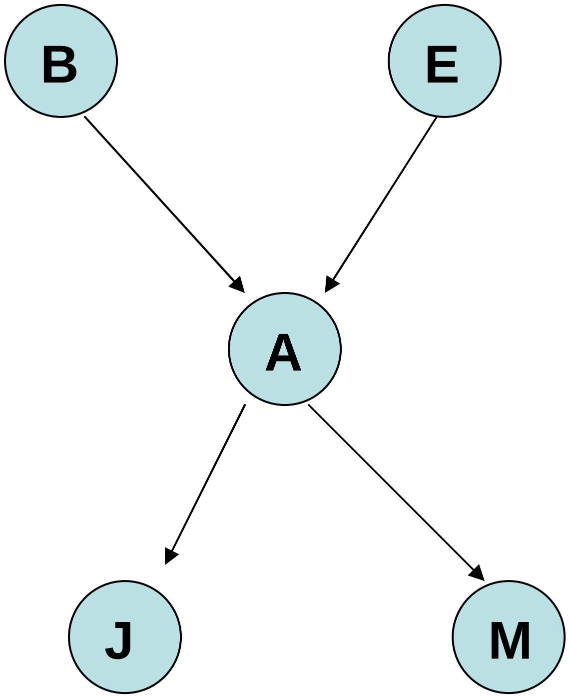
B: 1

E: 1

J: 2

M: 2

A total of 10 parameters



By relying on domain knowledge we saved 21 parameters!

Constructing a Bayesian network: Revisited

- Step 1: Identify the random variables
- Step 2: Determine the conditional dependencies

  - Select on ordering of the variables

  - Add them one at a time

  - For each new variable X added select the minimal subset of nodes as parents such that X is independent from all other nodes in the current network given its parents.
- Step 3: Populate the CPTs

  - We will discuss this when we talk about density estimations

Important points

- Random variables
- Chain rule
- Bayes rule
- Joint distribution, independence, conditional independence