Bayesian Inference

Uncertainty

Uncertainty

Classical logic

- Only permit exact reasoning
- Assume perfect knowledge exists
- Assume law of excluded middle can be applied

IF A is true	IF A is false
THEN A is not false	THEN A is not true

Uncertainty

Information

- Can be unsuitable for solving problem
  - Incomplete
  - Inconsistent
  - Uncertain

Uncertainty

- Lack of exact knowledge
  - Reach perfectly reliable conclusion

Uncertain Knowledge

Weak implications

- Domain experts and knowledge engineers
  - Difficult to establish concrete correlations for some rules
    - Between IF (condition) and THEN (action) parts
- Expert systems need to handle vague associations
  - Accept degree of correlations
    - Numerical certainty factors

Uncertain Knowledge

Imprecise language

- Natural language is ambiguous and imprecise
- Describe facts with terms
  - Often and sometimes, frequently and hardly ever
- Difficult to express knowledge
  - In precise IF-THEN form of production rules
- Quantify meaning of the facts
  - Can be used in expert systems

Uncertain Knowledge

- In 1944, Ray Simpson asked
  - 355 high school and college students
  - Place 20 terms like often
  - On a scale between 1 and 100
- In 1968, Milton Hakel repeated this experiment

Quantification of Ambiguous and Imprecise Terms on a Time-Frequency Scale

Ray Simpson (1944)		*Milton Hakel* (1968)	
Term	*Mean value*	*Term*	*Mean value*
Always	99	Always	100
Very often	88	Very often	87
Usually	85	Usually	79
Often	78	Often	74
Generally	78	Rather often	74
Frequently	73	Frequently	72
Rather often	65	Generally	72
About as often as not	50	About as often as not	50
Now and then	20	Now and then	34
Sometimes	20	Sometimes	29
Occasionally	20	Occasionally	28
Once in a while	15	Once in a while	22
Not often	13	Not often	16
Usually not	10	Usually not	16
Seldom	10	Seldom	9
Hardly ever	7	Hardly ever	8
Very seldom	6	Very seldom	7
Rarely	5	Rarely	5
Almost never	3	Almost never	2
Never	0	Never	0

Uncertain Knowledge

Combining views of different experts

- Large expert systems
  - Combine knowledge and expertise of a number of experts
  - Have contradictory opinions
  - Produce conflicting rules
    - To resolve the conflict
      - Attach a weight to each expert
      - Calculate the composite conclusion
      - No systematic method exists to obtain these weights

Uncertain Knowledge

Unknown data

- When data is incomplete or missing
- Only solution — Accept the value “unknown”
- Proceed to an approximate reasoning with this value

Basic Probability Theory

Basic Probability Theory

Concept of probability

- Long history, go back thousands of years
- When the following words were introduced into spoken languages
  - “probably”, “likely”, “maybe”, “perhaps” and “possibly”
- Mathematical theory of probability was formulated
  - Only in 17th century

Probability of an event

- Proportion of cases in which the event occurs
- Defined as a scientific measure of chance

Basic Probability Theory

Probability
- Expressed mathematically as a numerical index
- With a range between zero to unity
- From an absolute impossibility to an absolute certainty

Most events have a probability index
- Strictly between 0 and 1
- At least two possible outcomes
  - Favorable outcome or success
  - Unfavorable outcome or failure

Basic Probability Theory

$$P(success) = \frac{the\ number\ of\ successes}{the\ number\ of\ possible\ outcomes}$$

$$P(failure) = \frac{the\ number\ of\ failures}{the\ number\ of\ possible\ outcomes}$$

s: the number of times success can occur

f: the number of times failure can occur

$$P(success) = p = \frac{s}{s+f} \qquad P(failure) = q = \frac{f}{s+f}$$

Basic Probability Theory

Throw a coin

- Probability of getting a head = Probability of getting a tail

In a single throw

- $s = f = 1$
- Probability of getting a head (or a tail) is 0.5

$$P(success) = p = \frac{s}{s+f} \qquad P(failure) = q = \frac{f}{s+f}$$

$$p + q = 1$$

Conditional Probability

A: An event in the world B: Another event in the world

Events A and B are not mutually exclusive

Occur conditionally on the occurrence of the other

Conditional probability

- Probability that event A will occur if event B occurs
- Denoted mathematically as p(A|B)
  - Vertical bar represents GIVEN
  - Complete probability expression is interpreted as
    - "Conditional probability of event A occurring given that event B has occurred"

$$p(A|B) = \frac{\text{the number of times } A \text{ and } B \text{ can occur}}{\text{the number of times } B \text{ can occur}}$$

Conditional Probability

Joint probability of A and B

- The number of times A and B can occur
- Probability that both A and B will occur
- Represented mathematically as $p(A \cap B)$

Probability of B, $p(B)$

- The number of times B can occur

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Conditional probability of event B occurring

- Given that event A has occurred

$$p(B|A) = \frac{p(B \cap A)}{p(A)}$$

Conditional Probability

Joint probability

$$p(B \cap A) = p(B|A) \times p(A)$$

$$p(A \cap B) = p(B|A) \times p(A)$$

Substituting the last equation into the equation

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

yields the Bayesian rule

Bayesian Rule

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

p(A|B)
- Conditional probability that event A occurs
  - Given event B has occurred

p(B|A)
- Conditional probability of event B occurring
  - Given event A has occurred

p(A)
- Probability of event A occurring

p(B)
- Probability of event B occurring

Conditionality Probability

Concept of conditionality probability

- Event A is dependent upon event B

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Can be extended

- Event A is dependent on
  - A number of mutually exclusive events $B_1, B_2, \ldots, B_n$

$$p(A \cap B_1) = p(A|B_1) \times p(B_1)$$

$$p(A \cap B_2) = p(A|B_2) \times p(B_2)$$

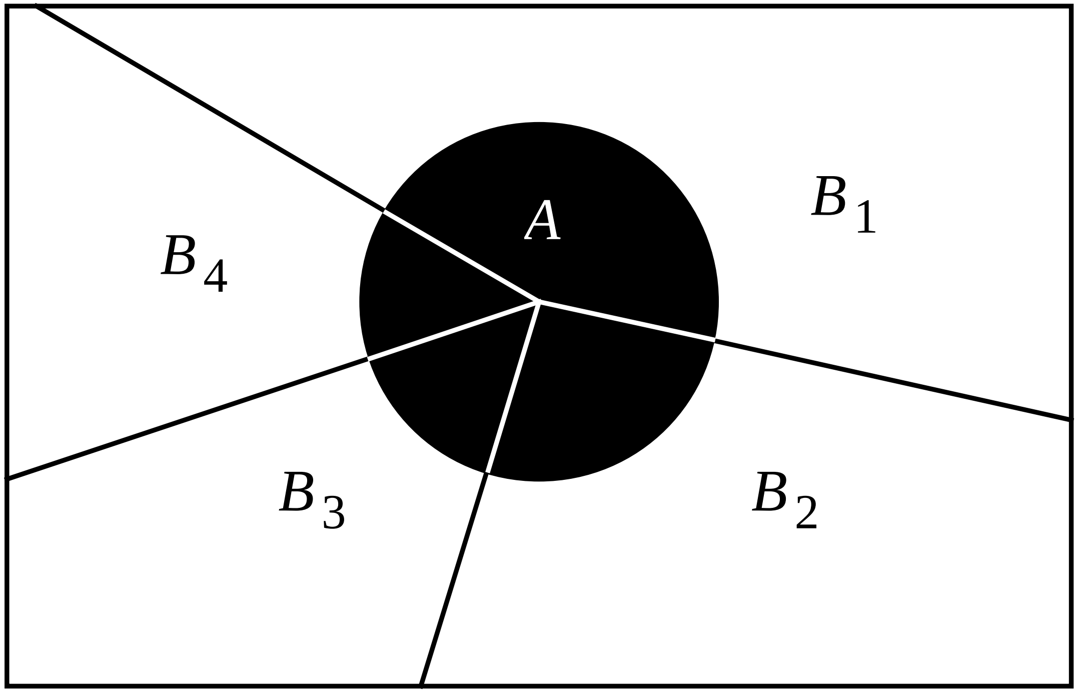
$$\cdot$$

$$p(A \cap B_n) = p(A|B_n) \times p(B_n)$$

$$\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

Joint Probability

$$\sum_{i=1}^{n} p\left(A \cap B_i\right) = \sum_{i=1}^{n} p\left(A \middle| B_i\right) \times p\left(B_i\right)$$



For All Events B_i

$$\sum_{i=1}^{n} p\left(A \cap B_i\right) = \sum_{i=1}^{n} p\left(A \middle| B_i\right) \times p\left(B_i\right)$$

$$\sum_{i=1}^{n} p\left(A \cap B_i\right) = p\left(A\right)$$

Therefore

$$p(A) = \sum_{i=1}^{n} p\left(A \middle| B_i\right) \times p\left(B_i\right)$$

For All Events B_i

If occurrence of event A depends on

- Only two mutually exclusive events
- B and NOT B

$$p(A) = p(A|B) \times p(B) + p(A|\neg B) \times p(\neg B)$$

$$p(B) = p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)$$

Substitute into Bayesian rule yields:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)}$$

Bayesian Reasoning

Bayesian Reasoning

Suppose all rules in KB are represented as:

IF E is true

THEN H is true {with probability p}

Imply that

- If event E occurs
- Then probability that event H will occur is p

In expert systems

- H usually represents hypothesis
- E denotes evidence to support the hypothesis

Bayesian Reasoning

Express Bayesian rule in terms of hypotheses and evidence

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E|H) \times p(H) + p(E|\neg H) \times p(\neg H)}$$

p(H)

- Prior probability of hypothesis H being true

p(E|H)

- Probability that hypothesis H being true will result in evidence E

p(¬H)

- Prior probability of hypothesis H being false

p(E|¬H)

- Probability of finding evidence E even when hypothesis H is false

Bayesian Reasoning

Probabilities required to solve a problem

- Provided by expert

Prior probabilities

- Possible hypotheses $p(H)$ and $p(\neg H)$

Conditional probabilities, observing evidence E

- If hypothesis H is true, $p(E|H)$
- If hypothesis H is false, $p(E|\neg H)$

Users provide information about evidence observed

Expert system computes probability $p(H|E)$

- Posterior probability of hypothesis H upon observing evidence E

Bayesian Reasoning

Take into account

- Multiple hypotheses H_1, H_2,..., H_m
- Multiple evidences E_1, E_2,..., E_n
- Must be mutually exclusive and exhaustive

Single evidence E and multiple hypotheses:

$$p(H_i|E) = \frac{p(E|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E|H_k) \times p(H_k)}$$

Multiple evidences and multiple hypotheses:

$$p(H_i|E_1\, E_2 \ldots E_n) = \frac{p(E_1\, E_2 \ldots E_n|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E_1\, E_2 \ldots E_n|H_k) \times p(H_k)}$$

Bayesian Reasoning

Multiple evidences and multiple hypotheses

- Obtain conditional probabilities
  - All possible combinations of evidences for all hypotheses
  - Enormous burden on expert

Assume conditional independence among evidences

$$p\left(H_i \middle| E_1\, E_2 \ldots E_n\right) = \frac{p\left(E_1 \middle| H_i\right) \times p\left(E_2 \middle| H_i\right) \times \ldots \times p\left(E_n \middle| H_i\right) \times p\left(H_i\right)}{\sum_{k=1}^{m} p\left(E_1 \middle| H_k\right) \times p\left(E_2 \middle| H_k\right) \times \ldots \times p\left(E_n \middle| H_k\right) \times p\left(H_k\right)}$$

Simple Example

Given to expert

- Three conditionally independent evidences
  - E_1, E_2 and E_3

Expert determines

- Three mutually exclusive and exhaustive hypotheses
  - H_1, H_2 and H_3
- Prior probabilities for these hypotheses
  - $p(H_1)$, $p(H_2)$ and $p(H_3)$
- Conditional probabilities of observing each evidence for all possible hypotheses

The Prior and Conditional Probabilities

Probability	*Hypothesis*		
	$i = 1$	$i = 2$	$i = 3$
$p(H_i)$	0.40	0.35	0.25
$p(E_1 \mid H_i)$	0.3	0.8	0.5
$p(E_2 \mid H_i)$	0.9	0.0	0.7
$p(E_3 \mid H_i)$	0.6	0.7	0.9

Posterior Probabilities

First observe evidence E_3

- Compute posterior probabilities for all hypotheses

$$p(H_i|E_3) = \frac{p(E_3|H_i) \times p(H_i)}{\sum_{k=1}^{3} p(E_3|H_k) \times p(H_k)}, \qquad i = 1, 2, 3$$

Belief in hypothesis H_1

- Decrease
- Equal to belief in hypothesis H_2

$$p(H_1|E_3) = \frac{0.6 \cdot 0.40}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.34$$

Belief in hypothesis H_3

- Increase
- Nearly reach beliefs in hypotheses H_1 and H_2

$$p(H_2|E_3) = \frac{0.7 \cdot 0.35}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.34$$

$$p(H_3|E_3) = \frac{0.9 \cdot 0.25}{0.6 \cdot 0.40 + 0.7 \cdot 0.35 + 0.9 \cdot 0.25} = 0.32$$

Posterior Probabilities

Observe evidence E_1

$$p(H_i|E_1E_3) = \frac{p(E_1|H_i)\times p(E_3|H_i)\times p(H_i)}{\sum_{k=1}^{3} p(E_1|H_k)\times p(E_3|H_k)\times p(H_k)}, \qquad i = 1, 2, 3$$

Hypothesis H_2 become most likely one

$$p(H_1|E_1E_3) = \frac{0.3\cdot 0.6\cdot 0.40}{0.3\cdot 0.6\cdot 0.40 + 0.8\cdot 0.7\cdot 0.35 + 0.5\cdot 0.9\cdot 0.25} = 0.19$$

$$p(H_2|E_1E_3) = \frac{0.8\cdot 0.7\cdot 0.35}{0.3\cdot 0.6\cdot 0.40 + 0.8\cdot 0.7\cdot 0.35 + 0.5\cdot 0.9\cdot 0.25} = 0.52$$

$$p(H_3|E_1E_3) = \frac{0.5\cdot 0.9\cdot 0.25}{0.3\cdot 0.6\cdot 0.40 + 0.8\cdot 0.7\cdot 0.35 + 0.5\cdot 0.9\cdot 0.25} = 0.29$$

Posterior Probabilities

Observe evidence E_2, final posterior probabilities

$$p(H_i|E_1E_2E_3) = \frac{p(E_1|H_i)\times p(E_2|H_i)\times p(E_3|H_i)\times p(H_i)}{\sum_{k=1}^{3} p(E_1|H_k)\times p(E_2|H_k)\times p(E_3|H_k)\times p(H_k)}, \qquad i = 1, 2, 3$$

Initial ranking is H_1, H_2 and H_3

After all evidences are observed

- Only H_1 and H_3 remain under consideration

$$p(H_1|E_1E_2E_3) = \frac{0.3\cdot 0.9\cdot 0.6\cdot 0.40}{0.3\cdot 0.9\cdot 0.6\cdot 0.40 + 0.8\cdot 0.0\cdot 0.7\cdot 0.35 + 0.5\cdot 0.7\cdot 0.9\cdot 0.25} = 0.45$$

$$p(H_2|E_1E_2E_3) = \frac{0.8\cdot 0.0\cdot 0.7\cdot 0.35}{0.3\cdot 0.9\cdot 0.6\cdot 0.40 + 0.8\cdot 0.0\cdot 0.7\cdot 0.35 + 0.5\cdot 0.7\cdot 0.9\cdot 0.25} = 0$$

$$p(H_3|E_1E_2E_3) = \frac{0.5\cdot 0.7\cdot 0.9\cdot 0.25}{0.3\cdot 0.9\cdot 0.6\cdot 0.40 + 0.8\cdot 0.0\cdot 0.7\cdot 0.35 + 0.5\cdot 0.7\cdot 0.9\cdot 0.25} = 0.55$$

Naïve Bayes Classifiers

Bayes Rules

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

p(A|B)
- Conditional probability that event A occurs
- Given event B has occurred

p(B|A)
- Conditional probability of event B occurring
- Given event A has occurred

p(A)
- Probability of event A occurring

p(B)
- Probability of event B occurring

Bayes Rules

Bayes Rules can be represented based on

- Hypothesis and evidence

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

Maximum A Posteriori (MAP)

Given a set of events E

Compute maximum a posteriori hypothesis

$$h_{MAP} = \arg\max_{h \in H} P(h \mid E)$$

$$h_{MAP} = \arg\max_{h \in H} \frac{P(E \mid h) P(h)}{P(E)}$$

$$h_{MAP} = \arg\max_{h \in H} P(E \mid h) P(h)$$

Omit P(E)

- Constant
- Independent of the hypothesis

Maximum A Posteriori (MAP)

A set of training examples

Records with conjunctive attributes (a_1, a_2, .. a_n)

Target function is finite set of classes V

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, .., a_n)$$

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, .., a_n \mid v_j) P(v_j)}{P(a_1, a_2, .., a_n)}$$

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, .., a_n \mid v_j) P(v_j)$$

Maximum A Posteriori (MAP)

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, .., a_n \mid v_j) P(v_j)$$

P(v_j) can easily be estimated

- Compute frequency of target class in training set

P(a_1,a_2,..a_n|v_j) is difficult to estimate

Naïve Bayes Classifiers

Assume attributes values are conditionally independent

$$P(a_1, a_2, .., a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

What we know

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, .., a_n \mid v_j) P(v_j)$$

Therefore

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Naïve Bayes Classifiers

Estimate $P(a_i|v_j)$ instead of $P(a_1,a_2,..,a_n|v_j)$

- Greatly reduce number of parameters

Learning step in Naïve Bayes

- Estimate $P(a_i|v_j)$ and $P(v_j)$
  - Based on frequencies in training data

No search during training

Classify an unseen instance

- Compute class that maximizes posterior

Naïve Bayes Classifiers Example

Day	*Outlook*	*Temp*	*Humidity*	*Wind*	*Play Tennis*
Dl	Sunny	Hot	High	Weak	**No**
D2	Sunny	Hot	High	Strong	**No**
D3	Overcast	Hot	High	Weak	**Yes**
D4	Rain	Mild	High	Weak	**Yes**
D5	Rain	Cool	Normal	Weak	**Yes**
D6	Rain	Cool	Normal	Strong	**No**
D7	Overcast	Cool	Normal	Strong	**Yes**
D8	Sunny	Mild	High	Weak	**No**
D9	Sunny	Cool	Normal	Weak	**Yes**
D10	Rain	Mild	Normal	Weak	**Yes**
D11	Sunny	Mild	Normal	Strong	**Yes**
D12	Overcast	Mild	High	Strong	**Yes**
D13	Overcast	Hot	Normal	Weak	**Yes**
D14	Rain	Mild	High	Strong	**No**

Naïve Bayes Classifiers Example

Outlook			Temperature			Humidity			Wind			Play	
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
Sunny	2	3	**Hot**	2	2	**High**	3	4	**Weak**	6	2	9	5
Overcast	4	0	**Mild**	4	2	**Normal**	6	1	**Strong**	3	3		
Rainy	3	2	**Cool**	3	1								
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
Sunny	2/9	3/5	**Hot**	2/9	2/5	**High**	3/9	4/5	**Weak**	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	**Mild**	4/9	2/5	**Normal**	6/9	1/5	**Strong**	3/9	3/5		
Rainy	3/9	2/5	**Cool**	3/9	1/5								

Naïve Bayes Classifiers Example

Classify following new instance:

Outlook = sunny

Temp = cool

Humidity = high

Wind = strong

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\arg\max}\ P(v_j)\prod_i P(a_i \mid v_j)$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\arg\max}\ P(v_j)P(outlook = sunny \mid v_j)P(temp = cool \mid v_j)P(humidity = high \mid v_j)P(wind = strong \mid v_j)$$

Naïve Bayes Classifiers Example

First calculate prior probabilities

$$P(play = yes) = 9/14$$

$$P(play = no) = 5/14$$

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes)$$

$$= 9/14 \times 2/9 \times 3/9 \times 3/9 \times 3/9 = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no)$$

$$= 5/14 \times 3/5 \times 1/5 \times 4/5 \times 3/5 = 0.0206$$

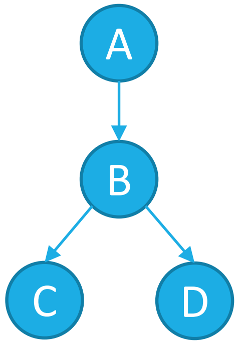
$$v_{NB} = \underset{v_j \in \{yes, no\}}{\arg\max}\ P(v_j)P(sunny \mid v_j)P(cool \mid v_j)$$

$$P(high \mid v_j)P(strong \mid v_j) = no$$

Bayesian Network

Bayesian Network

Directed acyclic graph



Set of tables for each node in the graph

A	P(A)
True	0.4
False	0.6

A	B	P(B|A)
True	True	0.3
True	False	0.7
False	True	0.99
False	False	0.01

B	C	P(C|B)
True	True	0.1
True	False	0.9
False	True	0.6
False	False	0.4

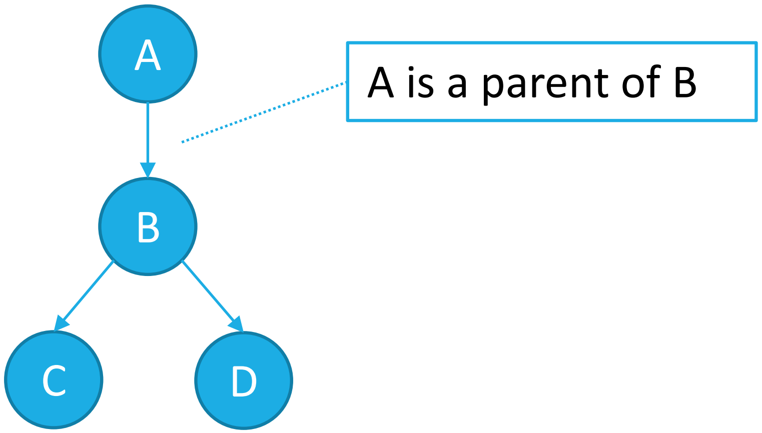
B	D	P(D|B)
True	True	0.95
True	False	0.05
False	True	0.98
False	False	0.02

Directed Acyclic Graph

Each node is a random variable

An arrow from node X to node Y

- Node X is a parent of node Y
- Node X has direct influence on node Y

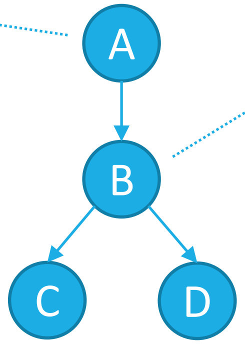


Set of Tables

Each node X_i has conditional probability distribution $P(X_i|Parent(X_i))$

- Quantify effect of parent on the node

A	P(A)
True	0.4
False	0.6



A	B	P(B|A)
True	True	0.3
True	False	0.7
False	True	0.99
False	False	0.01

B	C	P(C|B)
True	True	0.1
True	False	0.9
False	True	0.6
False	False	0.4

B	D	P(D|B)
True	True	0.95
True	False	0.05
False	True	0.98
False	False	0.02

Set of Tables

Conditional probability distribution for B given A

A	B	P(B|A)
True	True	0.3
True	False	0.7
False	True	0.99
False	False	0.01

For a given value of the parent

- Sum of entries for $P(X_i|\text{ConstantValue}) = 1$
- E.g. P(B = True | A = True) + P(B = False | A = True) = 0.3 + 0.7 = 1

  P(B = True | A = False) + P(B = False | A = False) = 0.99 + 0.01 = 1

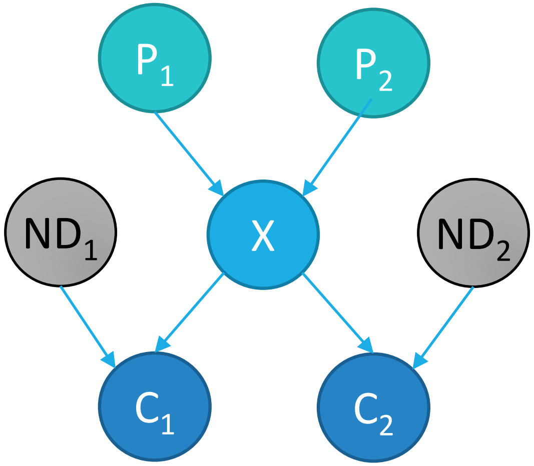
If a Boolean variable has k Boolean parents

- Table with 2^{k+1} rows of entries
- But only 2^k need to be stored

Conditional Independence

Given parents P_1 and P_2

- Node X is conditionally independent of
  - Its non descendants ND_1 and ND_2



Joint Probability Distribution

Joint probability distribution

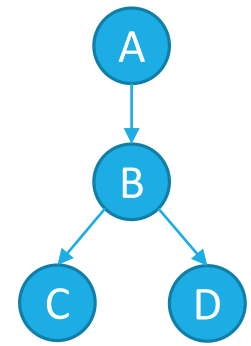
- Over all variables X_1, …, X_n

$$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v=1}^{n} \mathrm{P}(X_v = x_v \mid X_j = x_j \text{ for each } X_j \text{ which is a parent of } X_v)$$

Bayesian Network Example

Calculate joint probability

- A = True, B = True, C = True, D = True



$$P(A = \text{True}, B = \text{True}, C = \text{True}, D = \text{True})$$

$$= P(A = \text{True}) \times P(B = \text{True} \mid A = \text{True}) \times$$

$$P(C = \text{True} \mid B = \text{True}) \times P(D = \text{True} \mid B = \text{True})$$

From graph structure

$$= 0.4 \times 0.3 \times 0.1 \times 0.95$$

From tables

$$= 0.0114$$

A	P(A)
True	0.4
False	0.6

A	B	P(B|A)
True	True	0.3
True	False	0.7
False	True	0.99
False	False	0.01

B	C	P(C|B)
True	True	0.1
True	False	0.9
False	True	0.6
False	False	0.4

B	D	P(D|B)
True	True	0.95
True	False	0.05
False	True	0.98
False	False	0.02

Inference

Use Bayesian network to compute probability

Involves queries of the form

$$P(X|E)$$

- X = Query variable(s)
- E = Evidence variable(s)

Inference

Query example

- P(HasAnthrax = True | HasFever = True, HasCough = True)

HasDifficultyBreathing, HasWideMediastinum

- In Bayesian network
- Not given in query
- Not query variables nor evidence variables
- Treated as unobserved variables