March 24, 2017

Bayesian Learning

- Bayes Theorem
- MAP, ML hypotheses
- MAP learners
- Bayes optimal classifier
- Naive Bayes learner
- Bayesian belief networks

Bayesian Learning-Advantages

- Bayesian reasoning provides a probabilistic approach to inference.
- It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data.
- It is important to machine learning because it provides a quantitative approach to weighing the evidence supporting alternative hypotheses.
- Bayesian reasoning provides the basis for learning algorithms that directly manipulate probabilities, as well as a framework for analyzing the operation of other algorithms that do not explicitly manipulate probabilities.

Bayesian Learning -Relevance

- Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems.

- they provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$ = initial probability that hypothesis h holds before we have observed the training data (called *Prior Probability*).
- $P(D)$ = prior probability of training data D (prior probability that training data D will be observed (i.e., the probability of D given no knowledge about which hypothesis holds).
- $P(D|h)$ = probability of D given h (the probability of observing data D given some world in which hypothesis h holds.)
- $P(h|D)$ = probability of h given D (`posterior probability` of h: it reflects our confidence that h holds after we have seen the training data D)

Observation

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h|D)$ increases with P(h) and with $P(D|h)$ according to Bayes theorem.
- $P(h|D)$ decreases as P (D) increases, because the more probable it is that D will be observed independent of h, the less evidence D provides in support of h.

Choosing Hypothesis[1]

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis. We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis.

Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

[1] argmax f(x) $x \in X$: The value of x that maximises f(x), argmax x^2 = -3 where $x \in \{1, 2, -3\}$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)}$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.98 \times .008}{.0376} = .209$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.98 \times .008}{.0376} = .209$$
$$P(\neg cancer \mid +) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)}$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.98 \times .008}{.0376} = .209$$
$$P(\neg cancer \mid +) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)} = \frac{.03 \times .992}{.0376} = .791$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.98 \times .008}{.0376} = .209$$
$$P(\neg cancer \mid +) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)} = \frac{.03 \times .992}{.0376} = .791$$
$$P(+) = P(+ \mid c'r)P(c'r) + P(+ \mid \neg c'r)P(\neg c'r)$$

Bayes Theorem

Does patient have cancer or not?

> *A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.*

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(+|cancer) = .98 \quad P(-|cancer) = .02$$
$$P(+|\neg cancer) = .03 \quad P(-|\neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.98 \times .008}{.0376} = .209$$
$$P(\neg cancer \mid +) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)} = \frac{.03 \times .992}{.0376} = .791$$
$$P(+) = P(+ \mid c'r)P(c'r) + P(+ \mid \neg c'r)P(\neg c'r) = .0376$$

Most Probable Classification of New Instances

So far we've sought the most probable *hypothesis* given the data D (i.e., h_{MAP})

Given new instance x, what is its most probable *classification*?

- $h_{MAP}(x)$ is not the most probable classification!

Consider:

- Three possible hypotheses:

  $$P(h_1|D) = .4,\ P(h_2|D) = .3,\ P(h_3|D) = .3$$

- Given new instance x,

  $$h_1(x) = +,\ h_2(x) = -,\ h_3(x) = -$$

- What's most probable classification of x?

- Taking all hypotheses into account, the probability that x is positive is .4 (the probability associated with h i) , and
- The probability that it is negative is therefore 6.

- The most probable classification (negative) in this case is different from the classification generated by the MAP hypothesis.

- In general, the most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.

- If the possible classification of the new example can take on any value v j from some set V, then the probability $P(v_j \mid D)$ that the correct classification for the new instance is v;, is just

Bayes Optimal Classifier

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example:

Bayes Optimal Classifier

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example:

$$\begin{aligned} P(h_1|D) &= .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1 \\ P(h_2|D) &= .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0 \\ P(h_3|D) &= .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0 \end{aligned}$$

therefore

Bayes Optimal Classifier

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example:

$$\begin{aligned} P(h_1|D) &= .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1 \\ P(h_2|D) &= .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0 \\ P(h_3|D) &= .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0 \end{aligned}$$

therefore

$$\begin{aligned} \sum_{h_i \in H} P(+|h_i)P(h_i|D) &= .4 \\ \sum_{h_i \in H} P(-|h_i)P(h_i|D) &= .6 \end{aligned}$$

and

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

Naive Bayes Classifier

Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods.

When to use

- Moderate or large training set available
- Attributes that describe instances are conditionally independent given classification

Successful applications:

- Diagnosis
- Classifying text documents

Naive Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \ldots a_n \rangle$.
Most probable value of $f(x)$ is:

$$
\begin{aligned}
v_{MAP} &= \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n) \\
v_{MAP} &= \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \\
&= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)
\end{aligned}
$$

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

$$\textbf{Naive Bayes classifier: } v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

Naive Bayes Algorithm

Naive_Bayes_Learn(*examples*)

For each target value v_j

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value a_i of each attribute a

$\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance(x)

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Naive Bayes: Example

Traning Dataset

Age	Income	Student	Credit rating	Buys compter ?
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
$30 \ldots 40$	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
$31 \ldots 40$	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
$31 \ldots 40$	medium	no	excellent	yes
$31 \ldots 40$	high	yes	fair	yes
> 40	medium	no	excellent	no

Data Sample:

$X = (age \leq 30, Income = medium, Student = yes, Creditrating = fair)$

Class:

C1: Buys computer = 'yes'

C2: Buys computer = 'no'

Naive Bayes: Example

- Compute $P(X \mid Ci)$ for each class where **X = (age $\leq$ 30, Income = medium, Student = yes, Credit rating = fair)**

- $P(age =' < 30' \mid Buyscomputer =' yes') = 2/9 = 0.222$
- $P(age =' < 30' \mid Buyscomputer =' no') = 3/5 = 0.6$
- $P(income =' medium' \mid Buyscomputer =' yes') = 4/9 = 0.444$
- $P(income =' medium' \mid Buyscomputer =' no') = 2/5 = 0.4$
- $P(student =' yes' \mid Buyscomputer =' yes') = 6/9 = 0.667$
- $P(student =' yes' \mid Buyscomputer =' no') = 1/5 = 0.2$
- $P(Creditrating =' fair' \mid Buyscomputer =' yes') = 6/9 = 0.667$
- $P(Creditrating =' fair' \mid Buyscomputer =' no') = 2/5 = 0.4$

P(X | Ci) :
$P(X \mid Buyscomputer =' yes') = 0.222 \times 0.4444 \times 0.667 = 0.044$

$P(X \mid Buyscomputer =' no') = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$

P(X | Ci) $\times$ P(Ci) :
$P(X \mid Buyscomputer =' yes') \times P(Buyscomputer =' yes') = 0.028$ **(**$0.044 \times \frac{9}{14}$**)**

$P(X \mid Buyscomputer =' no') \times P(Buyscomputer =' no') = 0.007$ **(**$0.019 \times \frac{5}{14}$**)**
X belongs to class 'buys computer = 'yes'.

Naive Bayes: Subtleties

1. Conditional independence assumption is often violated

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

   - ...but it works surprisingly well anyway. Note don't need estimated posteriors $\hat{P}(v_j|x)$ to be correct; need only that

$$\underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) P(a_1 \ldots, a_n | v_j)$$

   - see [Domingos & Pazzani, 1996] for analysis
   - Naive Bayes posteriors often unrealistically close to 1 or 0

Naive Bayes: Subtleties

2. what if none of the training instances with target value v_j have attribute value a_i? Then

$$\hat{P}(a_i|v_j) = 0, \text{ and...}$$

$$\hat{P}(v_j)\prod_i \hat{P}(a_i|v_j) = 0$$

Typical solution is Bayesian estimate for $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

where

- n is number of training examples for which $v = v_j$,
- n_c number of examples for which $v = v_j$ and $a = a_i$
- p is prior estimate for $\hat{P}(a_i|v_j)$
- m is weight given to prior (i.e. number of "virtual" examples)

Learning to Classify Text

Why?

- Learn which news articles are of interest
- Learn to classify web pages by topic

Naive Bayes is among most effective algorithms

What attributes shall we use to represent text documents??

Learning to Classify Text

Target concept *Interesting*? : $Document \rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position i is w_k, given v_j

one more assumption: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

LEARN_NAIVE_BAYES_TEXT(*Examples*, V)

1. *collect all words and other tokens that occur in Examples*

- *Vocabulary* $\leftarrow$ all distinct words and other tokens in *Examples*

2. *calculate the required* $P(v_j)$ *and* $P(w_k|v_j)$ *probability terms*

- For each target value v_j in V do
  - $docs_j \leftarrow$ subset of *Examples* for which the target value is v_j
  - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
  - $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$
  - $n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)
  - for each word w_k in *Vocabulary*
    - $n_k \leftarrow$ number of times word w_k occurs in $Text_j$
    - $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

CLASSIFY_NAIVE_BAYES_TEXT(*Doc*)

- *positions* $\leftarrow$ all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB}, where

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

Bayesian Belief Networks

Interesting because:

- Naive Bayes assumption of conditional independence too restrictive
- But it's intractable without some such assumptions...
- Bayesian Belief networks describe conditional independence among *subsets* of variables

$\rightarrow$ allows combining prior knowledge about (in)dependencies among variables with observed training data

(also called Bayes Nets)

Conditional Independence

Definition: X *is* conditionally independent *of* Y *given* Z *if the probability distribution governing* X *is independent of the value of* Y *given the value of* Z*; that is, if*

$$(\forall x_i, y_j, z_k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

more compactly, we write

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

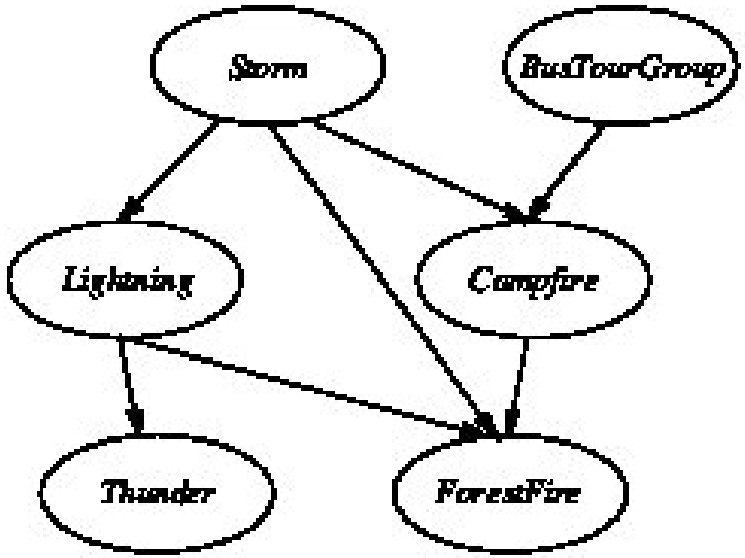
Example: *Thunder* is conditionally independent of *Rain*, given *Lightning*

$$P(\mathit{Thunder}|\mathit{Rain}, \mathit{Lightning}) = P(\mathit{Thunder}|\mathit{Lightning})$$

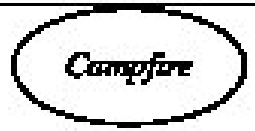
Naive Bayes uses cond. indep. to justify

$$\begin{aligned} P(X, Y|Z) &= P(X|Y, Z)P(Y|Z) \\ &= P(X|Z)P(Y|Z) \end{aligned}$$

Bayesian Belief Network



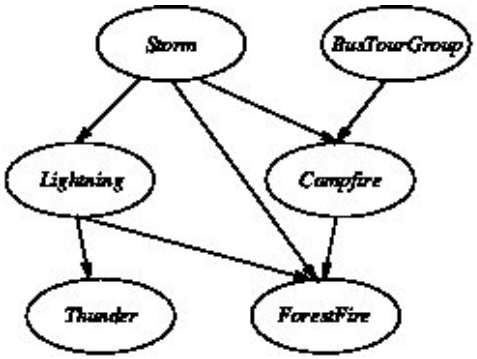
	S,B	$S,\neg B$	$\neg S,B$	$\neg S,\neg B$
C	0.4	0.1	0.8	0.2
$\neg C$	0.6	0.9	0.2	0.8



Network represents a set of conditional independence assertions:

- Each node is asserted to be conditionally independent of its nondescendants, given its immediate predecessors.
- Directed acyclic graph

Bayesian Belief Network



	S,B	S,¬B	¬S,B	¬S,¬B
C	0.4	0.1	0.8	0.2
¬C	0.6	0.9	0.2	0.8

Represents joint probability distribution over all variables

- e.g., $P(Storm, BusTourGroup, \ldots, ForestFire)$
- in general,

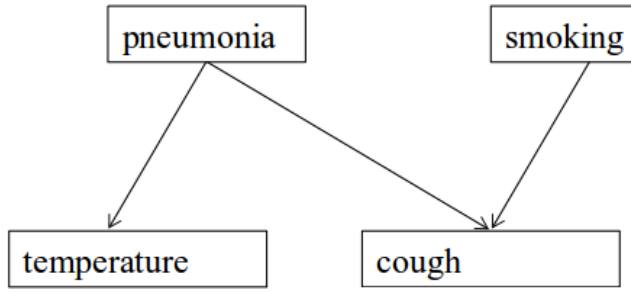
$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(y_i | Parents(Y_i))$$

  where $Parents(Y_i)$ denotes immediate predecessors of Y_i in graph
- so, joint distribution is fully defined by graph, plus the $P(y_i | Parents(Y_i))$

Example

pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

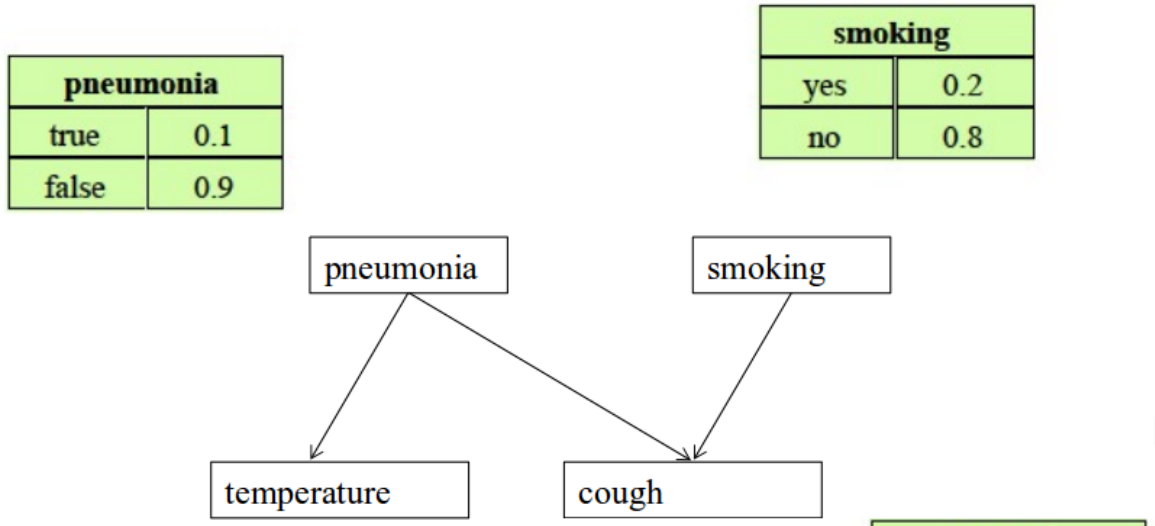


	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

		cough	
pneumonia	**smoking**	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(cough|smoking\ and\ pneumonia)$?

Example



pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

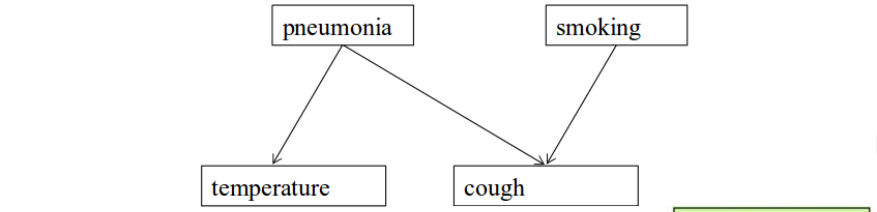
		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(cough|smoking\ and\ pneumonia)$?
  From table $P(C|S \wedge Pn) = .95$.

Example

pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

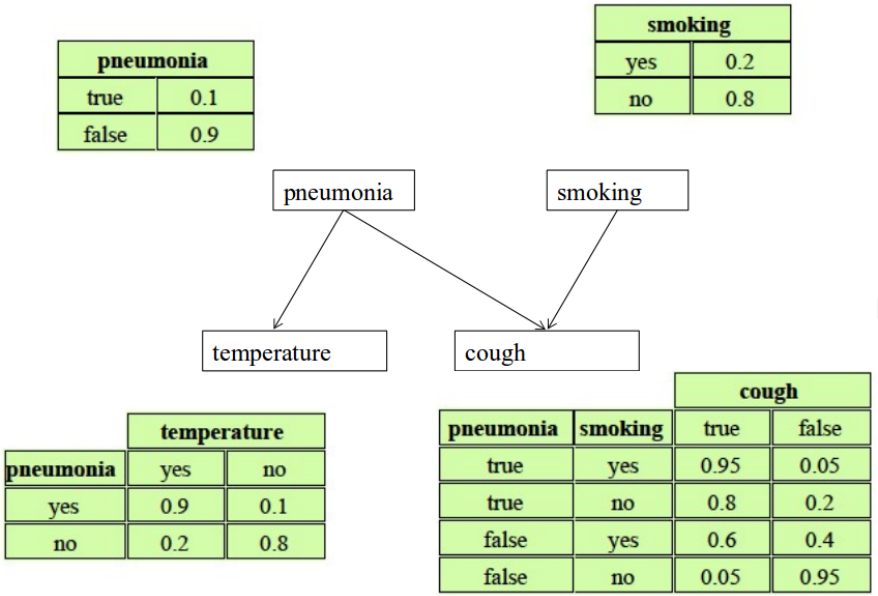


pneumonia	temperature	
	yes	no
yes	0.9	0.1
no	0.2	0.8

pneumonia	smoking	cough	
		true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?

Example



pneumonia	
true	0.1
false	0.9

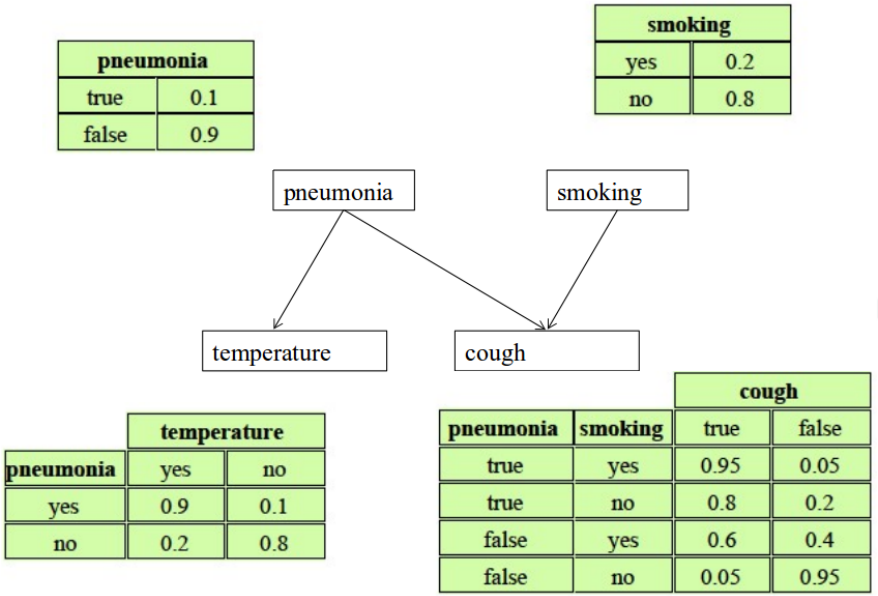
smoking	
yes	0.2
no	0.8

pneumonia	temperature	
	yes	no
yes	0.9	0.1
no	0.2	0.8

pneumonia	smoking	cough	
		true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$

Example



pneumonia	
true	0.1
false	0.9

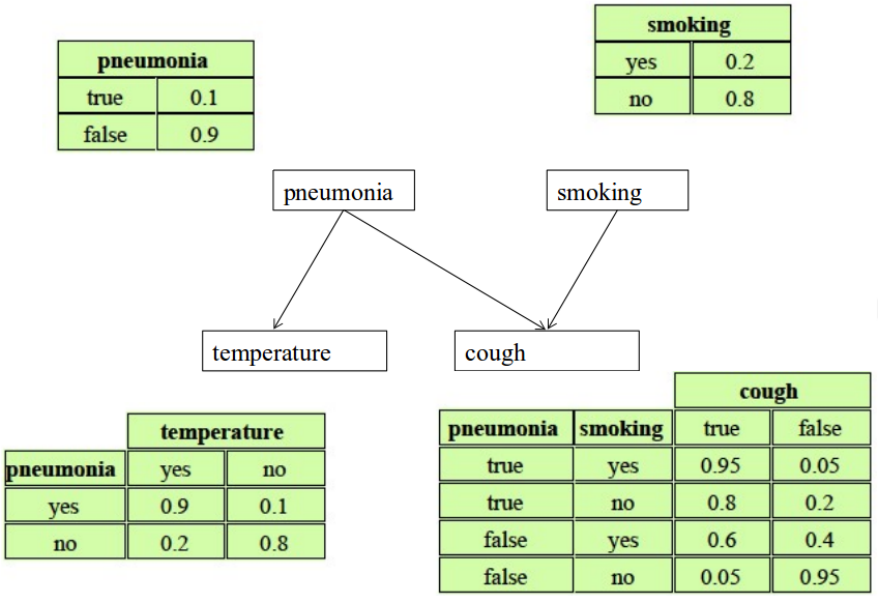
smoking	
yes	0.2
no	0.8

pneumonia	temperature	
	yes	no
yes	0.9	0.1
no	0.2	0.8

pneumonia	smoking	cough	
		true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$
  $P(C|S)P(S) = [P(C|S \wedge Pn)P(Pn) + P(C|S \wedge \neg Pn)P(\neg Pn)]P(S)$

Example



pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

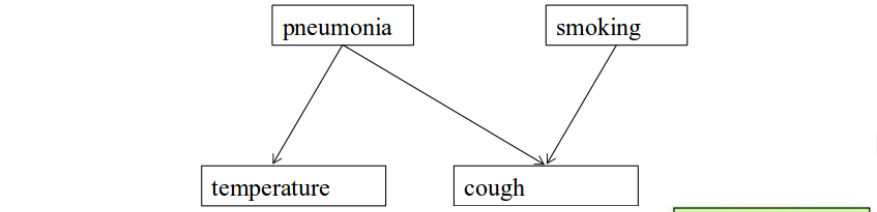
		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$
  $P(C|S)P(S) = [P(C|S \wedge Pn)P(Pn) + P(C|S \wedge \neg Pn)P(\neg Pn)]P(S)$
  $= [(.95)(.1) + (.6)(.9)](.2) = .127.$

Example

pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

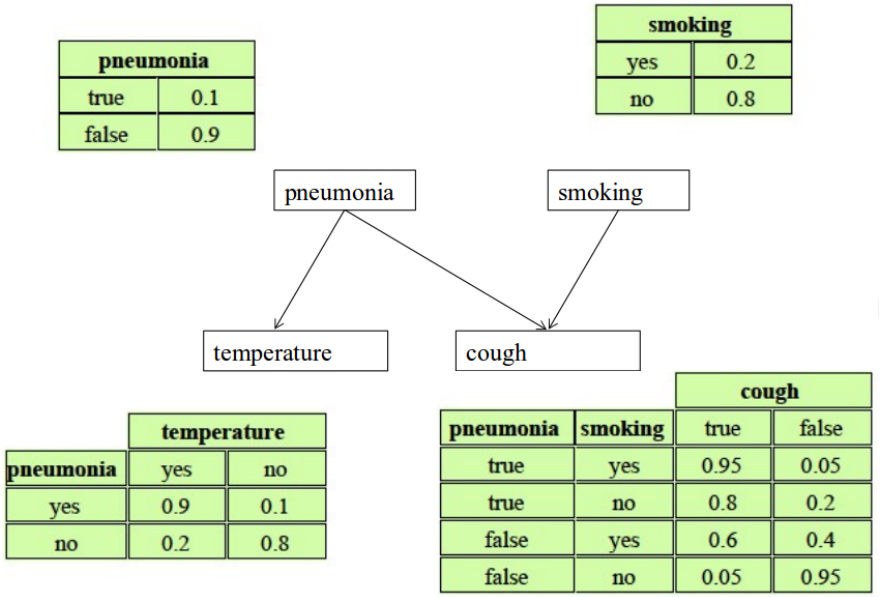


	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$
  $P(C|S)P(S) = [P(C|S \wedge Pn)P(Pn) + P(C|S \wedge \neg Pn)P(\neg Pn)]P(S)$
  $= [(.95)(.1) + (.6)(.9)](.2) = .127.$
  $P(C) = P(C|Pn \wedge S)P(Pn)P(S) + P(C|Pn \wedge \neg S)P(Pn)P(\neg S) +$
  $P(C|\neg Pn \wedge S)P(\neg Pn)P(S) + P(C|\neg Pn \wedge \neg S)P(\neg Pn)P(\neg S)$

Example



pneumonia	
true	0.1
false	0.9

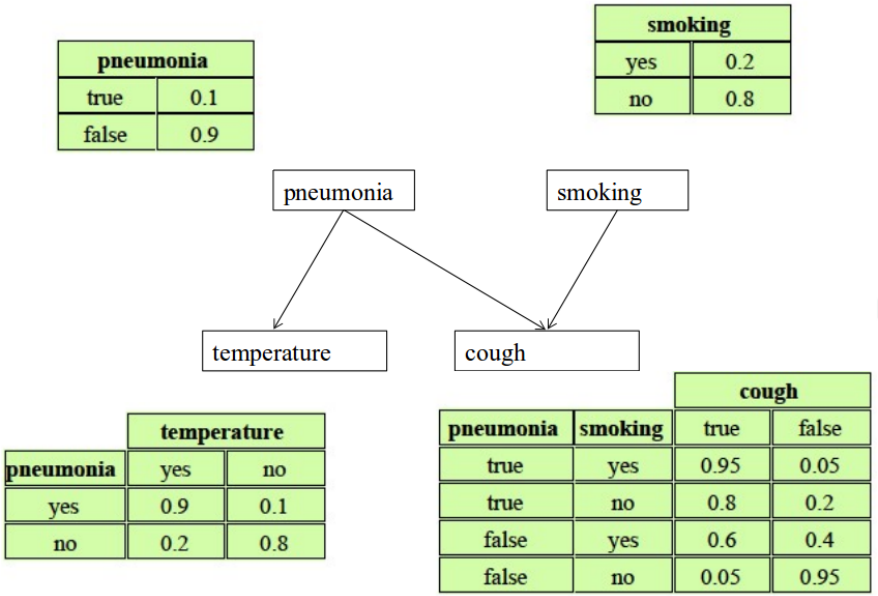
smoking	
yes	0.2
no	0.8

	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$
  $P(C|S)P(S) = [P(C|S \wedge Pn)P(Pn) + P(C|S \wedge \neg Pn)P(\neg Pn)]P(S)$
  $= [(.95)(.1) + (.6)(.9)](.2) = .127.$
  $P(C) = P(C|Pn \wedge S)P(Pn)P(S) + P(C|Pn \wedge \neg S)P(Pn)P(\neg S) +$
  $P(C|\neg Pn \wedge S)P(\neg Pn)P(S) + P(C|\neg Pn \wedge \neg S)P(\neg Pn)P(\neg S)$
  $= (.95)(.1)(.2) + (.8)(.1)(.8) + (.6)(.9)(.2) + (.05)(.9)(.8) = .227.$

Example



pneumonia	
true	0.1
false	0.9

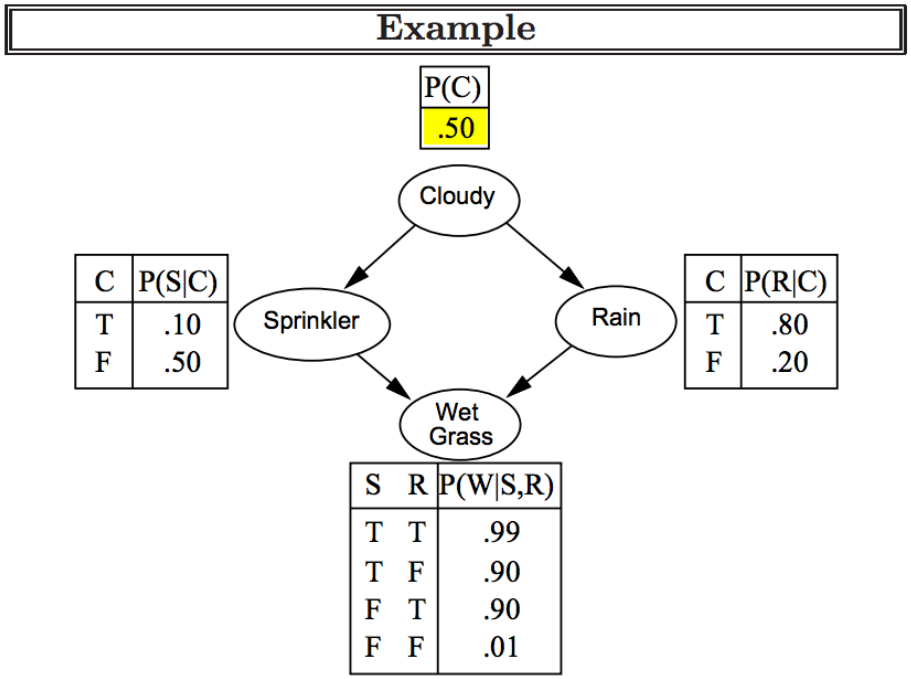
smoking	
yes	0.2
no	0.8

	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

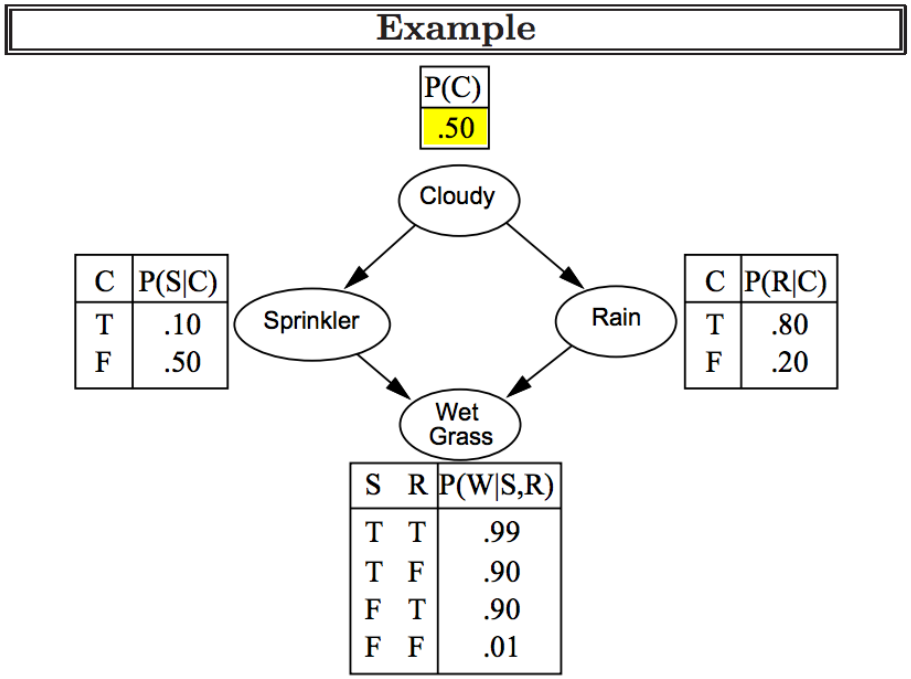
		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

- What is $P(smoking|cough)$?
  $P(S|C) = \frac{P(C|S)P(S)}{P(C)}$
  $P(C|S)P(S) = [P(C|S \wedge Pn)P(Pn) + P(C|S \wedge \neg Pn)P(\neg Pn)]P(S)$
  $= [(.95)(.1) + (.6)(.9)](.2) = .127.$
  $P(C) = P(C|Pn \wedge S)P(Pn)P(S) + P(C|Pn \wedge \neg S)P(Pn)P(\neg S) +$
  $P(C|\neg Pn \wedge S)P(\neg Pn)P(S) + P(C|\neg Pn \wedge \neg S)P(\neg Pn)P(\neg S)$
  $= (.95)(.1)(.2) + (.8)(.1)(.8) + (.6)(.9)(.2) + (.05)(.9)(.8) = .227.$
  $\mathbf{P(S|C) = \frac{.127}{.227} = .56.}$

Yet Another Example

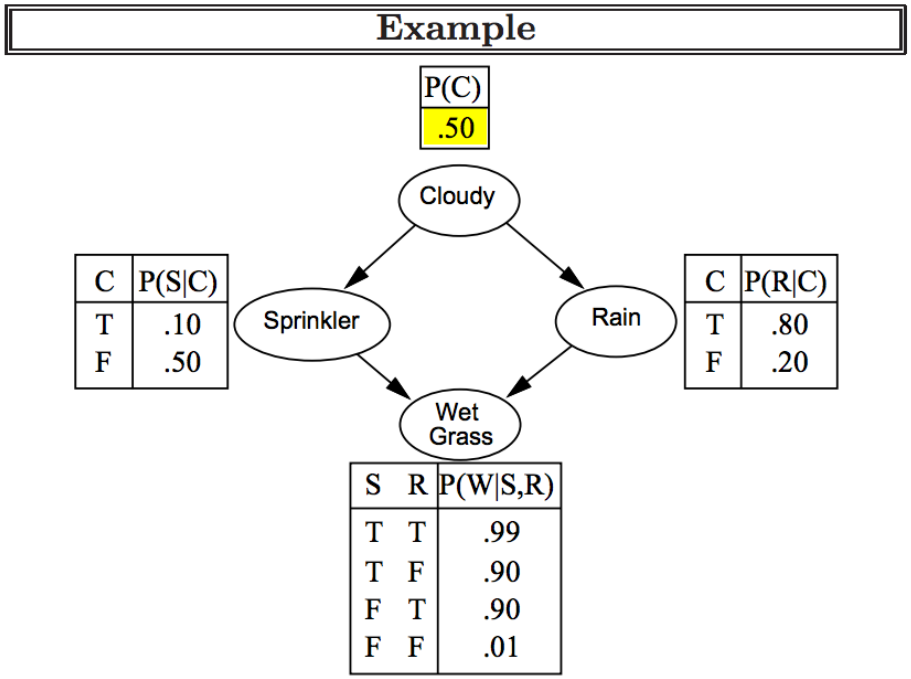


Yet Another Example



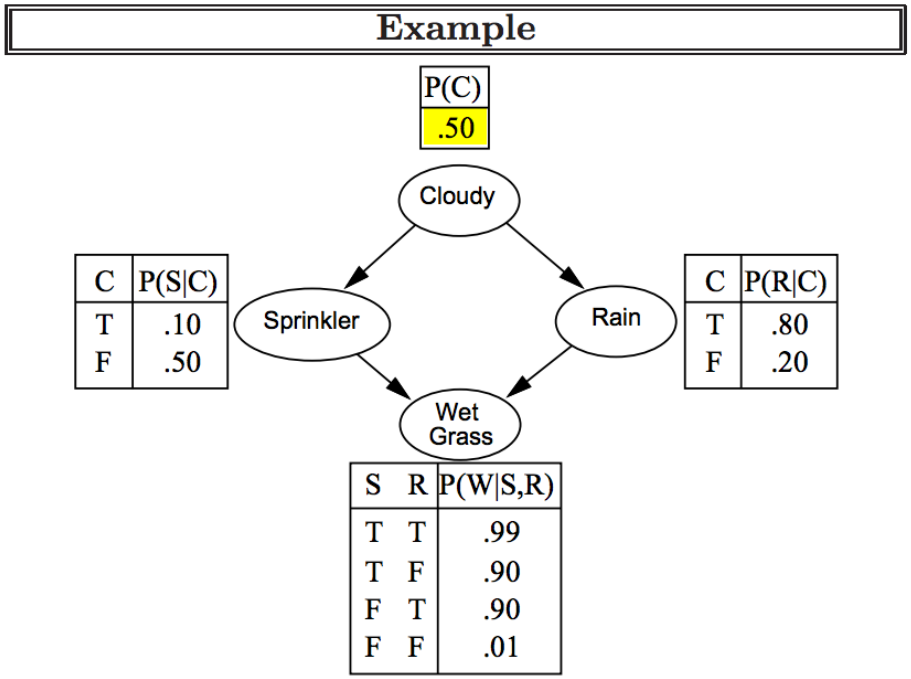
- What is $P(C, R, \neg S, W)$?

Yet Another Example

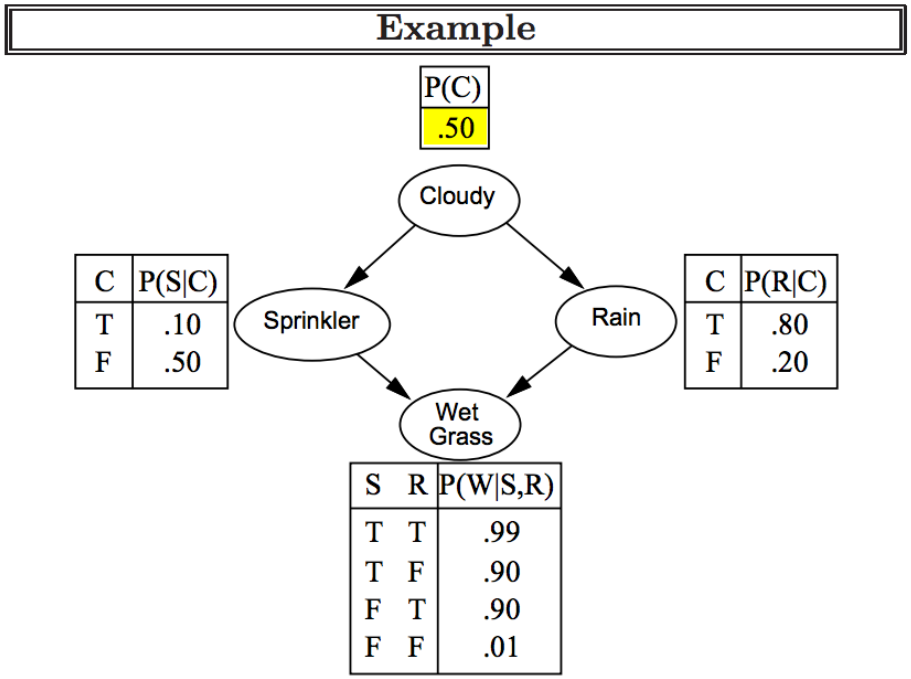


- What is $P(C, R, \neg S, W)$?
  $P(C, R, \neg S, W) = P(C)P(R|C)P(\neg S|C)P(W|R, \neg S$

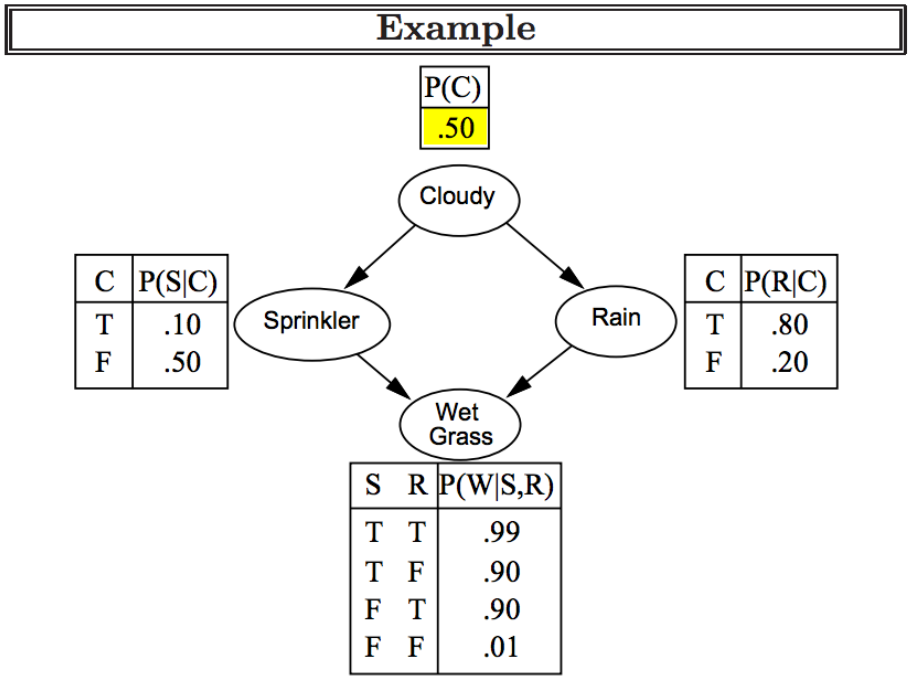
Yet Another Example



- What is $P(C, R, \neg S, W)$?
  $P(C, R, \neg S, W) = P(C)P(R|C)P(\neg S|C)P(W|R, \neg S = (.5)(.8)(.9)(.9) = .324$.

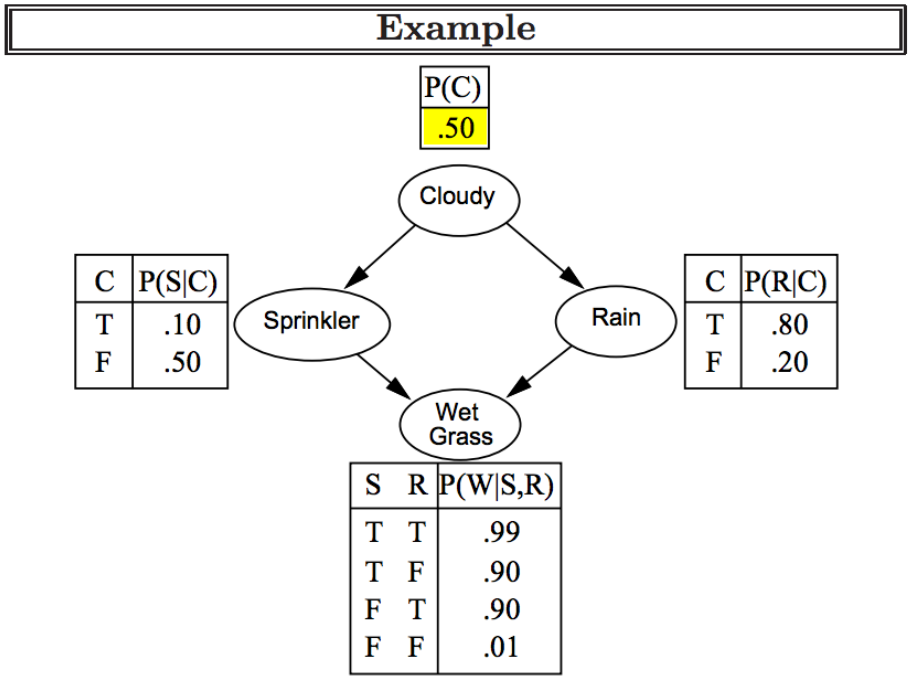


- Suppose you observe it is cloudy and raining. What is the probability that the grass is wet ?



- Suppose you observe it is cloudy and raining. What is the probability that the grass is wet ?
  Since *wet grass* is conditionally independent of *cloudy* given *rain* and *spinkler*, we have

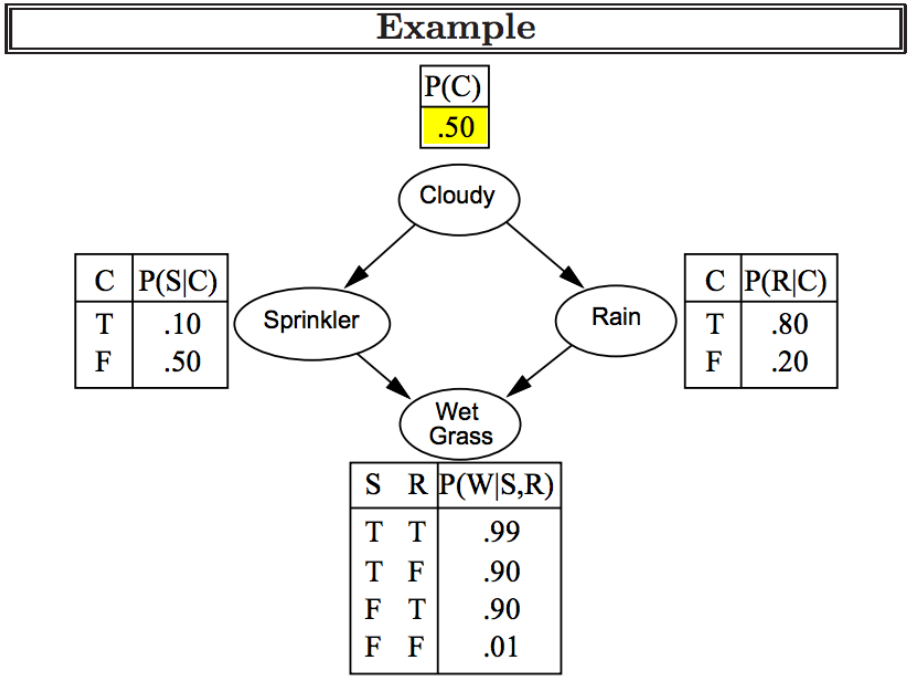
$$P(W|C,R) = P(W|R,S)P(S|C) + P(W|R,\neg S)P(\neg S|C)$$



- Suppose you observe it is cloudy and raining. What is the probability that the grass is wet ?
  Since *wet grass* is conditionally independent of *cloudy* given *rain* and *spinkler*, we have

$$P(W|C,R) = P(W|R,S)P(S|C) + P(W|R,\neg S)P(\neg S|C)$$

$$P(W|C,R) = (.99)(.1) + (.9)(.9) = .909.$$



- Suppose you observe the spinkler to be on and the grass is wet. What is the probability that it is raining ?
- Suppose you observe that the grass is wet and it is raining. What is the probability that it is cloudy ?

Inference in Bayesian Networks

How can one infer the (probabilities of) values of one or more network variables, given observed values of others?

- Bayes net contains all information needed for this inference
- If only one variable with unknown value, easy to infer it
- In general case, problem is NP hard

In practice, can succeed in many cases

- Exact inference methods work well for some network structures
- Monte Carlo methods "simulate" the network randomly to calculate approximate solutions

Learning of Bayesian Networks

Several variants of this learning task

- Network structure might be *known* or *unknown*
- Training examples might provide values of *all* network variables, or just *some*

If structure known and observe all variables

- Then it's easy as training a Naive Bayes classifier

Learning Bayes Nets

Suppose structure known, variables partially observable

e.g., observe *ForestFire, Storm, BusTourGroup, Thunder*, but not *Lightning, Campfire*...

- Similar to training neural network with hidden units
- In fact, can learn network conditional probability tables using gradient ascent!
- Converge to network h that (locally) maximizes $P(D|h)$

Summary: Bayesian Belief Networks

- Combine prior knowledge with observed data
- Impact of prior knowledge (when correct!) is to lower the sample complexity
- Active research area
  - Extend from boolean to real-valued variables
  - Parameterized distributions instead of tables
  - Extend to first-order instead of propositional systems
  - More effective inference methods
  - ...