CS(STAT)5525 : Data Analytics
Lecture : 10 Bayesian Classifier

Reza Jafari, Ph.D

Collegiate Associate Professor
rjafari@vt.edu

Bayes Theorem

- A probabilistic framework for solving classification problem:
- Conditional Probability

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

- Bayes theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Using Bayes Theorem for classification

- Consider each attribute and class label as random variables.
- Given a record with attributes $(X_1, X_2, ..., X_d)$, the goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes $P(Y|X_1, X_2, ..., X_d)$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Using Bayes Theorem for classification

- Approach
  - Compute the posterior probability $P(Y|X_1, X_2, ...X_d)$ using the Bayes theorem:

$$P(Y|X_1, X_2, ...X_d) = \frac{P(X_1, X_2, ...X_d|Y)P(Y)}{P(X_1, X_2, ...X_d)}$$

  - Maximum a-posteriori: choose Y that maximizes

$$P(Y|X_1, X_2, ...X_d)$$

  - Equivalent to choosing value of Y that maximizes

$$P(X_1, X_2, ...X_d|Y)P(Y)$$

- How to estimate

$$P(X_1, X_2, ...X_d|Y)$$

Naïve Bayes Classifier

- Assume conditional independence among attributes X_i when class is given:

$$P(X_1, X_2, ...X_d|Y_j) = P(X_1|Y_j)P(X_2|Y_j)...P(X_d|Y_j)$$

- Now we can estimate $P(X_i|Y_j)$ for all X_i and Yj combinations from the training dataset.
- New point is classified to Y_j if $P(Y_j)\prod P(X_i|Y_j)$ is maximal.

Example Data

- Given a Test Record:

$$X = (Redund = No, Divorced, Income = 120K)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Using Bayes Theorem:

$$P(Yes|X) = \frac{P(X|Yes)P(Yes)}{P(X)}$$

$$P(No|X) = \frac{P(X|No)P(No)}{P(X)}$$

where

- $P(X) = P(X|Yes)P(Yes) + P(X|No)P(No)$
- How to estimate $P(X|Yes)$ & $P(X|No)$?

Example

- Considering **conditional independency**:
- $$P(X|Yes) = P(Refund = No|Yes) \times P(Divorced = Yes|Yes)... \\ \times P(Income = 120K|Yes)$$
- $$P(X|No) = P(Refund = No|No) \times P(Divorced = Yes|No)... \\ \times P(Income = 120K|No)$$

Estimate Probabilities from Data

- Class: $P(Y_j) = \frac{N_j}{N}$ For example:
  - $P(No) = \frac{7}{10}$
  - $P(Yes) = \frac{3}{10}$
- For categorical attributes:

$$P(X_i = x_k | Y_j) = \frac{|X_{ik}|}{N_j}$$

- where $|X_{ik}|$ is number of instances having attribute value $X_i = x_k$ and belonging yo class Y_j
- For example:

$$P(Status = Married | No) = \frac{4}{7}$$

$$P(Refund = Yes | Yes) = 0$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Estimate Probabilities form Data

- For continuous attributes:

Discretization

- Partition the range into bins
- Replace continuous value with bin value
- Attribute changed from continuous to ordinal

Probability density estimation

- Assume attribute follows a Normal distribution
- Use data to estimate parameters of distribution (e.g mean and standard deviation)
- Once the probability distribution os known, use it to estimate the conditional probability $P(X_i|Y)$

Estimate Probabilities from Data

- Normal distribution:

$$P(X_i|Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (X_i, Y_j)
- For (Income, Class = No), if Class = No
  - sample mean = 110
  - sample variance = 2975

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi(2975)}} e^{\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

- Given a Test record:

$$x = (Refund = No, Divorced, Income = 120K)$$

P(No) = 7/10, P(Yes) = 3/10
P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(x | No) = P(Refund=No | No)
  × P(Divorced | No)
  × P(Income=120K | No)
  = 4/7 × 1/7 × 0.0072 = 0.0006

- P(x | Yes) = P(Refund=No | Yes)
  × P(Divorced | Yes)
  × P(Income=120K | Yes)
  = 1 × 1/3 × 1.2 × 10^{-9} = 4 × 10^{-10}

Since P(x|No)P(No) > P(x|Yes)P(Yes)

Therefore P(No|x) > P(Yes|x)
=> Class = No

Incrementally Updating Predictions

- Given a Test record:

$$x = (Refund = No, Divorced, Income = 120K)$$

P(No) = 7/10, P(Yes) = 3/10
P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10
- P(Yes | Divorced) = 1/3 x 3/10 / P(Divorced)
  P(No | Divorced) = 1/7 x 7/10 / P(Divorced)
- P(Yes | Refund = No, Divorced) = 1 x 1/3 x 3/10 / P(Divorced, Refund = No)
  P(No | Refund = No, Divorced) = 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No)

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

P(No) = 7/10, P(Yes) = 3/10
P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes | Married) = **0** x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

If one of the conditional probabilities is zero, then the entire expression becomes zero

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

m-estimate of probability

- Note that we estimated conditional probabilities $\Pr(A_i|C)$ by $\frac{N_{ic}}{N_c}$ where N_{ic} is the number of instances having attribute A_i in class c and N_c is number of instances in class c.
- This can cause trouble if $N_{ic} = 0$ and need to use other estimates of conditional probabilities that simple fraction.
- Laplace smoothing:

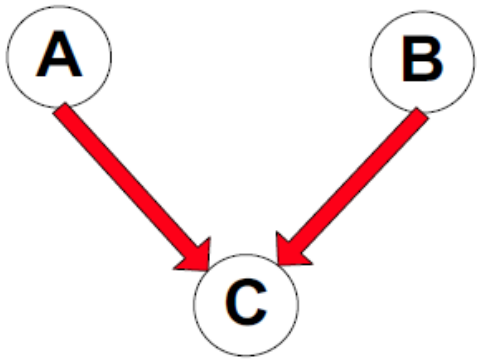
$$P(A_i|C) = \frac{N_{ic} + \alpha}{N_c + p * \alpha}$$

where p is number of classes, α smoothing parameter (most of the time $\alpha = 1$).

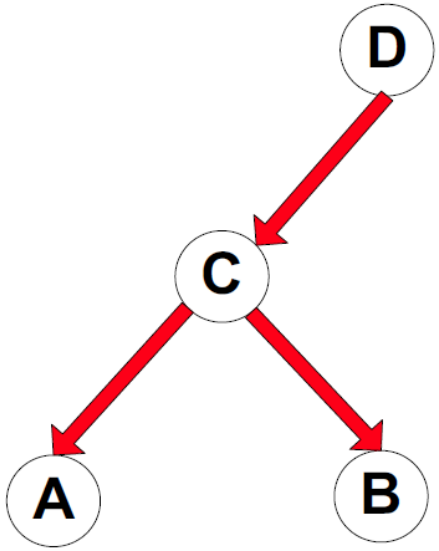
- As α increases, the likelihood probability moves toward uniform distribution (0.5) in the binary classification.

Bayesian Networks

- Provides graphical representation of probabilistic relationships among a set of random variables.
- Consists of
  - A directed acyclic graph (DAG)
  - Corresponds to dependance relationship between a pair of variables.
- A probability table associating each node to its immediate parents(s)

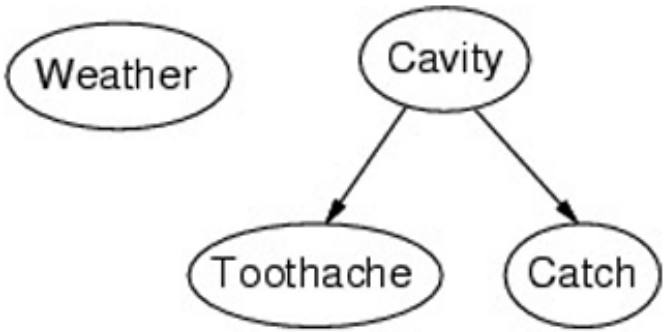


Conditional Independence



- D is parent of C.
- A is child of C.
- B is descendant of D.
- D is ancestor of A.
- The semantics of a Bayesian Networks are simple: Each node is conditionally independent from its non-descendant given its parent.

Bayesian Networks: Structure



- Nodes: random variables.
- Arcs: interactions
  - An arrow from one variable to another indicates direct casual influence of variables #1 on variable #2
  - Must form a direct, acyclic graph.

Conditional independence and the joint distribution

- Key properties: each node is conditionally independent of its non-descendants given its parents
- Suppose the nodes $X_1, X_2, ..., X_n$ are sorted in topological order.
- To get the joint distribution $P(X_1, X_2, ..., X_n)$ use the chain rule.

Chain rule

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, ..., X_{i-1})$$
$$= \prod_{i=1}^{n} P(X_i | parent(X_i))$$

Example: Los Angles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm.

- What are the random variables?

- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Los Angles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm.
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?

- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Los Angles Burglar Alarm

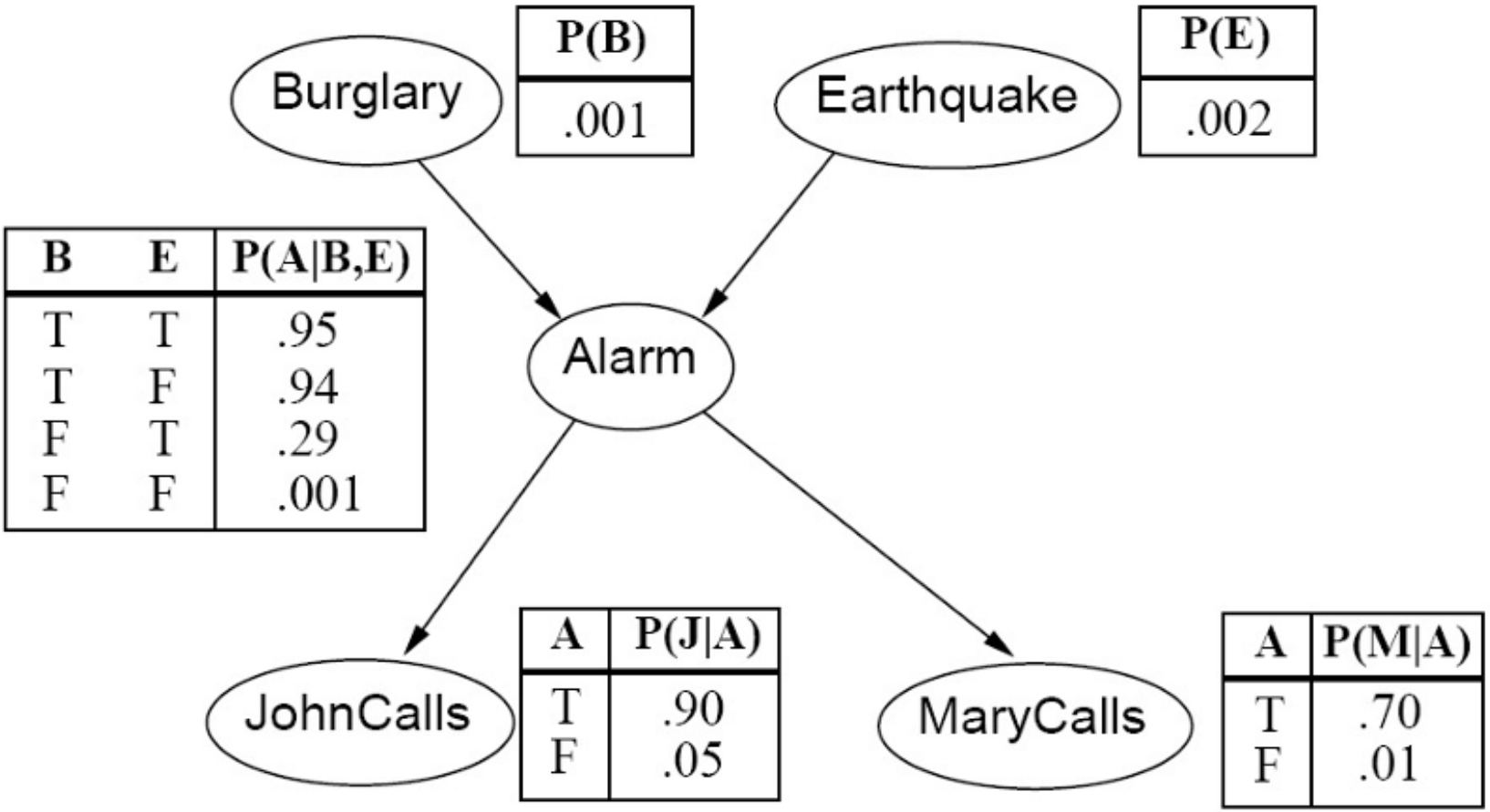
- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm.
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?

- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Los Angles Burglar Alarm

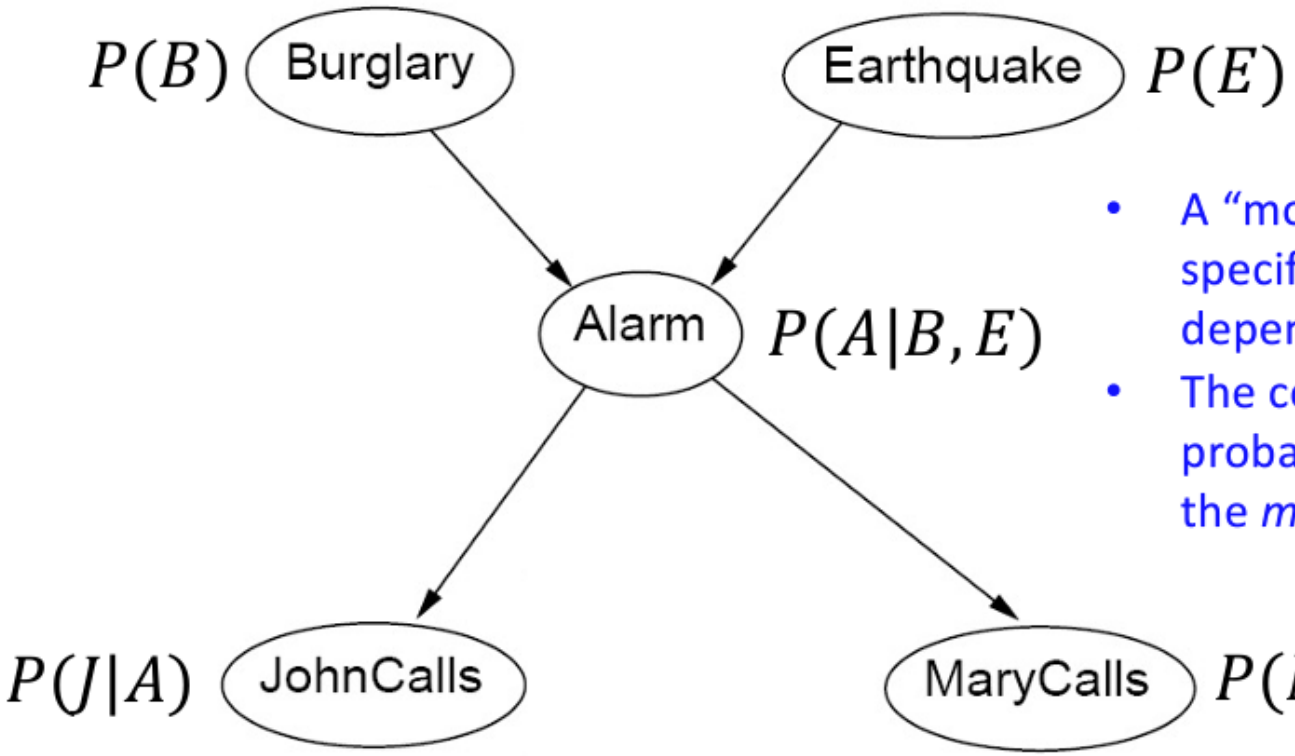
- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm.
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - Burglary, Earthquake, Alarm, JohnCalls, MaryCalls
- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Burglar Alarm



Example: Burglar Alarm

Example: Burglar Alarm



- A "model" is a complete specification of the dependencies.
- The conditional probability tables are the *model parameters.*

Classification using probabilities

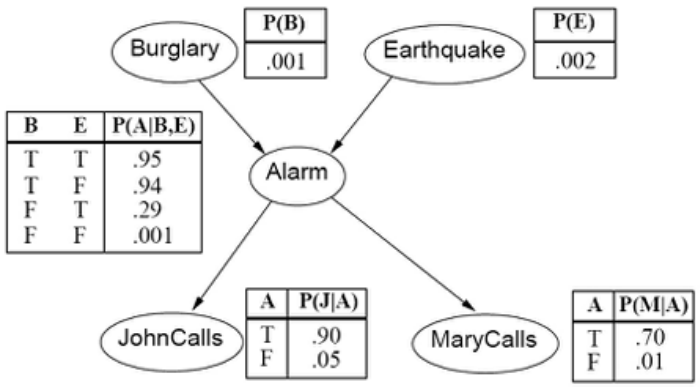
- Suppose Mary has called to tell you that you had a burglar alarm. Should you call the police?
  - Make a decision that maximizes the probability of being correct This is called a MAP (maximum a posteriori) decision. You decide that you have a burglar in your house if and only if

$$P(Burglary|Mary) > P(\overline{Burglary}|Mary)$$

Using Bayes Network to estimate a posteriori probabilities

- We don't know $P(B|M)$! we have to figure out what it is.
- This is called inference
- *First step* : find the joint probability of B (and $\overline{B}$), M (and $\overline{M}$) and any other variables that are necessary in order to link these two together.

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

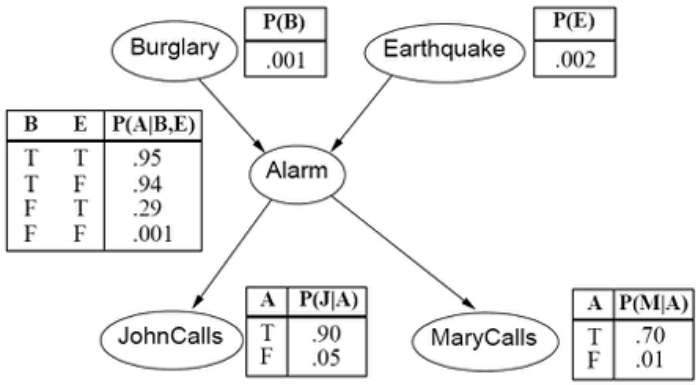


$P(BEAM)$	$\neg M, \neg A$	$\neg M, A$	$M, \neg A$	M, A
$\neg B, \neg E$	0.986045	2.99×10^{-4}	9.96×10^{-3}	6.98×10^{-4}
$\neg B, E$	1.4×10^{-3}	1.7×10^{-4}	1.4×10^{-5}	4.06×10^{-4}
$B, \neg E$	5.93×10^{-5}	2.81×10^{-4}	5.99×10^{-7}	6.57×10^{-4}
B, E	9.9×10^{-8}	5.7×10^{-7}	10^{-9}	1.33×10^{-6}

Using Bayes Network to estimate a posteriori probabilities

- *Second step* : Marginalize (add) to get ride of the variables you don't care about.

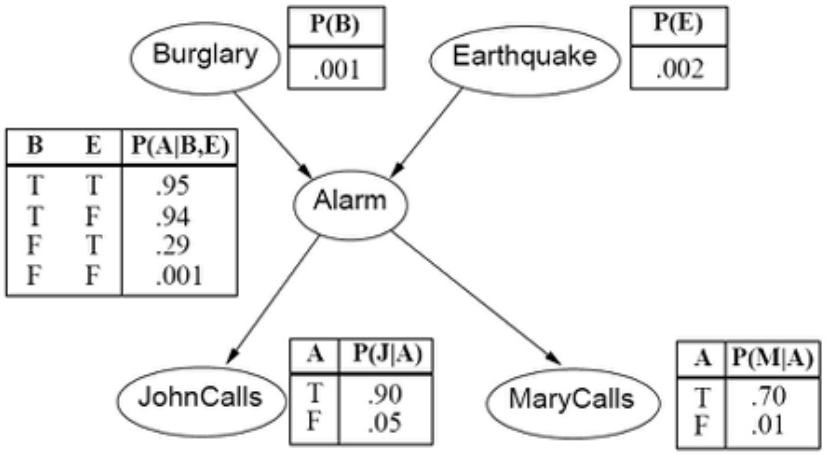
$$P(B, M) = \sum_{e \in F,T} \sum_{a \in F,T} P(B, E = e, A = a, M)$$



$P(B, M)$	$\neg M$	M
$\neg B$	0.987922	0.011078
B	0.000341	0.000659

Using Bayes Network to estimate a posteriori probabilities

- *Third step* : Ignore (delete) the column that did not happen.

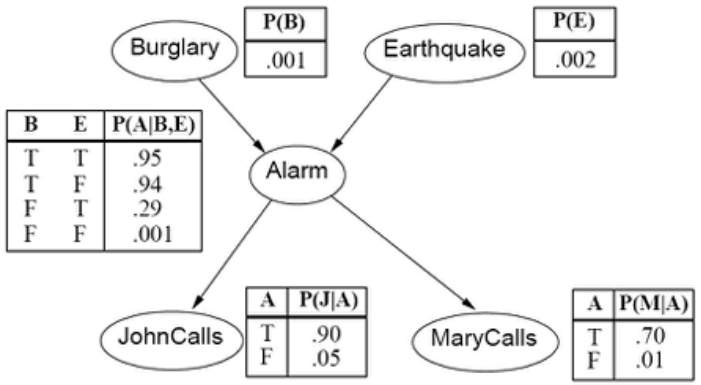


$P(B, M)$	M
$\neg B$	0.011078
B	0.000659

Using Bayes Network to estimate a posteriori probabilities

- *Fourth step* : Use the definition of conditional probability

$$P(B|M) = \frac{P(B, M)}{P(B, M) + P(\overline{B}, M)}$$

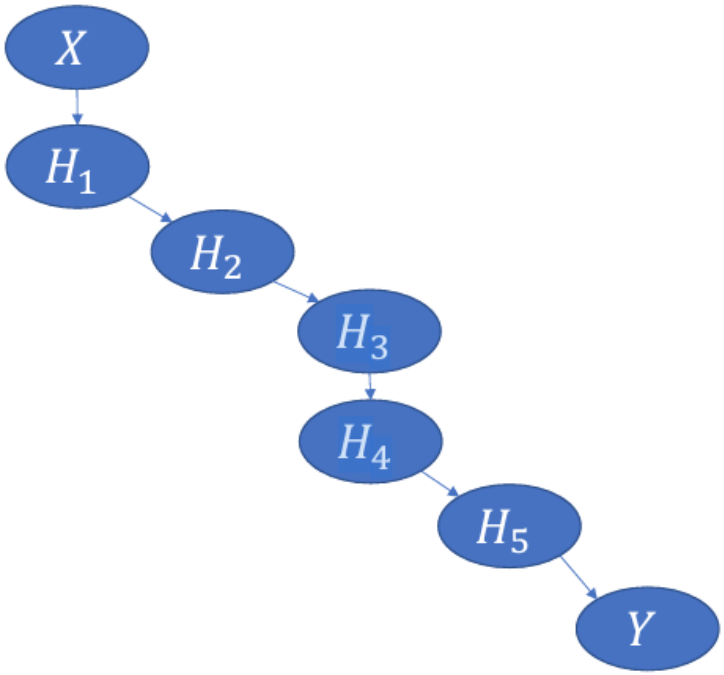


$P(B\|M)$	M
$\neg B$	0.943883
B	0.056117

- Since $P(B|M) < P(\overline{B}|M)$ hence the burglary is unlikely if only Mary calls or John calls,
- The probability of burglary is still only about 5%

The general Algorithm

- Given an arbitrary Bayes net, you want to find the joint probability of two variables, , X and Y, that are connected by a chain of intermediate variables, H_1 through H_N



Initialize:

Start with P(X)

Iterate:

1. PRODUCT: Multiply in the next variable
2. SUM: Marginalize out any variables you no longer need

Terminate:

When you have P(X,Y)

The general Algorithm

$$
\begin{aligned}
P(X, H_1) &= P(X)P(H_1|X) \\
P(X, H_1, H2) &= P(X, H_1)P(H_2|H_1) \\
P(X, H_2) &= \sum_{h_1} P(X, H_1 = h_1, H_2) \\
P(X, H2, H_3) &= P(X, H_2)P(H_3|H_2) \\
P(X, H_3) &= \sum_{h_2} P(X, H_2 = h_2, H_3) \\
&\vdots \\
P(X, H_4, H5) &= P(X, H_4)P(H_5|H_4) \\
P(X, H_5) &= \sum_{h_4} P(X, H_4 = h_4, H_5) \\
P(X, H_5, Y) &= P(X, H_5)P(Y|H_5) \\
P(X, Y) &= \sum_{h_5} P(X, H_5 = h_5, Y)
\end{aligned}
$$

Play tennis dataset

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Example

- Using the Naiive Bayes classifier, find out if a person will play tennis if the weather is $x_1 =$ Sunny, $x_2 =$ Cool , $x_3 =$ High, $x_4 =$Strong

- We have to calculate the following two conditional probabilities:

$$P(Yes|x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4|Yes)}{P(x_1, x_2, x_3, x_4)} = \frac{\prod_{i=1}^{i=4} P(x_i|Yes) * P(Yes)}{P(x_1, x_2, x_3, x_4)}$$

$$P(No|x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4|No)}{P(x_1, x_2, x_3, x_4)} = \frac{\prod_{i=1}^{i=4} P(x_i|No) * P(No)}{P(x_1, x_2, x_3, x_4)}$$

- Since the denominators are identical hence you need to compare the numerator for the classification .

Example

- Using the Naiive Bayes classifier, find out if a person will play tennis if the weather is $x_1 =$ Sunny, $x_2 =$ Cool , $x_3 =$ High, $x_4 =$Strong
- Let the attributes in the dataset to be defined as : $X_1 = Oulook$, $X_2 = Temprerature$, $X_3 = Humidity$, $X_4 = Wind$
- We have to calculate the following two conditional probabilities:

$$P(Yes|x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4|Yes)}{P(x_1, x_2, x_3, x_4)} = \frac{\prod_{i=1}^{i=4} P(x_i|Yes) * P(Yes)}{P(x_1, x_2, x_3, x_4)}$$

$$P(No|x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4|No)}{P(x_1, x_2, x_3, x_4)} = \frac{\prod_{i=1}^{i=4} P(x_i|No) * P(No)}{P(x_1, x_2, x_3, x_4)}$$

- Since the denominators are identical hence you need to compare the numerator for the classification .

Example

- We can estimate each term using the dataset , for example:
  - $\Pr(\text{Yes}) = \frac{9}{14}$
  - $\Pr(\text{No}) = \frac{5}{14}$
  - $\Pr(\text{Outlook=Sunny}|\text{Yes}) = \frac{2}{9}$
  - $\Pr(\text{Outlook=Sunny}|\text{No}) = \frac{3}{5}$

$$\prod_{i=1}^{i=4} P(x_i|Yes) * P(Yes) = \frac{2}{9}.\frac{3}{9}.\frac{3}{9}.\frac{3}{9}.\frac{9}{14} = 0.0053$$

$$\prod_{i=1}^{i=4} P(x_i|Yes) * P(Yes) = \frac{3}{5}.\frac{1}{5}.\frac{4}{5}.\frac{3}{5}.\frac{5}{14} = 0.0206$$

- We thus predict that NO is the output.