Probability, Naïve Bayes Model, Bayesian networks

Note: this material partly contains the slides provided by Prof. Padhraic Smyth

Bayes' Rule

- Bayes' Rule
  - $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$

Bayes' Rule

- Bayes' Rule

  P(Aㅅ B) = p(A) P(B|A)
  P(Aㅅ B)/P(A) = P(B|A) = P(B)P(A|B)
  { P(A)P(B|A) } /P(B) = P(A|B)

  - P(A | B) = P(B | A) P(A) / P(B) 역추론을 할 수 있다는 것
  - P(disease | symptom) 증상이 있으면 감기가 있을 확률은?

  = P(symptom | disease) P(disease)/ P(symptom)

  감기가 있으면 두통이 있을 확률

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)

  = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE (광우병)
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US

Bayes' Rule

- Bayes' Rule
  - $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$
  - P(disease | symptom)

    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)

Bayes' Rule

$P(B|A) = \{ P(B)P(A|B) \} / P(A) = \{ P(B)P(A|B) \} / \{ P(B)P(A|B) + P(B)P(A|B) \}$

- Bayes' Rule
  - $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$
  - P(disease | symptom)

    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)
- Conditioning
  - $P(A) = P(A \mid B)\, P(B) + P(A \mid \neg B)\, P(\neg B)$

Bayes' Rule

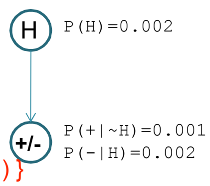
- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
  = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)
- Conditioning
  - $P(A) = P(A \mid B)\, P(B) + P(A \mid \neg B)\, P(\neg B)$
    $= P(A \wedge B) + P(A \wedge \neg B)$

7

Bayes' Rule (Revisited!)

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- ⇨ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) / P(b)$ P(A|B), P(B|A) 다름을 알기

- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis(뇌수막염), *S* be stiff neck,
    P(m)=0.01%, P(s|m)=80%, P(s)=10%
    Then, stiff neck 증상이 있는 사람 중 뇌수막염이 있을 확률
    $P(m|s) = P(s|m)\, P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008 = 0.08\%$
  - Note: posterior probability of meningitis still very small!



- Another Example:
  - False Positive & False Negative for Hepatitis C Test
  - E.g., let P(H)=0.2% and P(+|~H)=1%, P(-|H)=2%
    Then, 맞으면 맞다 하고 틀리면 틀리다 할 확률
    P(H|+)=? P(H|+) = { P(H) P(+|H) } / P(+)
    = { P(H) P(+|H) } / { P(H)P(+|H) + P(~H)(+|~H) }
    = (0.002 * 0.98) / (0.002 * 0.98) + (0.998 * 0.01)

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$

11

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$

12

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Independence is essential for efficient probabilistic reasoning

13

Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Independence is essential for efficient probabilistic reasoning
- A and B are conditionally independent given C iff
  - $P(A \mid B, C) = P(A \mid C)$

14

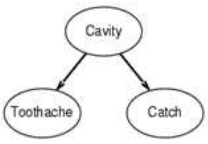
Independence

- A and B are independent iff
  - $P(A \wedge B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Independence is essential for efficient probabilistic reasoning
- A and B are conditionally independent given C iff
  - $P(A \mid B, C) = P(A \mid C)$ A,B는 서로간에 영향을 미치지 않는다
  - $P(B \mid A, C) = P(B \mid C)$
  - $P(A \wedge B \mid C) = P(A \mid C) \cdot P(B \mid C)$

15

Examples of Conditional Independence

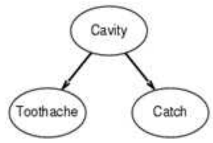
- Toothache (T)
- Spot in Xray (X)
- Cavity (C)



16

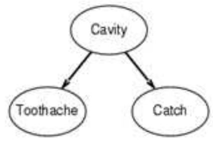
Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other

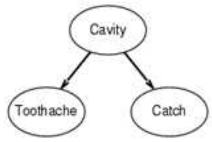


Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C



Examples of Conditional Independence

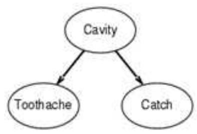


- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)

19

Examples of Conditional Independence

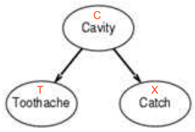


- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another

20

Examples of Conditional Independence



- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

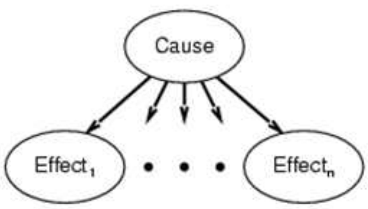
- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another
- R and S are conditionally independent given B

21

Naïve Bayes Model

P(A,B,C,D) = { P(A)P(B|A)(P(C|A,B) } / P(D|A,B,C)

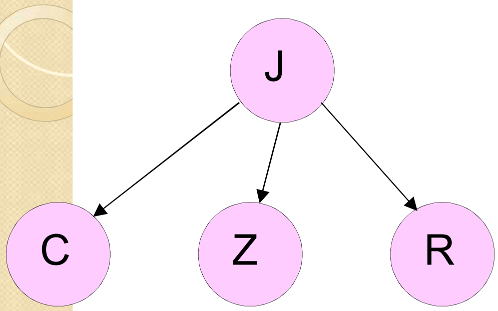
- $\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause}) \cdot \pi_i \mathbf{P}(\text{Effect}_i|\text{Cause})$



- Total number of parameters is linear in *n*

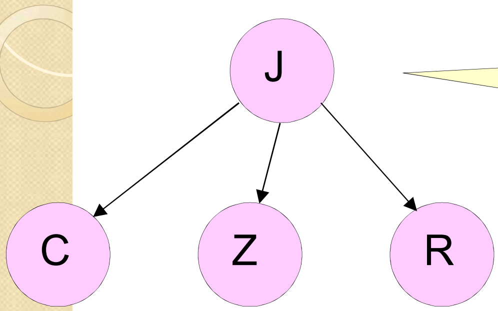
2^n개의 경우의 수
linear : 2n+1
linear하게 확률 갖고있어도
정확히 계산할 수 있는건 익스포젠셜…
101개의 확률로 2^50개의 확률을 계산할 수 있다는 것

Naïve Bayes Model



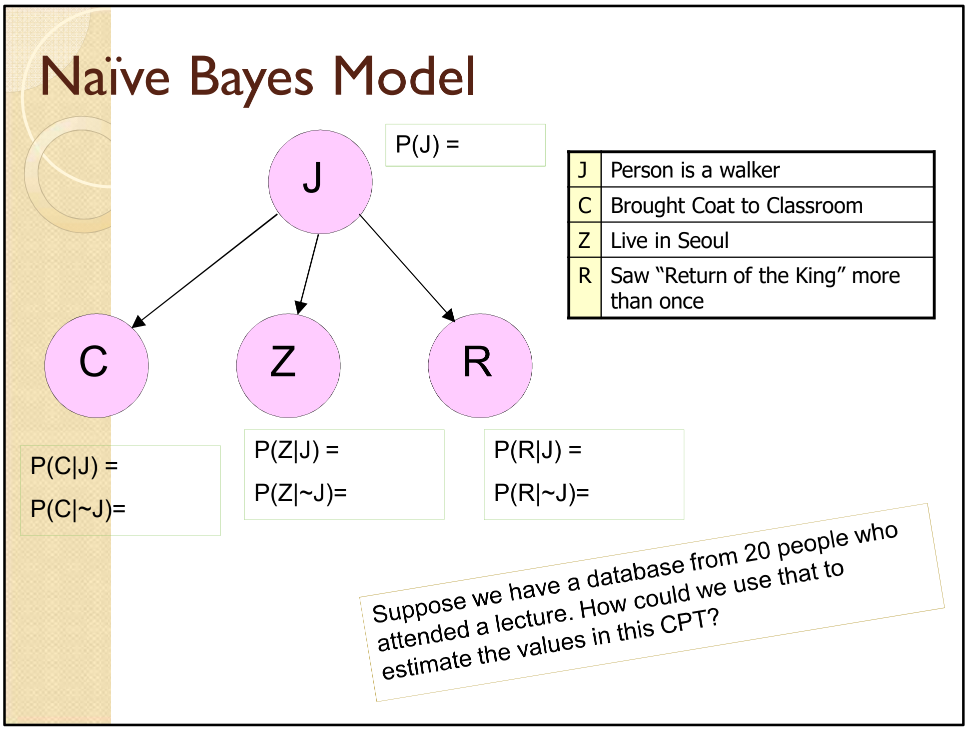
J	Person is a walker
C	Brought Coat to Classroom
Z	Live in Seoul
R	Saw "Return of the King" more than once

Naïve Bayes Model

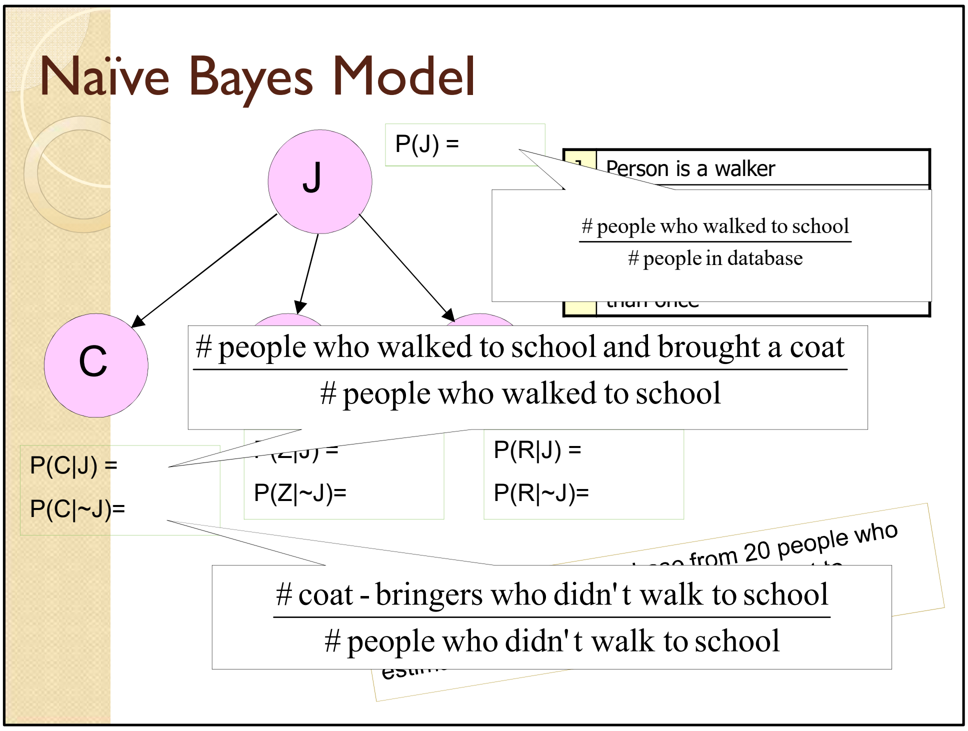


What parameters are stored in the CPTs of this Bayes Net?

J	Person is a walker
C	Brought Coat to Classroom
Z	Live in Seoul
R	Saw "Return of the King" more than once

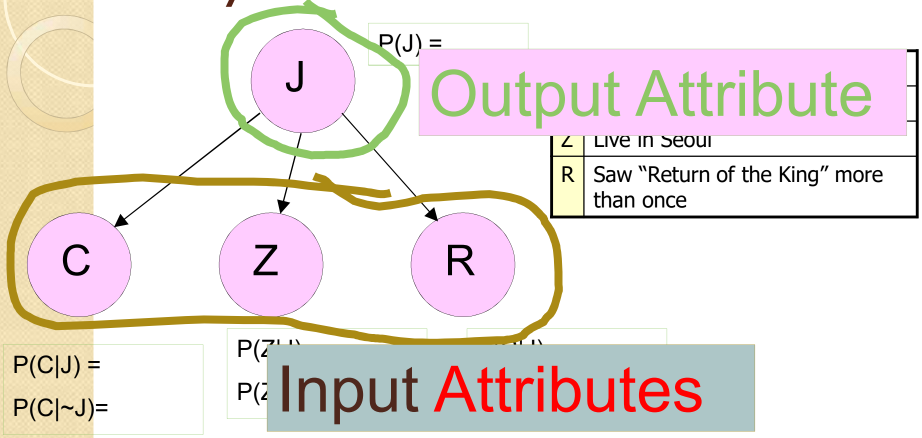


Naïve Bayes Model
J
C
Z
R
P(J) =
J Person is a walker
C Brought Coat to Classroom
Z Live in Seoul
R Saw "Return of the King" more than once
P(C|J) =
P(C|~J)=
P(Z|J) =
P(Z|~J)=
P(R|J) =
P(R|~J)=
Suppose we have a database from 20 people who attended a lecture. How could we use that to estimate the values in this CPT?



Naïve Bayes Model
J
C
P(J) =
J Person is a walker
people who walked to school
people in database
people who walked to school and brought a coat
people who walked to school
P(C|J) =
P(C|~J)=
P(Z|~J)=
P(R|J) =
P(R|~J)=
coat - bringers who didn't walk to school
people who didn't walk to school

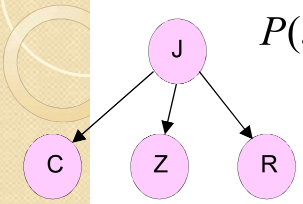
Naïve Bayes Classifier



A new person shows up at class wearing an overcoat saying "I live in Bundang and I saw all the Lord of The Rings Movies every night".

What is the probability that he is a walker?

Naïve Bayes Classifier Inference



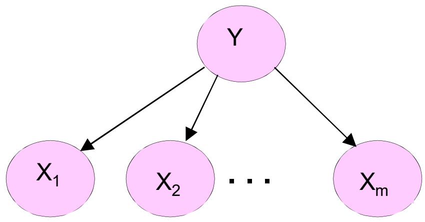
$$P(J \mid C \wedge \neg Z \wedge R) =$$

$$= \frac{P(J \wedge C \wedge \neg Z \wedge R)}{P(C \wedge \neg Z \wedge R)}$$

$$= \frac{P(J \wedge C \wedge \neg Z \wedge R)}{P(J \wedge C \wedge \neg Z \wedge R) + P(\neg J \wedge C \wedge \neg Z \wedge R)}$$

$$= \frac{P(C \mid J)P(\neg Z \mid J)P(R \mid J)P(J)}{\begin{pmatrix} P(C \mid J)P(\neg Z \mid J)P(R \mid J)P(J) \\ + \\ P(C \mid \neg J)P(\neg Z \mid \neg J)P(R \mid \neg J)P(\neg J) \end{pmatrix}}$$

Naïve Bayes General Case



1. Estimate P(Y=v) as fraction of records with Y=v
2. Estimate P(X_i=u | Y=v) as fraction of “Y=v” records that also have X=u.
3. To predict the Y value given observations of all the X_i values, compute

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y=v \mid X_1 = u_1 \cdots X_m = u_m)$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y=v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(Y=v \wedge X_1 = u_1 \cdots X_m = u_m)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y=v) P(Y=v)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(X_1 = u_1 \cdots X_m = u_m \mid Y=v) P(Y=v)$$

Because of the structure of the Bayes Net

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}}\, P(Y=v) \prod_{j=1}^{n_Y} P(X_j = u_j \mid Y=v)$$

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

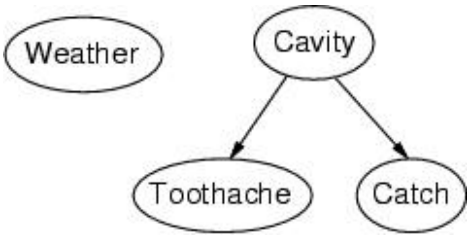
- **Syntax:**
  - **a set of nodes, one per variable**
  -
  - **a directed, acyclic graph (link ≈ "directly influences")**
  - **a conditional distribution for each node given its parents:**

$$\mathbf{P\ (X_i \mid Parents\ (X_i))}$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

A Very Simple Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

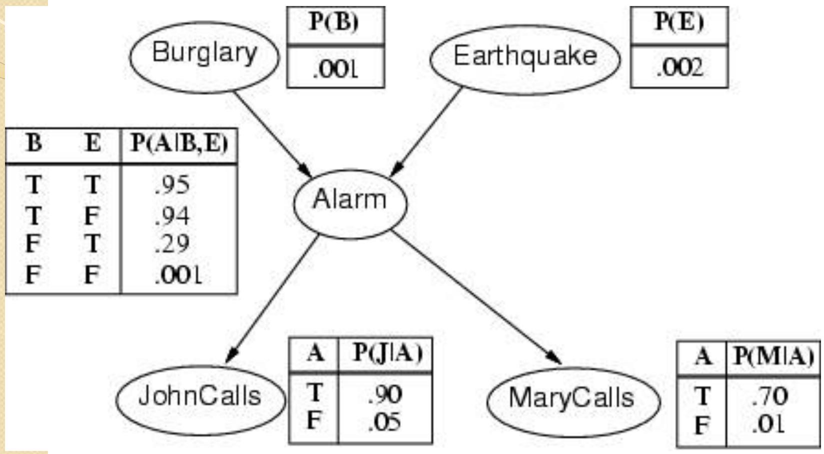
Another Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

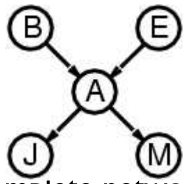
- Network topology reflects "causal" knowledge:
  - A burglar can make the alarm ringing
  - An earthquake can sometimes make the alarm ringing
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.

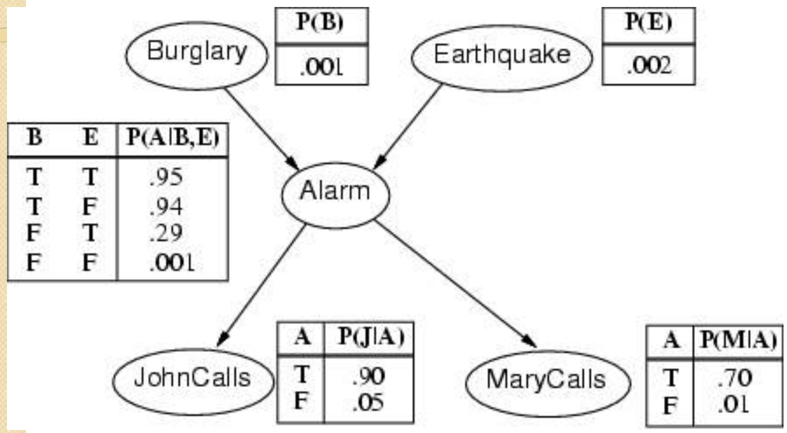


Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for X_i = *true* (the number for X_i = *false* is just $1-p$)



- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)



Query Types

Given a Bayesian network, what questions might we want to ask?

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)
- Maximum a posteriori probability:
  - What value of x maximizes P(x | e) ?

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)
- Maximum a posteriori probability:
  - What value of x maximizes P(x | e) ?
  - General question: What's the whole probability distribution over variable X given evidence e ? i.e. P(X | e)?

Using the joint distribution

To answer any query involving a conjunction of variables, sum over the variables not involved in the query.

Using the joint distribution

To answer any query involving a conjunction of variables, sum over the variables not involved in the query. 무식하게~ 정확하게~ 빠르게~ 계산

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c \wedge D = d)$$

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

둘 다 더해줘야 한다

P(cavity) = 0.04 + 0.06 = 0.1

[add elements of cavity row]

Using the joint distribution

To answer any query involving a conjunction of variables, sum over the variables not involved in the query.

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c \wedge D = d)$$

Using the joint distribution

To answer any query involving a conjunction of variables, sum over the variables not involved in the query.

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

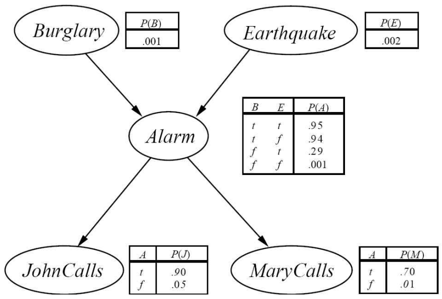
$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c \wedge D = d)$$

$$\Pr(d \mid b) = \frac{\Pr(b,d)}{\Pr(b)} = \frac{\sum_{AC} \Pr(a,b,c,d)}{\sum_{ACD} \Pr(a,b,c,d)}$$

Inference (Reasoning) in Bayesian Networks

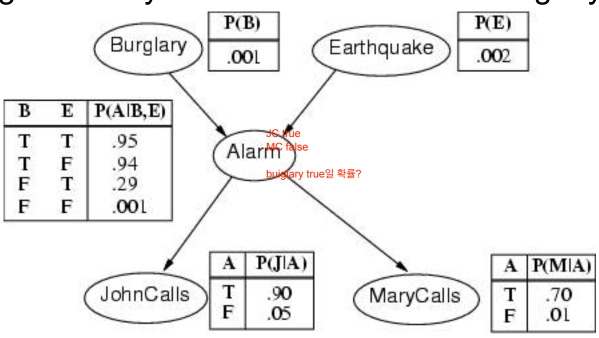
- Consider answering a query in a Bayesian Network
  - Q = set of query variables
  - e = evidence (set of instantiated variable-value pairs)
  - Inference = computation of conditional distribution P(Q | e)

- Examples
  - P(burglary | alarm)
  - P(earthquake | JCalls, MCalls)
  - P(JCalls, MCalls | burglary, earthquake)



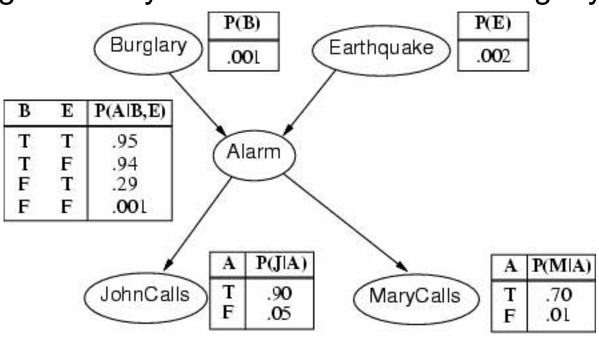
Example - Revisited

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Is there a burglary ?



Example - Revisited

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Is there a burglary ?



Pr(Burglary | JohnCalls, ¬MaryCalls) = 0.0495

Pr(¬Burglary | JohnCalls, ¬MaryCalls) = 0.9505

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easier for domain experts to construct BN because they may have prior knowledge of causal denpendencies