Artificial Intelligence

CS 165A

Oct 15, 2020

Instructor: Prof. Yu-Xiang Wang

Today

→ Probability notations

→ Counting number of parameters

→ Probabilistic modeling

→ Factorization and conditional independence

Recap: Last lecture

- Logistic loss and its gradient
- Stochastic gradient descent
- From linear logistic regression to neural networks
- Discriminative vs. Generative modelling

Plan for today

- Basics
  - Probability notations
  - Joint distributions, marginal, conditional
  - Representing these quantities as arrays / matrices
- Modeling:
  - Case study of “author classification”
- Conditional independences and factorization
- Introduction to BayesNet

- Tuesday next week:
  - BayesNet examples
  - d-separation, reasoning and inference, probabilistic modelling

Probability notation and notes

- Probabilities of *propositions / events*
  - P(A), P(the sun is shining)
- Probabilities of *random variables (r.v.)*
  - $P(X = x_1)$, $P(Y = y_1)$, $P(x_1 < X < x_2)$

Probability notation and notes

- Probabilities of *propositions / events*
  - P(A), P(the sun is shining)
- Probabilities of *random variables (r.v.)*
  - $P(X = x_1)$, $P(Y = y_1)$, $P(x_1 < X < x_2)$
- P(A) usually means P(A = True) (A is a proposition, not a variable)
  - This is a probability **value**
  - Technically, P(A) is a probability *function*

P(¬A)

- $P(X = x_1)$
  - This is a probability **value** (P(X) is a probability *function*)
- P(X)
  - This is a **probability distribution** function, a.k.a probability mass function (**p.m.f.**) for discrete r.v. or a probability density function (**p.d.f.**) for continuous r.v.

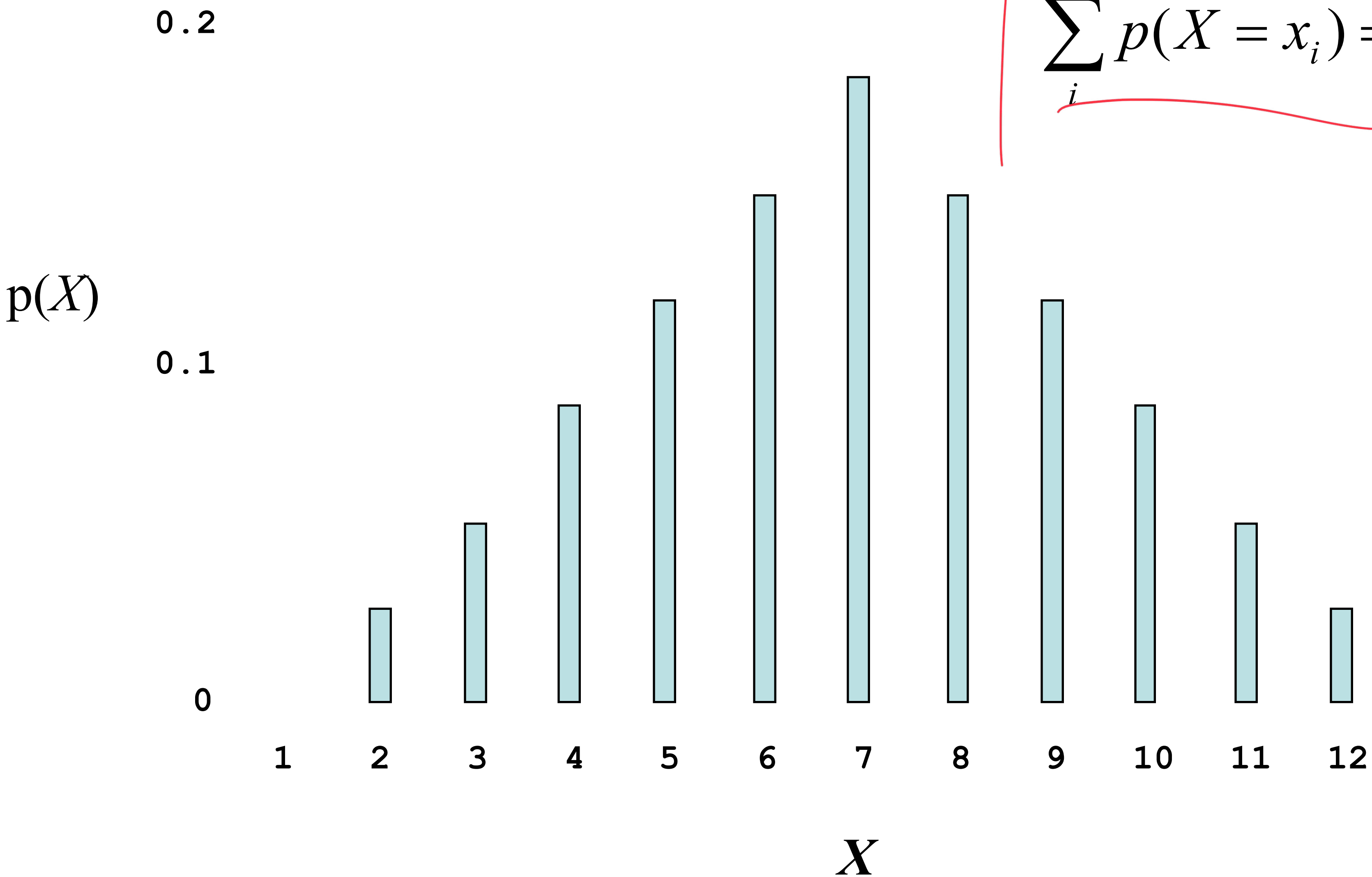
Probability notation and notes

- Probabilities of *propositions / events*
  - P(A), P(the sun is shining)
- Probabilities of *random variables (r.v.)*
  - $P(X = x_1)$, $P(Y = y_1)$, $P(x_1 < X < x_2)$
- P(A) usually means P(A = True) (A is a proposition, not a variable)
  - This is a probability **value**
  - Technically, P(A) is a probability *function*
- $P(X = x_1)$
  - This is a probability **value** (P(X) is a probability *function*)
- P(X)
  - This is a **probability distribution** function, a.k.a probability mass function (**p.m.f.**) for discrete r.v. or a probability density function (**p.d.f.**) for continuous r.v.
- Technically, if X is an r.v., we should not write $P(X) = 0.5$
  - But rather $P(X = x_1) = 0.5$

$P(X) = [P(X=0), P(X=1)]$

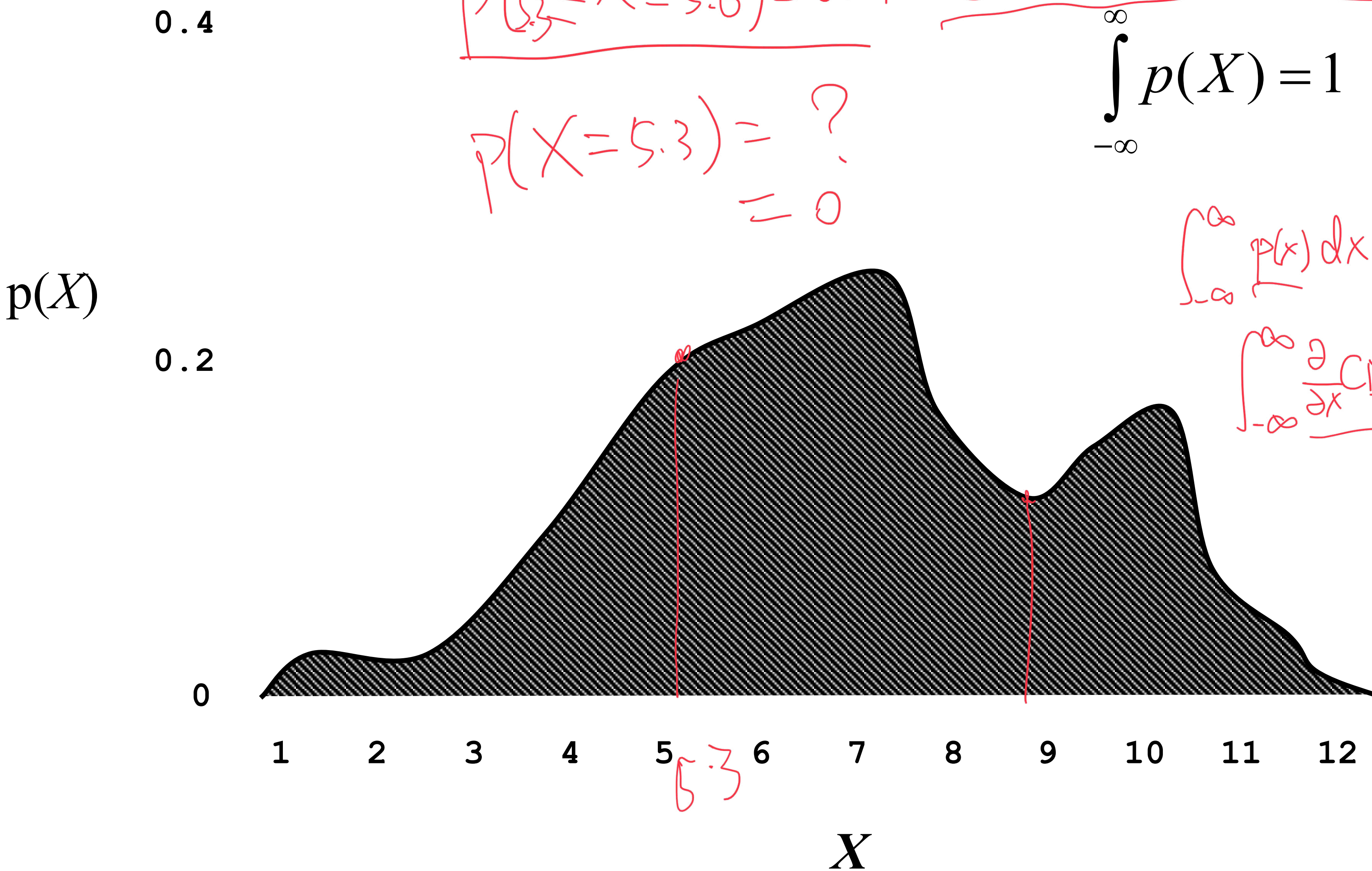
Discrete probability distribution

$$\sum_i p(X = x_i) = 1$$



Continuous probability distribution

$$\int_{-\infty}^{\infty} p(X) = 1$$



Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1{=}x_1 \wedge X_2{=}x_2 \wedge \ \ldots \wedge X_i{=}x_i \wedge \ \ldots) = p$

Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1=x_1 \wedge X_2=x_2 \wedge \ \ldots \wedge X_i=x_i \wedge \ \ldots) = p$
- Rather than stating knowledge as
  - Raining $\Rightarrow$ WetGrass

Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1{=}x_1 \wedge X_2{=}x_2 \wedge \ldots \wedge X_i{=}x_i \wedge \ldots) = p$
- Rather than stating knowledge as
  - Raining ⇒ WetGrass
- We can state it as
  - P(Raining, WetGrass) = 0.15
  - P(Raining, ¬WetGrass) = 0.01
  - P(¬Raining, WetGrass) = 0.04
  - P(¬Raining, ¬WetGrass) = 0.8

Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1=x_1 \wedge X_2=x_2 \wedge \ldots \wedge X_i=x_i \wedge \ldots) = p$
- Rather than stating knowledge as
  - Raining ⇒ WetGrass
- We can state it as
  - P(Raining, WetGrass) = 0.15
  - P(Raining, ¬WetGrass) = 0.01
  - P(¬Raining, WetGrass) = 0.04
  - P(¬Raining, ¬WetGrass) = 0.8

	¬WetGrass	WetGrass
¬Raining	0.8	0.04
Raining	0.01	0.15

Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \ldots, X_i, \ldots)$
- A particular instance of the world has the probability
  - $P(X_1{=}x_1 \wedge X_2{=}x_2 \wedge \ldots \wedge X_i{=}x_i \wedge \ldots) = p$
- Rather than stating knowledge as
  - Raining ⇒ WetGrass
- We can state it as
  - P(Raining, WetGrass) = 0.15
  - P(Raining, ¬WetGrass) = 0.01
  - P(¬Raining, WetGrass) = 0.04
  - P(¬Raining, ¬WetGrass) = 0.8

	¬WetGrass	WetGrass
¬Raining	0.8	0.04
Raining	0.01	0.15

Marginal and Conditional Probability

- Marginal Probability
  - Marginal probability (distribution) of X: $P(X) = \sum_Y P(X, Y)$
  - **Bayesian interpretation:** Probabilities associated with one proposition or variable, **prior** to any evidence
  - E.g., P(WetGrass), P(¬Raining)
- Conditional Probability
  - P(A | B) – "The probability of A given that we know B"
  - **Bayesian interpretation:** After (**posterior** to) procuring evidence
  - E.g., P(WetGrass | Raining)

Marginal and Conditional Probability

- Marginal Probability
  - Marginal probability (distribution) of X: $P(X) = \sum_Y P(X, Y)$
  - **Bayesian interpretation:** Probabilities associated with one proposition or variable, **prior** to any evidence
  - E.g., P(WetGrass), P(¬Raining)
- Conditional Probability
  - P(A | B) – "The probability of A given that we know B"
  - **Bayesian interpretation:** After (**posterior** to) procuring evidence
  - E.g., P(WetGrass | Raining)

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} \quad \text{or} \quad P(X \mid Y)\, P(Y) = P(X, Y)$$

Assumes P(Y) nonzero

The chain rule: factorizing a joint distribution into marginal and conditionals

$$P(X,Y) = P(X \mid Y)\, P(Y)$$

By the Chain Rule

$$P(X,Y,Z) = P(X \mid Y,Z)\, P(Y,Z)$$

= P(Y|Z) P(Z)

= P(Z|Y) P(Y)

$$= P(X \mid Y,Z)\, P(Y \mid Z)\, P(Z)$$

or, equivalently

$$= P(X) P(Y \mid X) P(Z \mid X,Y)$$

Notes:

- Precedence: '|' is lowest
- E.g., P(X | Y, Z) means which?

  P((X | Y), Z)

  P(X | (Y, Z))

Chain Rule implies Bayes' Rule

Thomas Bayes: 1701 - 1761

- Since $P(X,Y) = P(X \mid Y)\, P(Y)$

  and $P(X,Y) = P(Y \mid X)\, P(X)$

- Then $P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$

Posterior of X|Y; Conditional of Evidence; Prior

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

$$P(Y) = \sum_{X} P(Y \mid X) \cdot P(X)$$

Chain Rule implies Bayes' Rule

Thomas Bayes: 1701 - 1761

- Since $P(X,Y) = P(X \mid Y)\,P(Y)$

  and $P(X,Y) = P(Y \mid X)\,P(X)$

- Then $P(X \mid Y)\,P(Y) = P(Y \mid X)\,P(X)$

$$P(X \mid Y) = \frac{P(Y \mid X)\,P(X)}{P(Y)}$$ Bayes' Rule

Chain Rule implies Bayes' Rule

Thomas Bayes: 1701 - 1761

- Since $P(X,Y) = P(X \mid Y)\, P(Y)$

  and $P(X,Y) = P(Y \mid X)\, P(X)$

- Then $P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$

$$P(X \mid Y) = \frac{P(Y \mid X)\ P(X)}{P(Y)}$$ Bayes' Rule

Funny fact: Thomas Bayes is arguably a frequentist.

Stephen Fienberg. "When did Bayesian inference become 'Bayesian'?." *Bayesian analysis* 1.1 (2006): 1-40.
https://projecteuclid.org/euclid.ba/1340371071

Representing Probability Distributions using linear algebraic data structures (in python)

	Continuous vars	Discrete vars
P(X)	Function (of one variable)	m vector
P(X=x)	Scalar*	Scalar
P(X,Y)	Function of two variables	m×n matrix
P(X|Y)	Function of two variables	m×n matrix
P(X|Y=y)	Function of one variable	m vector
P(X=x|Y)	Function of one variable	n vector
P(X=x|Y=y)	Scalar*	Scalar

* - actually zero. Should be $P(x_1 < X < x_2)$

Example: Joint probability distribution

From $P(X,Y)$, we can always calculate:

$P(X)$	$P(X=x_1)$
$P(Y)$	$P(Y=y_2)$
$P(X\|Y)$	$P(X\|Y=y_1)$
$P(Y\|X)$	$P(Y\|X=x_1)$
	$P(X=x_1\|Y)$
	etc.

Example: Joint probability distribution

From P(X,Y), we can always calculate:

$P(X)$ $P(X=x_1)$

$P(Y)$ $P(Y=y_2)$

$P(X|Y)$ $P(X|Y=y_1)$

$P(Y|X)$ $P(Y|X=x_1)$

$P(X=x_1|Y)$

etc.

X

Y		x_1	x_2	x_3
	y_1	0.2	0.1	0.1
	y_2	0.1	0.2	0.3

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

$$\frac{0.2}{0.2+0.1+0.1} =$$

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

$P(X=x_1, Y=y_2) = ?$

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

$P(X|Y=y_1) = ?$

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

$P(X|Y=y_1) = ?$

$P(X=x_1|Y) = ?$

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

Quick checkpoint

- Probability notations
  - P(A) is a number when A is an event / predicate.
  - P(X) is a vector/function when X is a random variable.

- Joint probability distribution
  - Enumerating all combinations of events.
  - All values the random variables can take.
  - Assign a non-negative value to each.

- Marginals, conditionals
  - How they are related: Chain rule, Bayes rule

You should know HOW TO do the following:

- For discrete probability distributions for multiple random variables
  - Know **the number of possible values** these RVs can take
  - Know the **shape of the numpy arrays** that you need to represent Joint-distribution,   conditional distribution
  - Know the **number of independent parameters** you need to specify these distributions. (we often need to -1 here or there. Why is that?)
- More generally: Know the distinctions between
  - p.m.f -- probability mass function (for discrete distribution)
  - p.d.f. -- probability density function (for continuous distribution)
  - CDF -- cumulative distribution function (for both)

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

- With the joint distribution, we can do anything we want

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

- With the joint distribution, we can do anything we want
  - Design classifiers

$$\hat{y} = \text{arg}max_y P(Y = y \,|X)$$

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

- With the joint distribution, we can do anything we want
  - Design classifiers

$$\hat{y} = \text{arg}max_y P(Y = y \,|X)$$

  - We can make Bayesian inference (probabilistic reasoning)
    - Sherlock Holmes: P(Murderer | Observed Evidence)

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

- With the joint distribution, we can do anything we want
  - Design classifiers

$$\hat{y} = \text{arg}max_y P(Y = y \,|X)$$

  - We can make Bayesian inference (probabilistic reasoning)
    - Sherlock Holmes: P(Murderer | Observed Evidence)
    - Doctor: P(Disease | Symptoms) , P(Effect | Treatment)

How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge

- With the joint distribution, we can do anything we want
  - Design classifiers

$$\hat{y} = \text{arg}max_y P(Y = y \mid X)$$

  - We can make Bayesian inference (probabilistic reasoning)
    - Sherlock Holmes: P(Murderer | Observed Evidence)
    - Doctor: P(Disease | Symptoms) , P(Effect | Treatment)
    - Parenting:
      - P(Dirty Diaper, Hungry, Lonely | 5 a.m., Baby crying)
      - P(Baby crying at 5 a.m. | feeding at 2 a.m.)
      - P(Baby crying at 5 a.m. | feeding at 1 a.m.)

(3 min discussion) Modeling the world with probability distribution

- Example: Author attribution task as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Kafka, Jane Austen, Tolkien, George RR. Martin, … , Xueqin Cao, Douglas Adams}
  - A vocabulary of size 3000

(3 min discussion) Modeling the world with probability distribution

- Example: Author attribution task as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Kafka, Jane Austen, Tolkien, George RR. Martin, … , Xueqin Cao, Douglas Adams}
  - A vocabulary of size 3000

$P(W_1, \dots W_N, Author)$

3000×15 ?

$W_1 \in \{w_1, \dots w_{3000}\}$

- Questions:
  - What is the dimension(s) of the joint distribution?
  - How many free parameters are needed to represent this distribution?

$3000^N \cdot 15$

$3000^N \cdot 15 - 1$

Statistical Independences

(Marginal / absolute) Independence

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

- X and Y are independent iff
  - $P(X,Y) = P(X)\,P(Y)$ [by definition]
  - $P(X \mid Y) = P(X)$ Since $P(X \mid Y) = P(X, Y)/P(Y) = P(X)\,P(Y)/P(Y)$

Statistical Independences

(Marginal / absolute) Independence

- X and Y are independent iff
  - $P(X,Y) = P(X)\,P(Y)$ [by definition]
  - $P(X \mid Y) = P(X)$ Since $P(X \mid Y) = P(X, Y)/P(Y) = P(X)\,P(Y)/P(Y)$

Conditional Independence

- If X and Y are (conditionally) independent given Z, then
  - $P(X \mid Y, Z) = P(X \mid Z)$

Statistical Independences

(Marginal / absolute) Independence

- X and Y are independent iff
  - $P(X,Y) = P(X)\,P(Y)$ [by definition]
  - $P(X \mid Y) = P(X)$ Since $P(X \mid Y) = P(X, Y)/P(Y) = P(X)\,P(Y)/P(Y)$

Conditional Independence

- If X and Y are (conditionally) independent given Z, then
  - $P(X \mid Y, Z) = P(X \mid Z)$
  - Example:
    - P(WetGrass | Season, Rain) = P(WetGrass | Rain)

Example of Conditional Independence

- In practice, conditional independence is more common than marginal independence.
  - P(Final exam grade | Weather) ≠ P(Final exam grade)
    - i.e., they are not independent
  - P(Final exam grade | Weather, Effort) = P(Final exam grade | Effort)
    - But they are conditionally independent given Effort

(Example continued) Modeling the world with probability distribution

- Example: Author attribution as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N*, *Author*
  - 15 authors in total: {Dickens, Shakespeare, Tolkien, George RR. Martin, … ,Douglas Adams}
  - A vocabulary of size 3000

(Example continued) Modeling the world with probability distribution

- Example: Author attribution as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N*, *Author*
  - 15 authors in total: {Dickens, Shakespeare, Tolkien, George RR. Martin, … ,Douglas Adams}
  - A vocabulary of size 3000

- In addition, assume that: *Word 1, …, Word N* are **mutually independent** given *Author*
  - $P(Word\ 1, \ldots, Word\ N \mid Author) = P(Word\ 1 \mid Author) \times \ldots \times P(Word\ N \mid Author)$

(Example continued) Modeling the world with probability distribution

- Example: Author attribution as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Tolkien, George RR. Martin, … ,Douglas Adams}
  - A vocabulary of size 3000
- In addition, assume that: *Word 1, …, Word N* are **mutually independent** given *Author*
  - $P(\text{Word } 1, \ldots, \text{Word } N \mid \text{Author}) = P(\text{Word } 1 \mid \text{Author}) \times \ldots \times P(\text{Word } N \mid \text{Author})$
- Question:
  - What are the dimensions of each factor?
  - How many “free parameters” are needed in total?

(Example continued) Modeling the world with probability distribution

- Example: Author attribution as in HW1
  - Variables: *Word 1, Word 2, Word 3, …, Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Tolkien, George RR. Martin, … ,Douglas Adams}
  - A vocabulary of size 3000

- In addition, assume that: *Word 1, …, Word N* are **mutually independent** given *Author*
  - $P(Word\ 1, \ldots, Word\ N \mid Author) = P(Word\ 1 \mid Author) \times \ldots \times P(Word\ N \mid Author)$

- Question:
  - What are the dimensions of each factor?
  - How many "free parameters" are needed in total?

Quiz time: Representing a joint Probability

- Joint probability: $P(X_1, X_2, \ldots, X_N)$
  - Defines the probability for any possible state of the world

Quiz time: Representing a joint Probability

- Joint probability: $P(X_1, X_2, \ldots, X_N)$
  - Defines the probability for any possible state of the world
  - Let the variables be binary. How many numbers ("free parameters") does it take to define the joint distribution?

$2^N - 1$

Quiz time: Representing a joint Probability

- Joint probability: $P(X_1, X_2, \ldots, X_N)$
  - Defines the probability for any possible state of the world
  - Let the variables be binary. How many numbers ("free parameters") does it take to define the joint distribution?

- If the variables are independent, then

  $P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2) \ldots P(X_N)$

Quiz time: Representing a joint Probability

- Joint probability: $P(X_1, X_2, \ldots, X_N)$
  - Defines the probability for any possible state of the world
  - Let the variables be binary. How many numbers ("free parameters") does it take to define the joint distribution?

- If the variables are independent, then

  $$P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2) \ldots P(X_N)$$

  - How many numbers does it take to define the joint distribution?

Tradeoffs in our model choices

Fully Independent

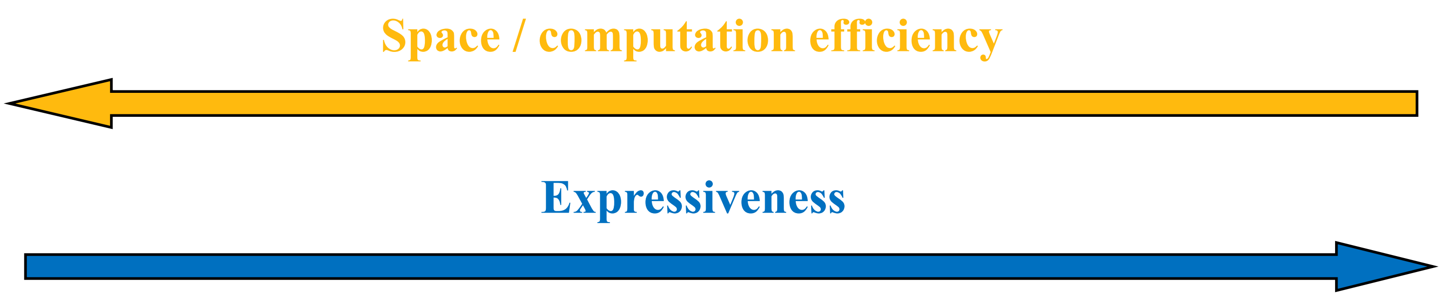
$P(X_1, X_2, \ldots, X_N)$
$= P(X_1)\, P(X_2) \ldots P(X_N)$

$$O(N)$$

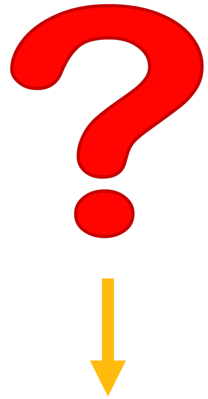
Fully general

$P(X_1, X_2, \ldots, X_N)$

$$O(e^N)$$



Tradeoffs in our model choices



Fully Independent

$P(X_1, X_2, \ldots, X_N)$
$= P(X_1)\, P(X_2) \ldots P(X_N)$

$$O(N)$$

Fully general

$P(X_1, X_2, \ldots, X_N)$

$$O(e^N)$$

Space / computation efficiency

Expressiveness

Tradeoffs in our model choices

?

Fully Independent

$P(X_1, X_2, \ldots, X_N)$

$= P(X_1)\, P(X_2) \ldots P(X_N)$

$O(N)$

Fully general

$P(X_1, X_2, \ldots, X_N)$

$O(e^N)$

Space / computation efficiency ←

Expressiveness →

Idea:

1. Independent groups of variables?
2. **Conditional independences?**

Benefit of conditional independence

$(X \perp Y \mid Z) \Leftrightarrow P(X \mid Y, Z) = P(X \mid Z)$

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)
    - 1 + 2 + 4 + 8 = 15 numbers to specify

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)
    - 1 + 2 + 4 + 8 = 15 numbers to specify
  - But if Y and W are independent given X, and Z is independent of W and X given Y, then

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)
    - 1 + 2 + 4 + 8 = 15 numbers to specify
  - But if Y and W are independent given X, and Z is independent of W and X given Y, then

Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)
    - 1 + 2 + 4 + 8 = 15 numbers to specify
  - But if Y and W are independent given X, and Z is independent of W and X given Y, then
    - P(W,X,Y,Z) = P(W) P(X|W) P(Y|X) P(Z|Y)
      - 1 + 2 + 2 + 2 = 7 numbers

Benefit of conditional independence

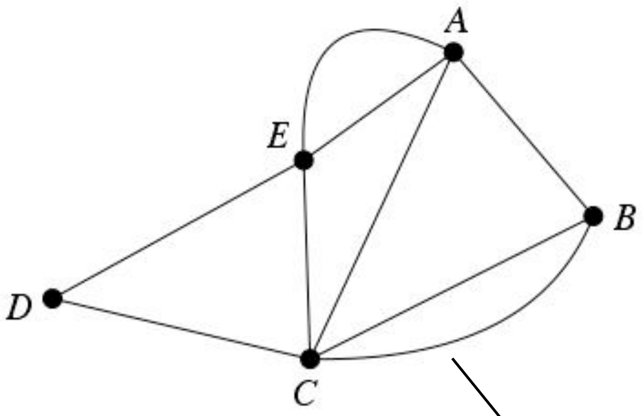
- If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)
    - 1 + 2 + 4 + 8 = 15 numbers to specify
  - But if Y and W are independent given X, and Z is independent of W and X given Y, then
    - P(W,X,Y,Z) = P(W) P(X|W) P(Y|X) P(Z|Y)
      - 1 + 2 + 2 + 2 = 7 numbers
- This is often the case in real problems.

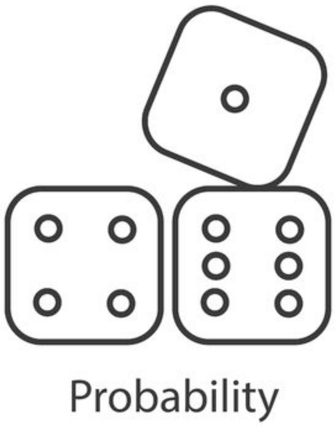
When given a problem with many variables.

[CS165A Lecture attendance,
HW1,HW2, HW3, HW4,
Readings, Piazza, Final Grade,
Weather, Election Result, Job Offer]

How do we know **which conditional independence**(s) to include in the joint distribution?

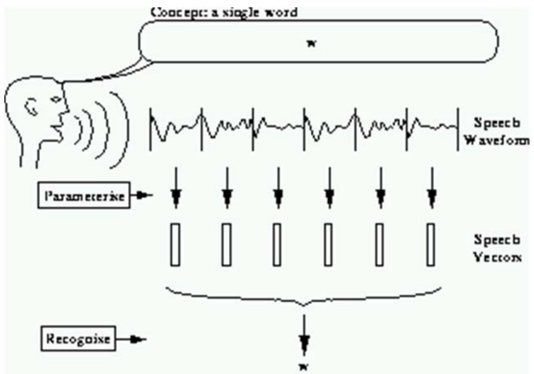
Graphical models come out of the marriage of graph theory and probability theory



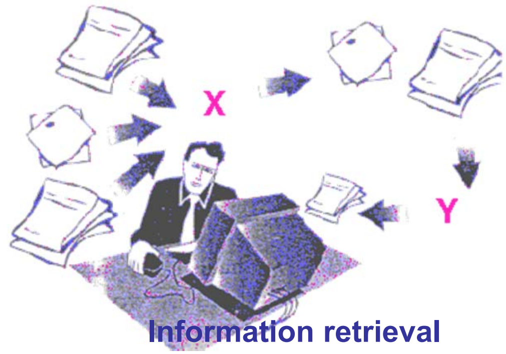


Directed Graph => Bayesian Networks / Belief Networks
Undirected Graph => Markov Random Fields

Used as a modeling tool. Many applications!

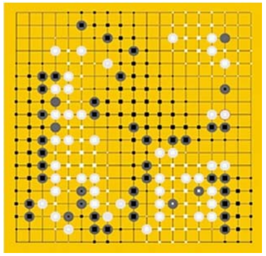


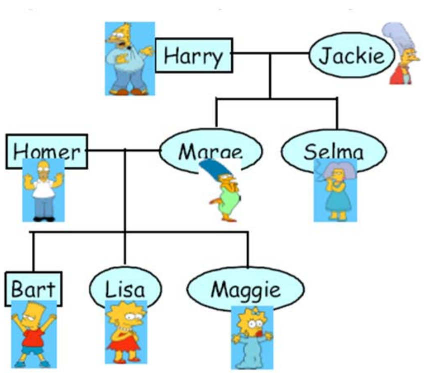
Speech recognition



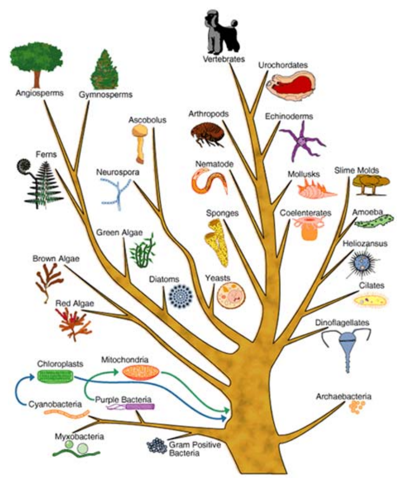
Information retrieval


Computer vision


Games

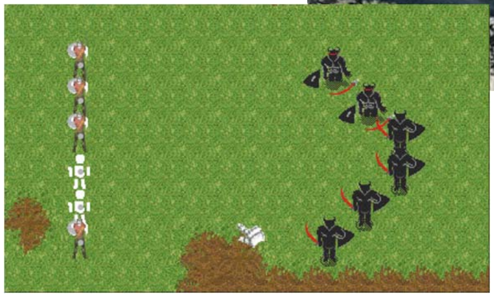


Pedigree



Evolution


Robotic control


Planning

(Slides from Prof. Eric Xing)

Two ways to think about Graphical Models

- A particular factorization of a joint distribution
  - $P(X,Y,Z) = P(X)\ P(Y|X)\ P(Z|Y)$
- A collection of conditional independences
  - $\{ X \perp Z \mid Y, \dots \}$

Two ways to think about Graphical Models

- A particular factorization of a joint distribution
  - P(X,Y,Z) = P(X) P(Y|X) P(Z|Y)
- A collection of conditional independences
  - { X ⊥ Z | Y, … }

Represented using a graph!

Belief Networks

a.k.a. Probabilistic networks, Belief nets, Bayes nets, etc.

- Belief network
  - A data structure (depicted as a graph) that represents the dependence among variables and allows us to concisely specify the joint probability distribution
  - The graph itself is known as an "influence diagram"

Belief Networks

a.k.a. Probabilistic networks, Belief nets, Bayes nets, etc.

- Belief network
  - A data structure (depicted as a graph) that represents the dependence among variables and allows us to concisely specify the joint probability distribution
  - The graph itself is known as an "influence diagram"

- A belief network is a **directed acyclic graph** where:
  - The nodes represent the set of random variables (one node per random variable)
  - Arcs between nodes represent *influence*, or *causality*
    - A link from node X to node Y means that X "directly influences" Y
  - Each node has a ***conditional probability table*** (CPT) that defines **P(node | parents)**

Example

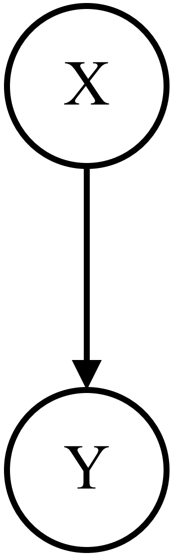
- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

  Or, Y is a *symptom* of X
- Draw two nodes and link them

Example

- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

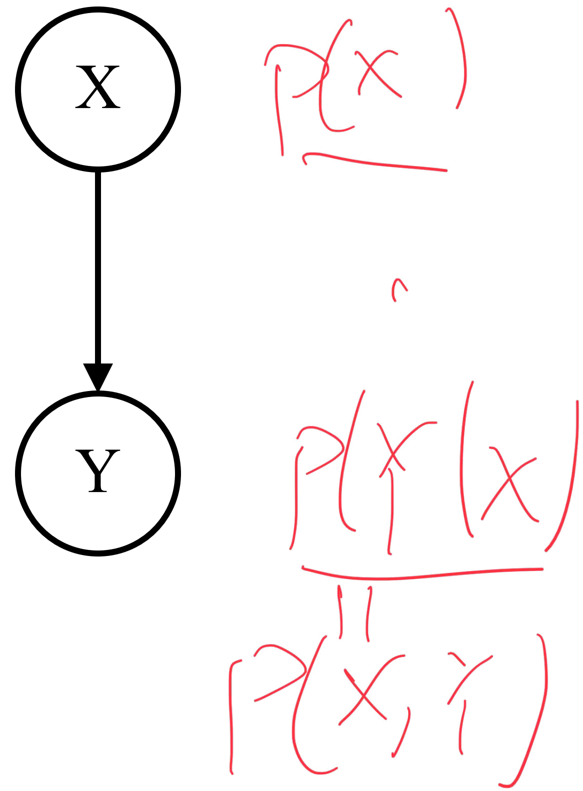
  Or, Y is a *symptom* of X
- Draw two nodes and link them



Example

- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

  Or, Y is a *symptom* of X
- Draw two nodes and link them



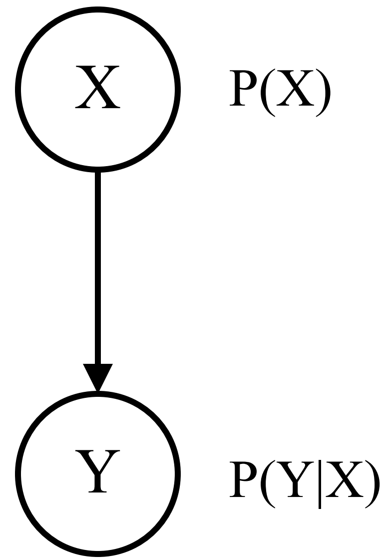
- Define the CPT for each node
  - P(X) and P(Y | X)

Example

- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

  Or, Y is a *symptom* of X
- Draw two nodes and link them

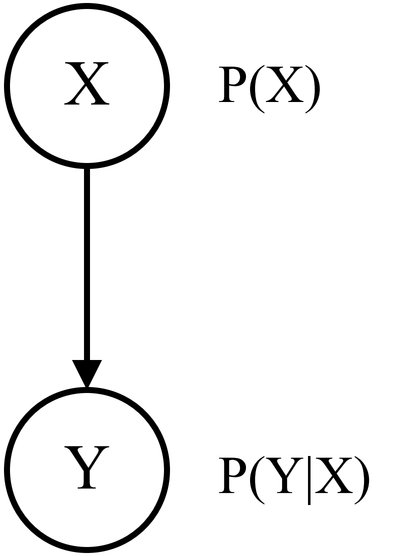
- Define the CPT for each node
  - P(X) and P(Y | X)



Example

- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

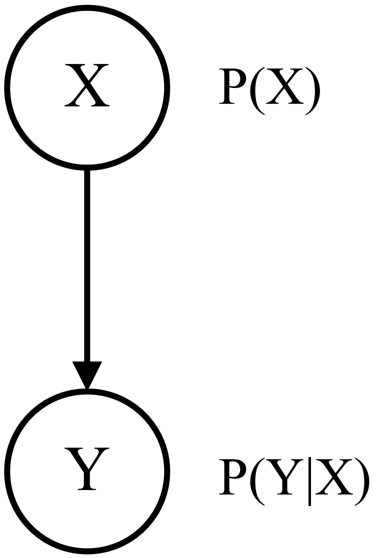
  Or, Y is a *symptom* of X
- Draw two nodes and link them



- Define the CPT for each node
  - P(X) and P(Y | X)
- Typical use: we observe Y and we want to query P(X | Y)
  - Y is an *evidence variable*
  - X is a *query variable*

We can write everything we want as a function of the CPTs. Try it!

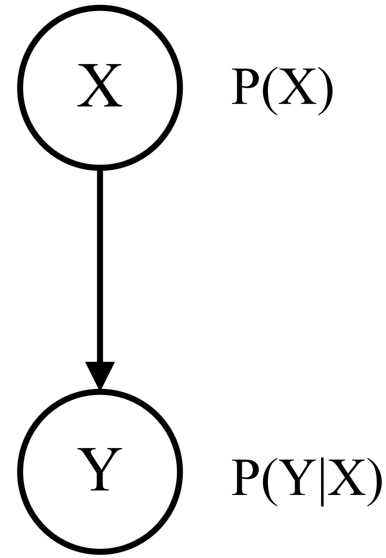
- What is P(X | Y)?
  - Given that we know the CPTs of each node in the graph



We can write everything we want as a function of the CPTs. Try it!

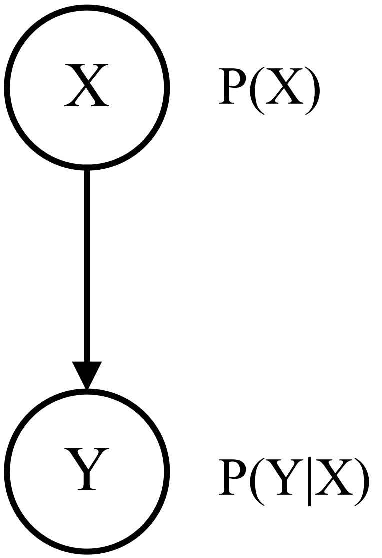
- What is P(X | Y)?
  - Given that we know the CPTs of each node in the graph

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



We can write everything we want as a function of the CPTs. Try it!

- What is P(X | Y)?
  - Given that we know the CPTs of each node in the graph

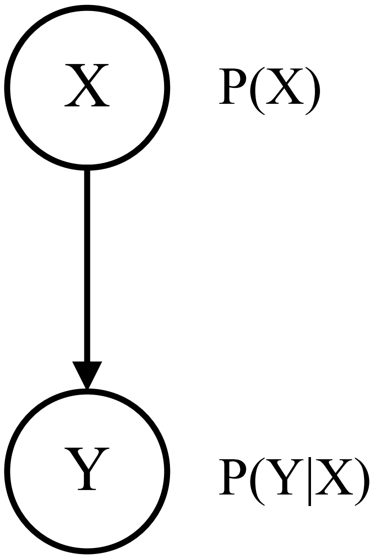


$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

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

We can write everything we want as a function of the CPTs. Try it!

- What is P(X | Y)?
  - Given that we know the CPTs of each node in the graph



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

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

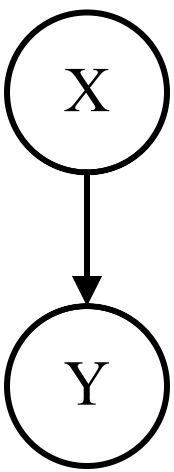
$$= \frac{P(Y \mid X)P(X)}{\sum_X P(Y \mid X)P(X)}$$

Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i|Parents(var_i))$

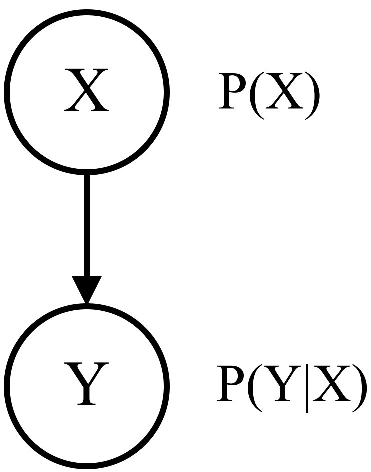
Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i|Parents(var_i))$



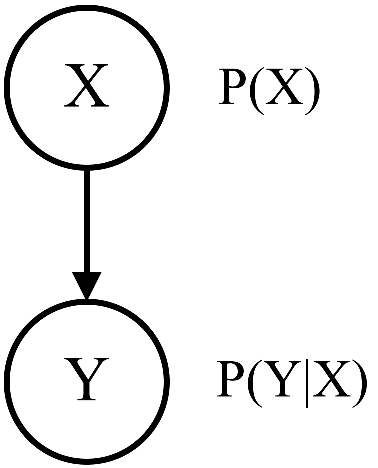
Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i|Parents(var_i))$



Belief nets represent the joint probability

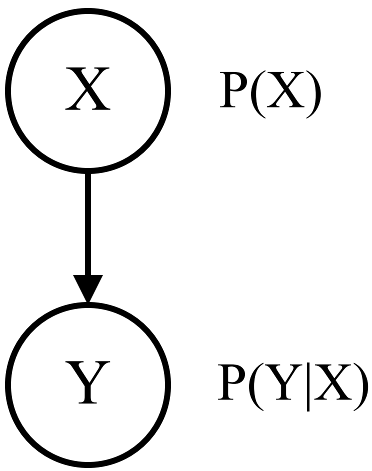
- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i | Parents(var_i))$

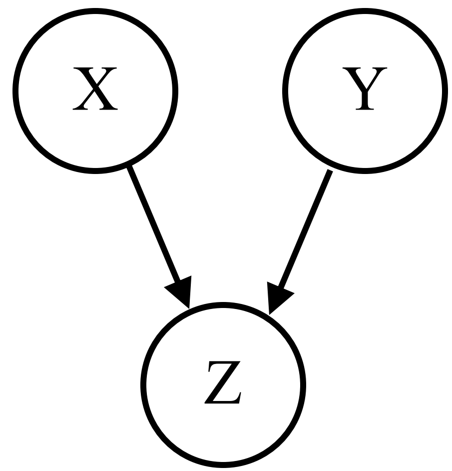


$P(X,Y) = P(X)\, P(Y|X)$

Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i | Parents(var_i))$

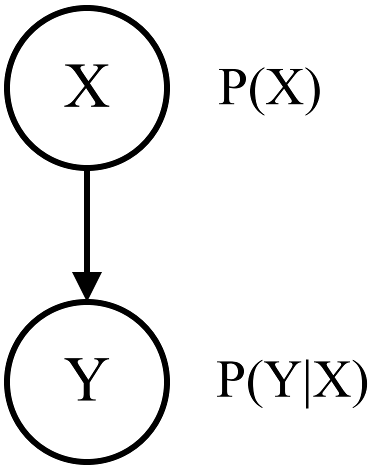


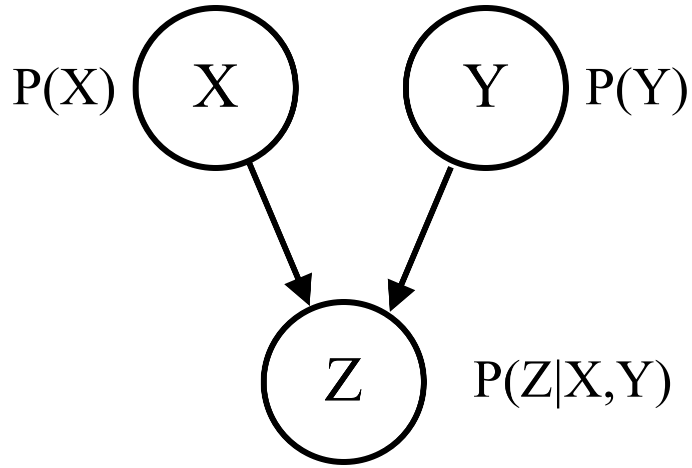


P(X,Y) = P(X) P(Y|X)

Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i | Parents(var_i))$

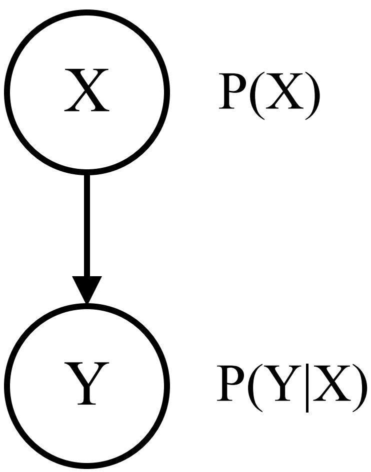


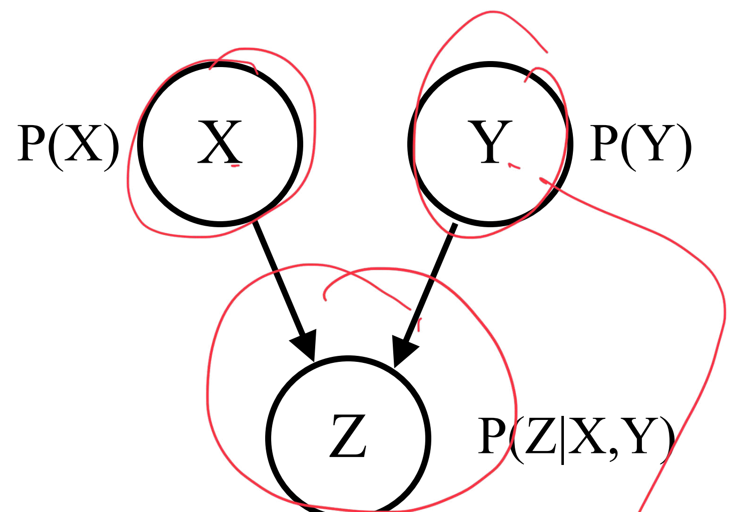


P(X,Y) = P(X) P(Y|X)

Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i | Parents(var_i))$





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

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

Three steps in modelling with Belief Networks

1. Choose variables in the environments, represent them as nodes.

2. Connect the variables by inspecting the "direct influence": cause-effect

3. Fill in the probabilities in the CPTs.

Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm

Example: Modelling with Belief Net

I’m at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn’t call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links

Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links
- This defines the joint probability
  - P(JohnCalls, MaryCalls, Earthquake, Burglar, Alarm)

Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links
- This defines the joint probability
  - P(JohnCalls, MaryCalls, Earthquake, Burglar, Alarm)
- What do we want to know?

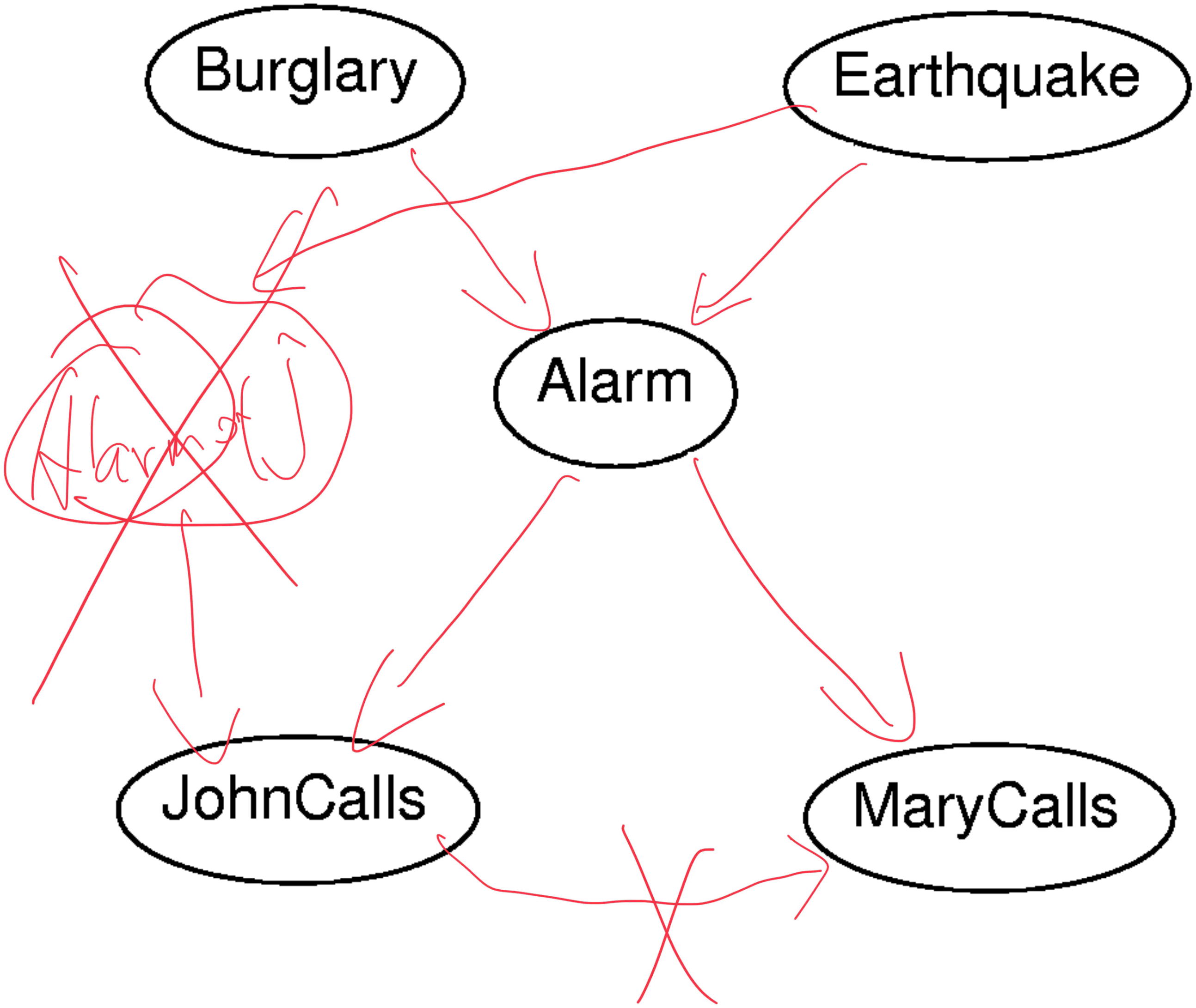
Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

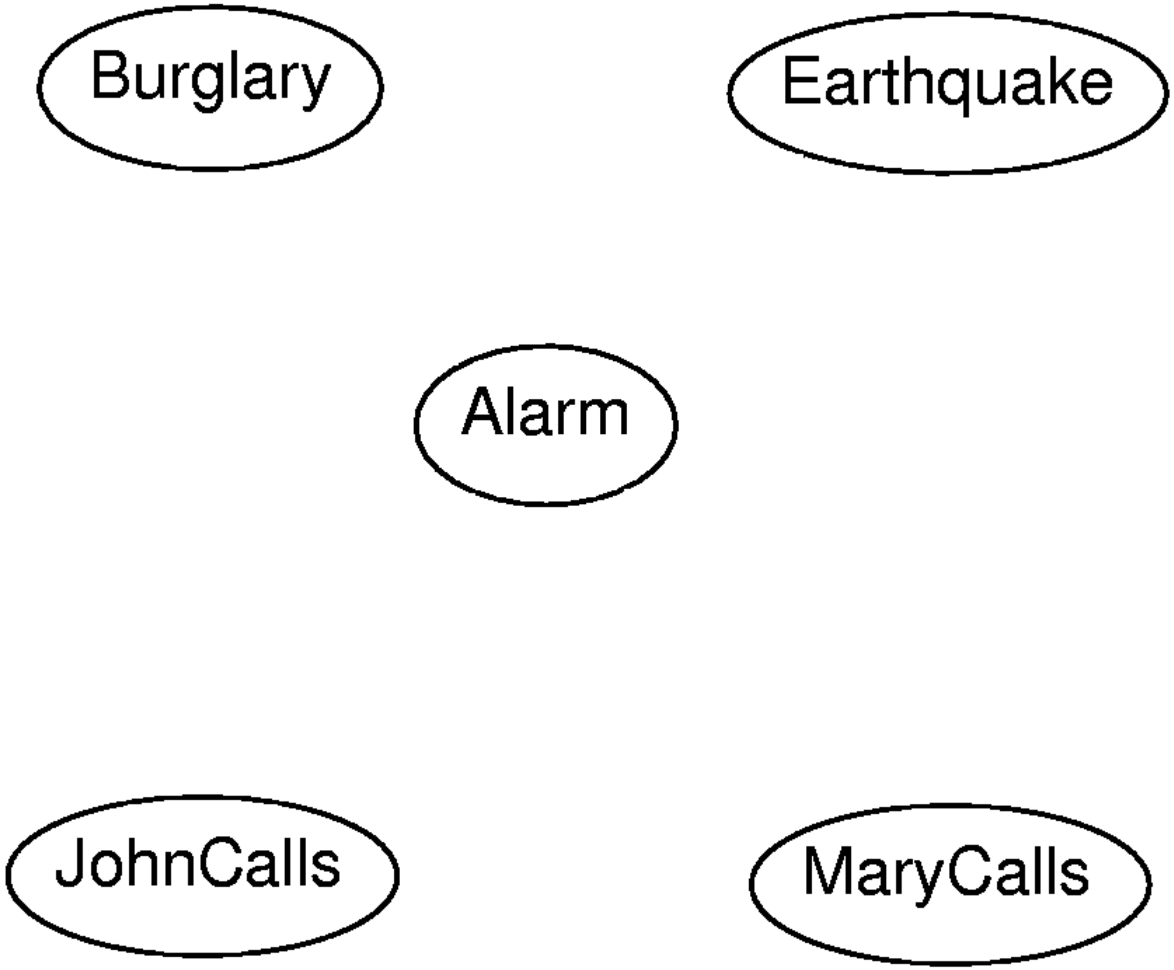
- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links
- This defines the joint probability
  - P(JohnCalls, MaryCalls, Earthquake, Burglar, Alarm)
- What do we want to know?

$P(B \mid J, \neg M)$

How should we connect the nodes? (3 min discussion)

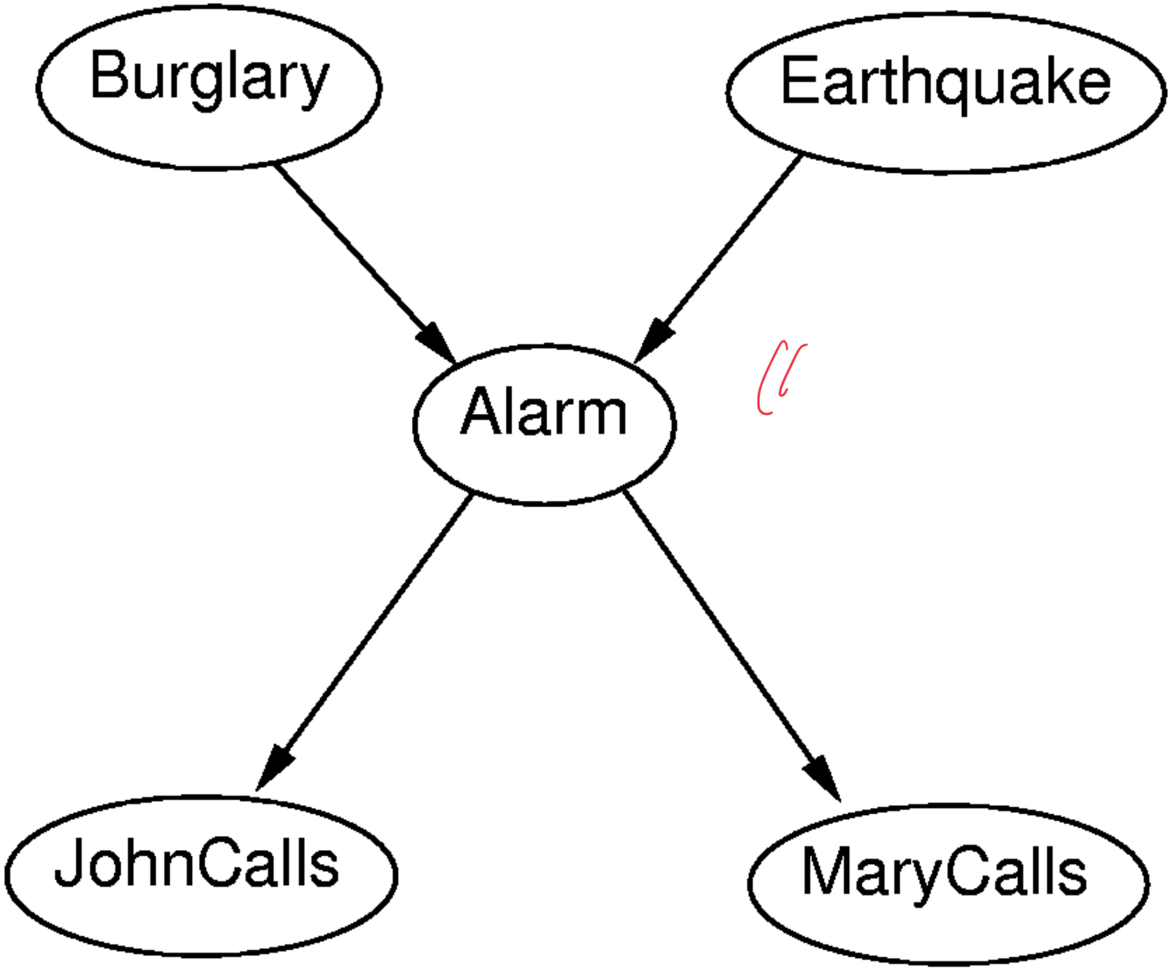


How should we connect the nodes? (3 min discussion)

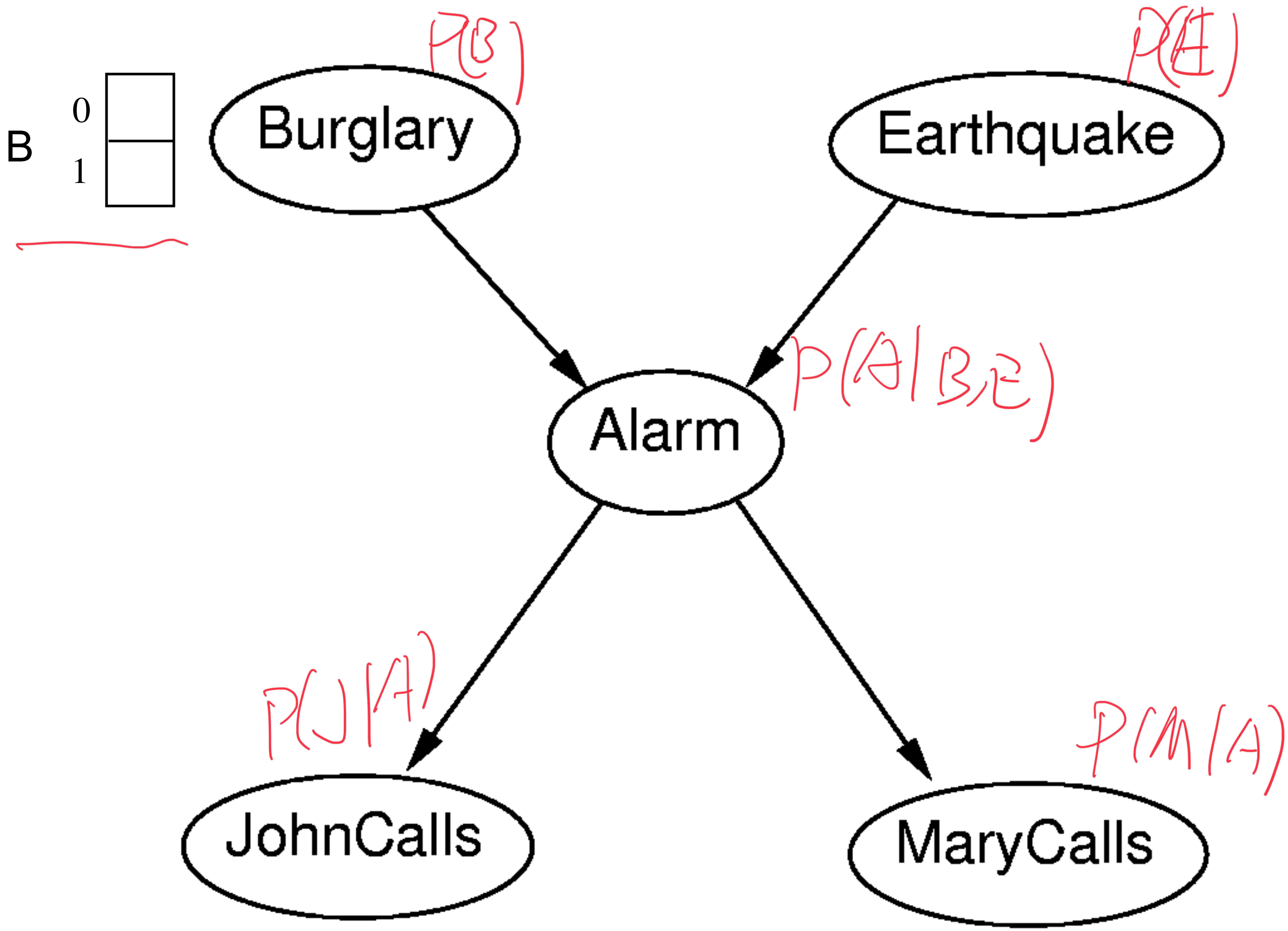


Links and CPTs?

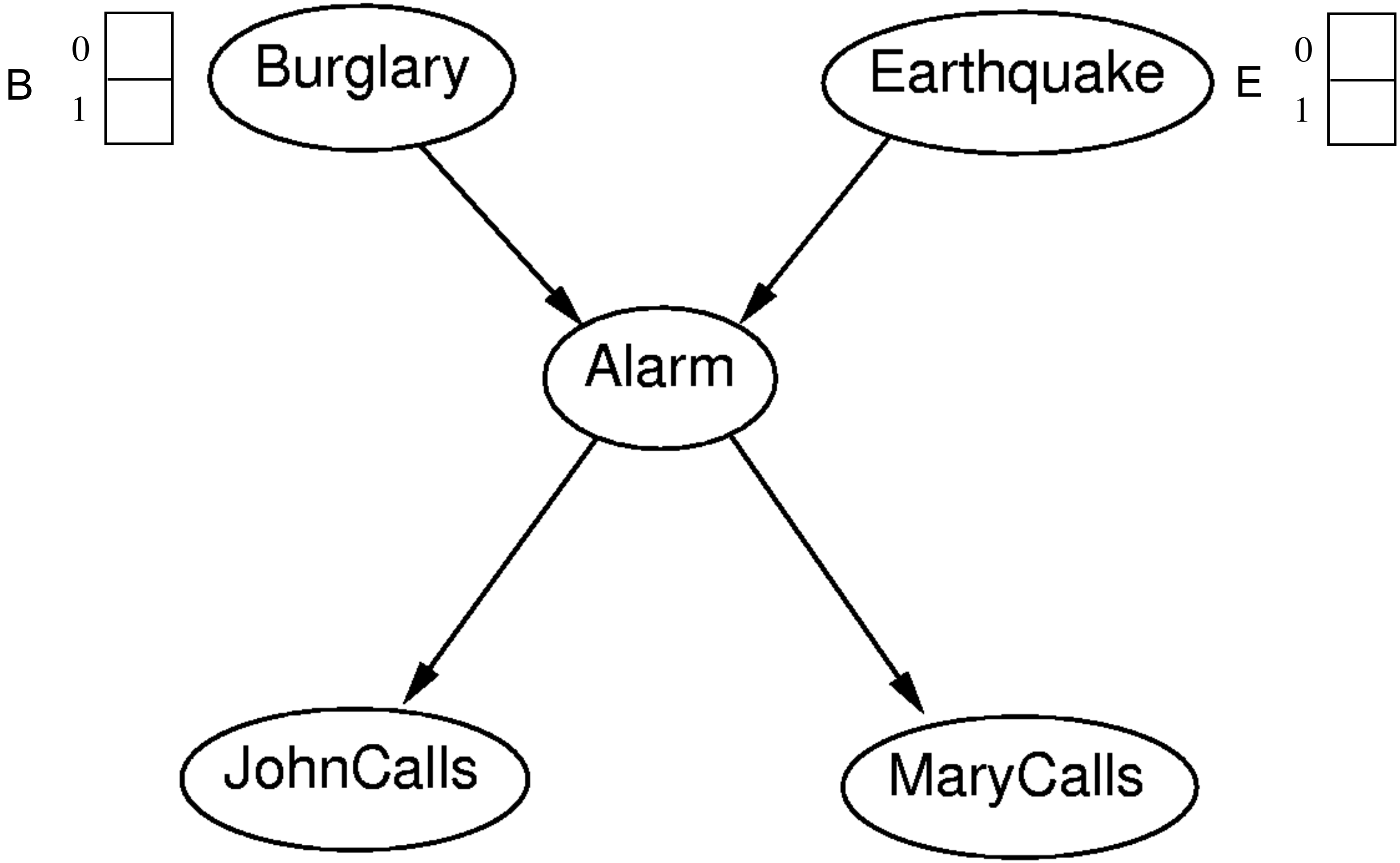
What are the CPTs? What are their dimensions?



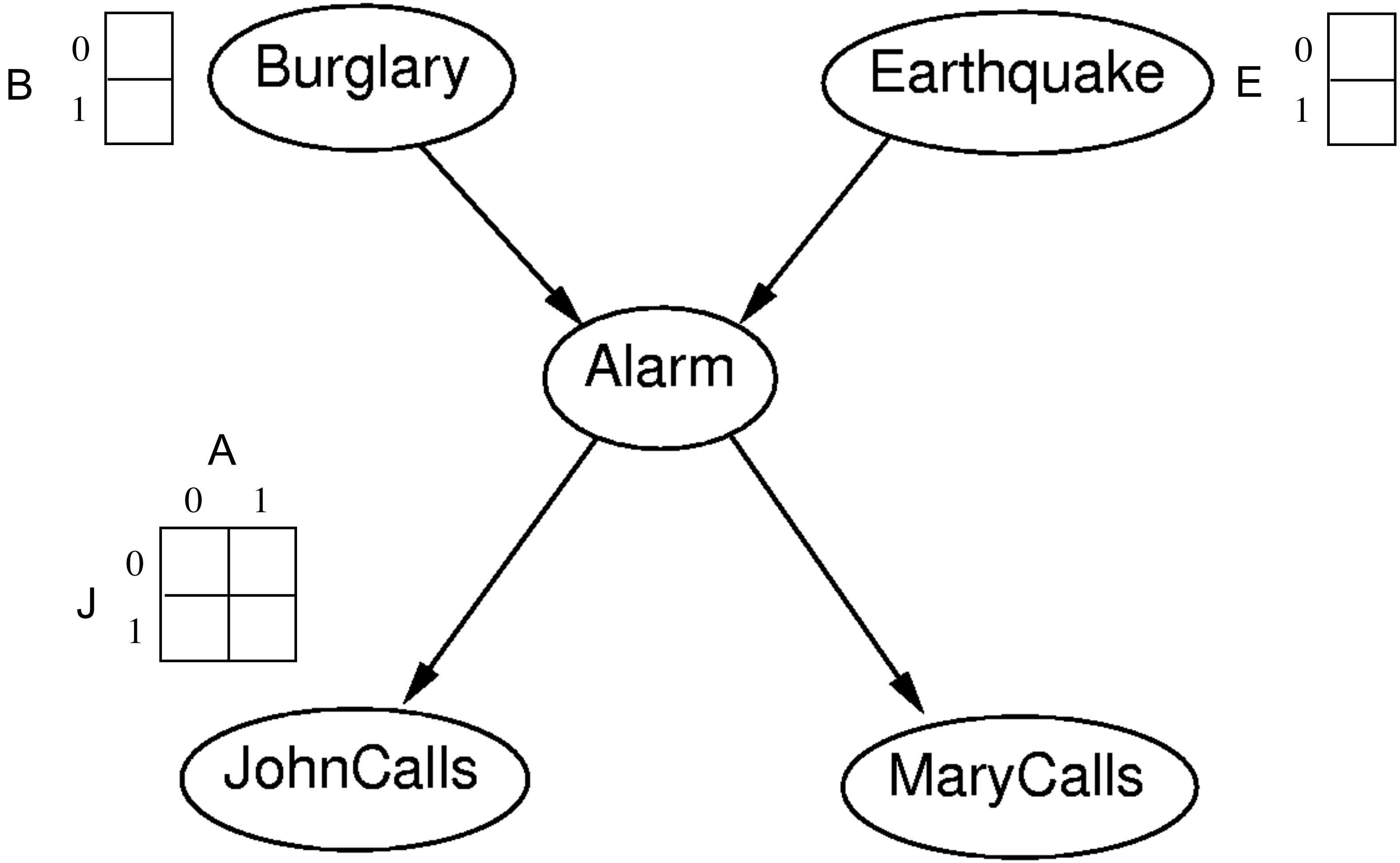
What are the CPTs? What are their dimensions?



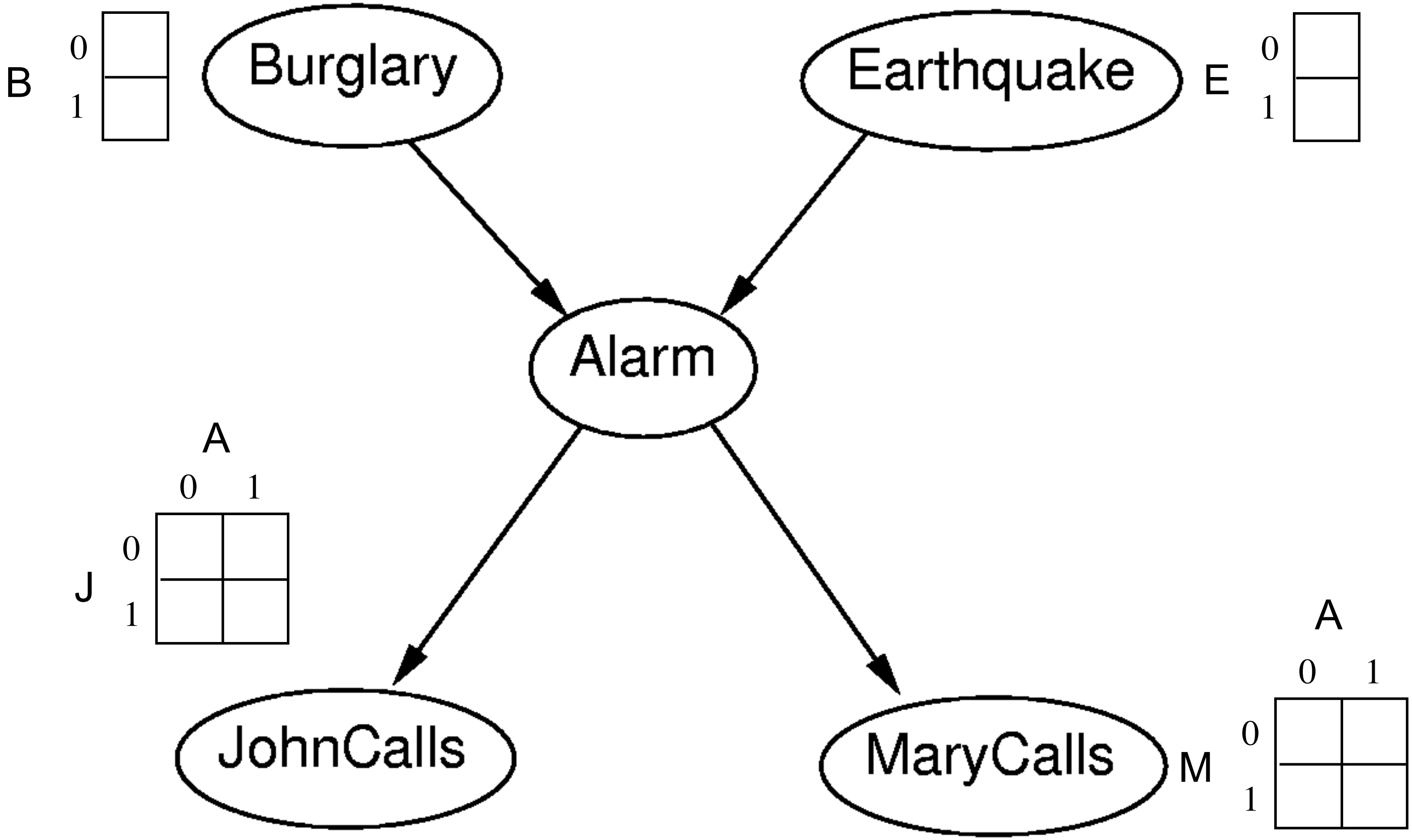
What are the CPTs? What are their dimensions?



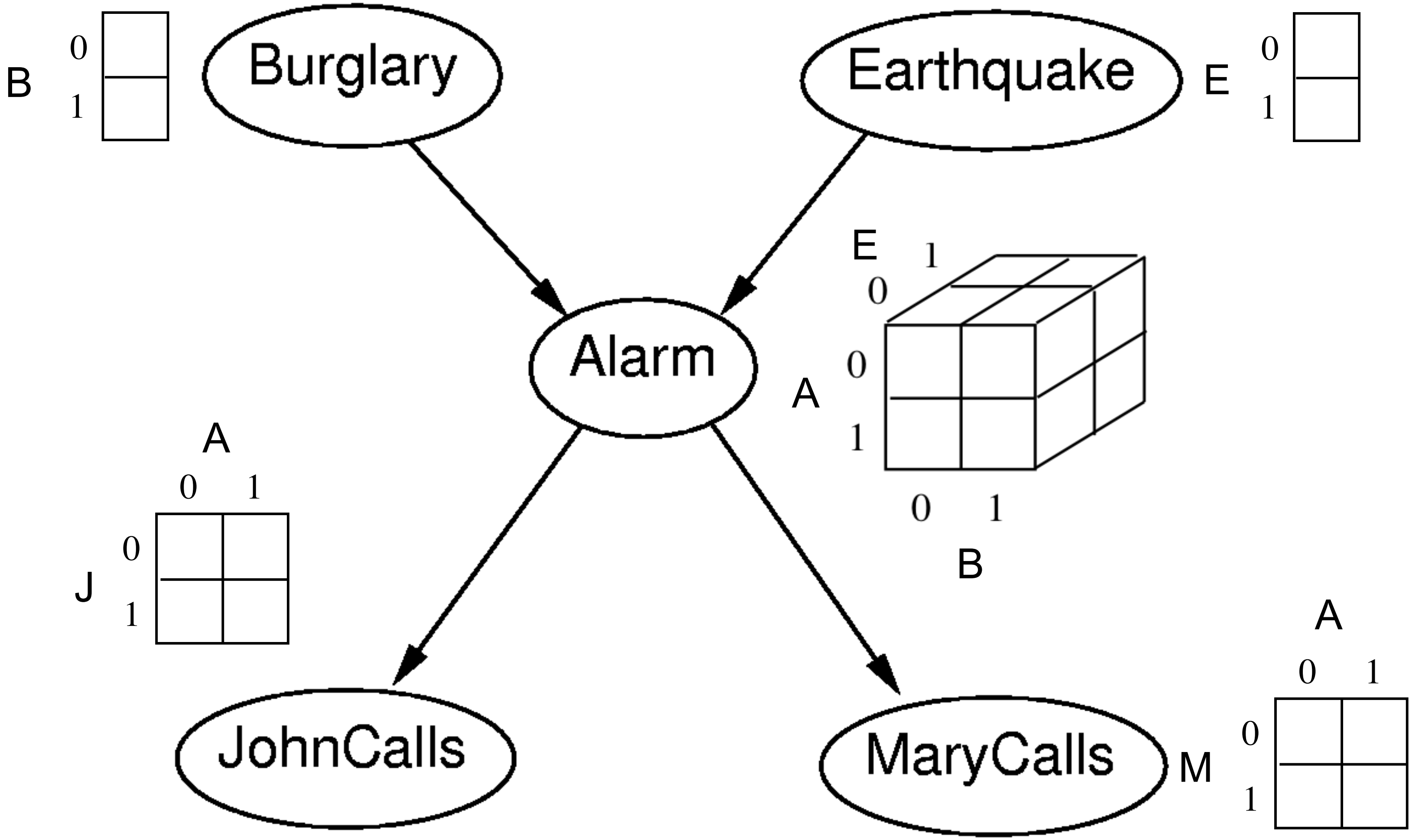
What are the CPTs? What are their dimensions?



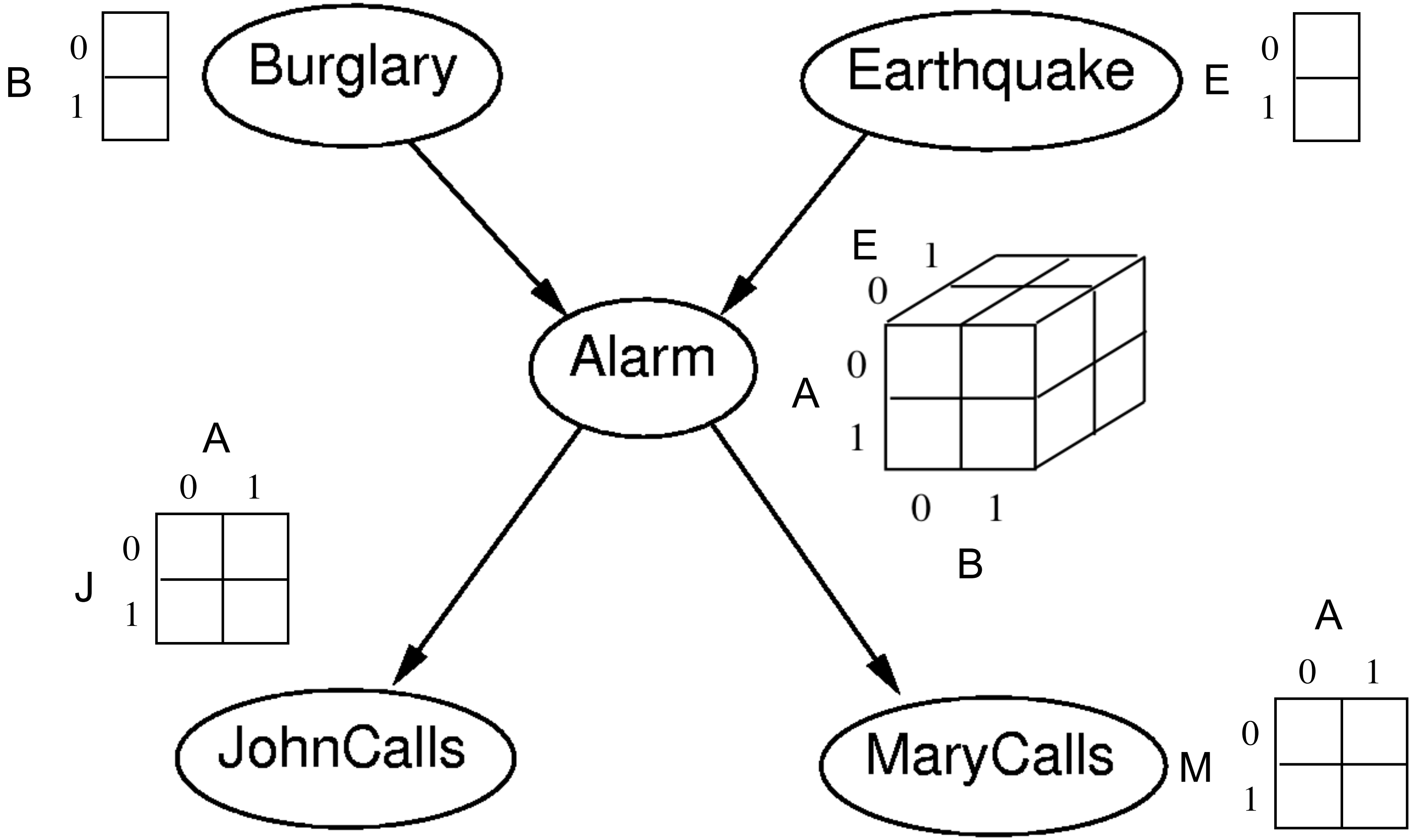
What are the CPTs? What are their dimensions?



What are the CPTs? What are their dimensions?

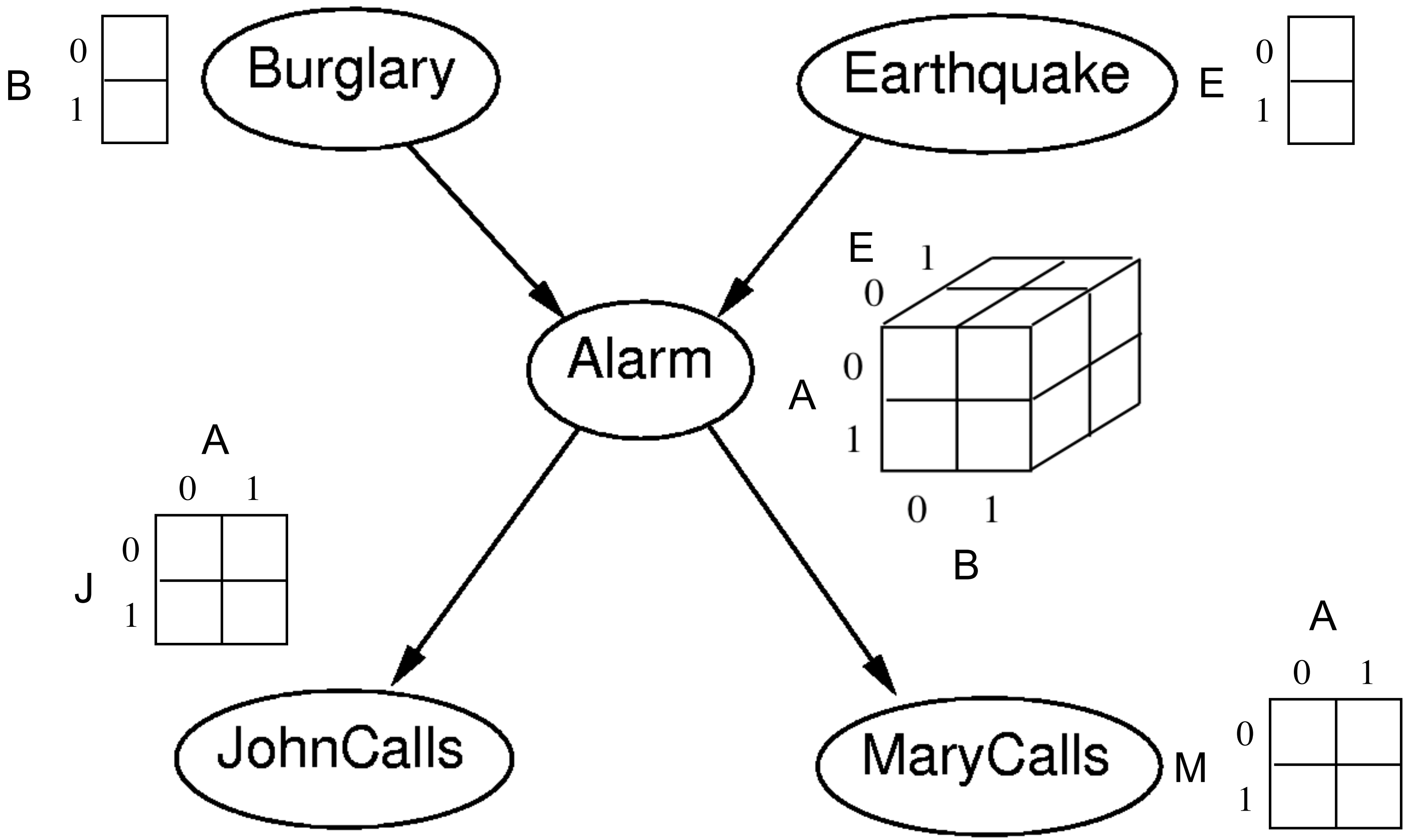


What are the CPTs? What are their dimensions?



Question: How to fill values into these CPTs?

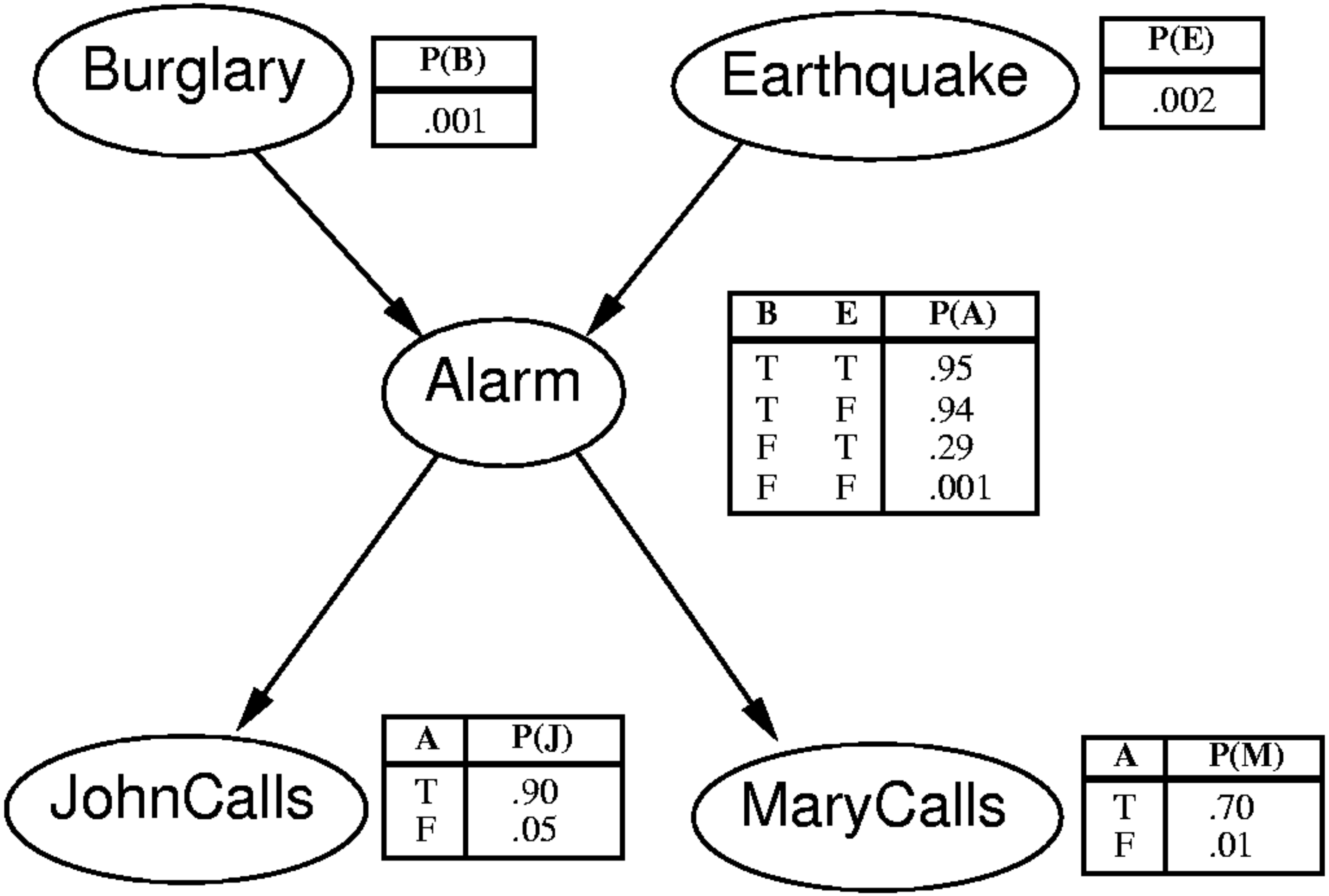
What are the CPTs? What are their dimensions?



Question: How to fill values into these CPTs?

Ans: Specify by hands. Learn from data (e.g., MLE).

Example



P(B)
.001

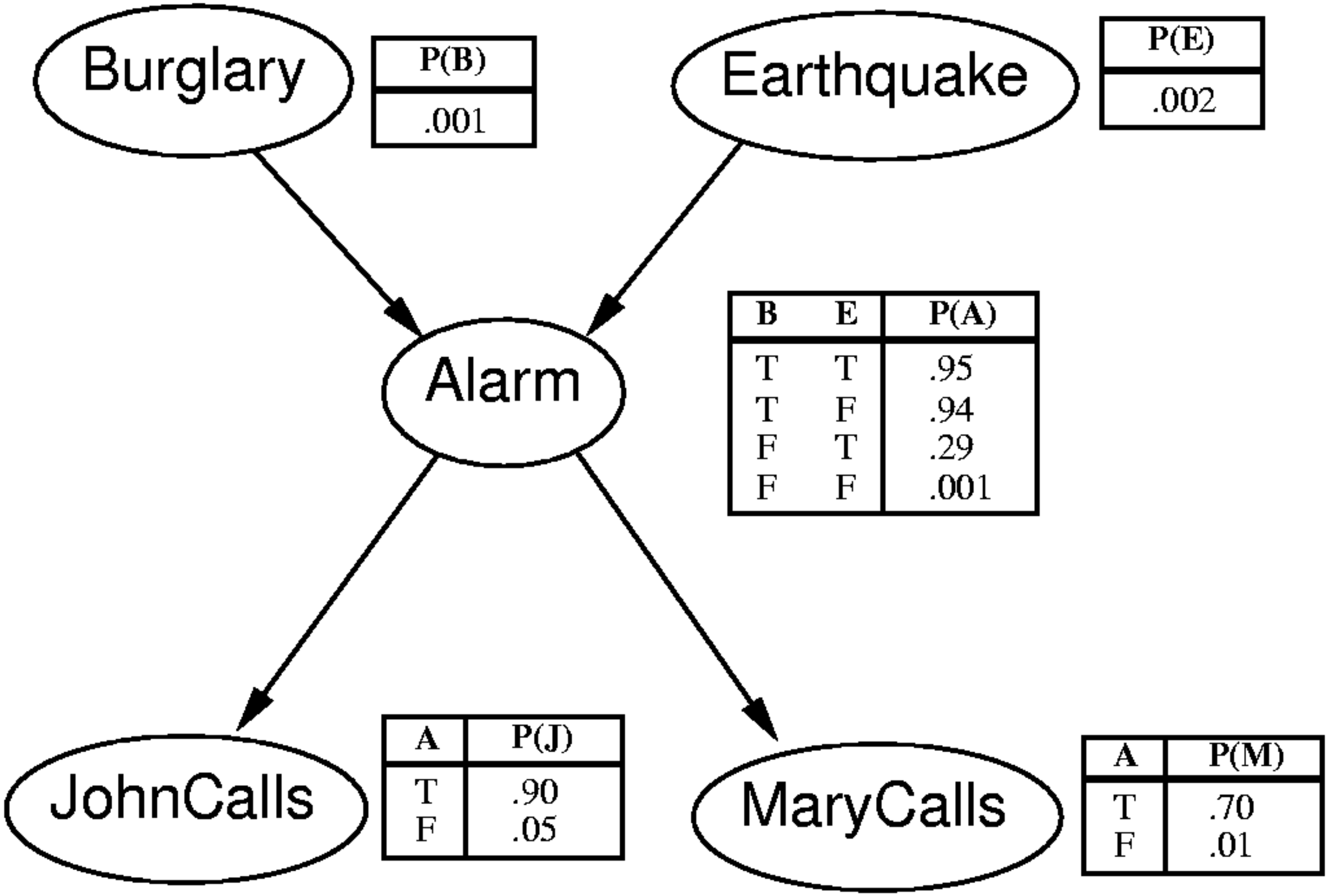
P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

Example



P(B)
.001

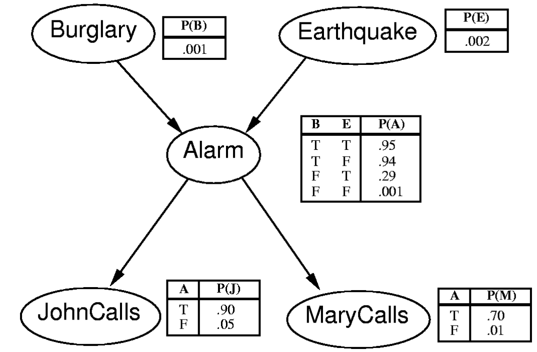
P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

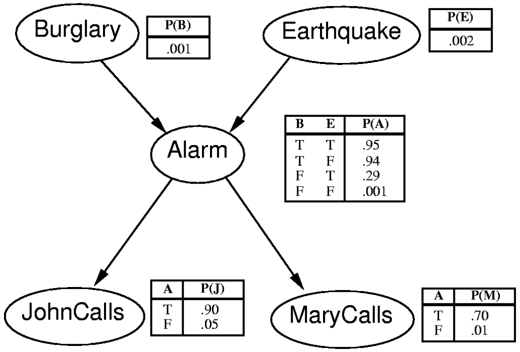
Joint probability? P(J, ¬M, A, B, ¬E)?



Calculate $P(J, \neg M, A, B, \neg E)$

Read the joint pf from the graph:

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

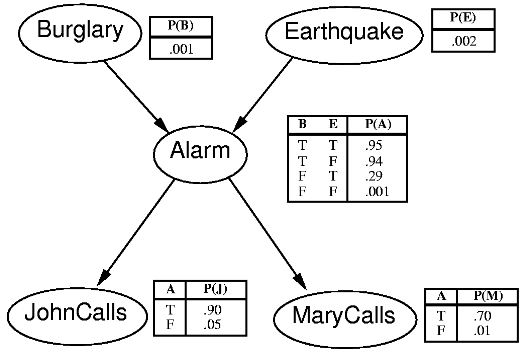


Calculate $P(J, \neg M, A, B, \neg E)$

Read the joint pf from the graph:

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

Plug in the desired values:



Calculate $P(J, \neg M, A, B, \neg E)$

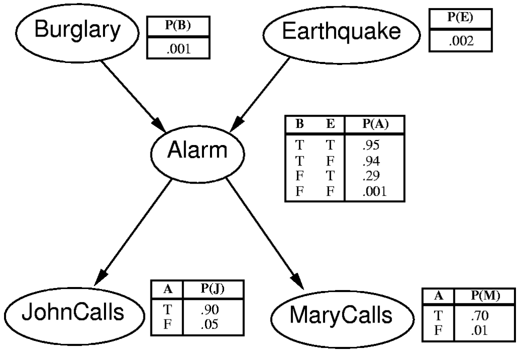
Read the joint pf from the graph:

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

Plug in the desired values:

$$P(J, \neg M, A, B, \neg E) = P(B)\ P(\neg E)\ P(A|B,\neg E)\ P(J|A)\ P(\neg M|A)$$

Calculate P(J, ¬M, A, B, ¬E)



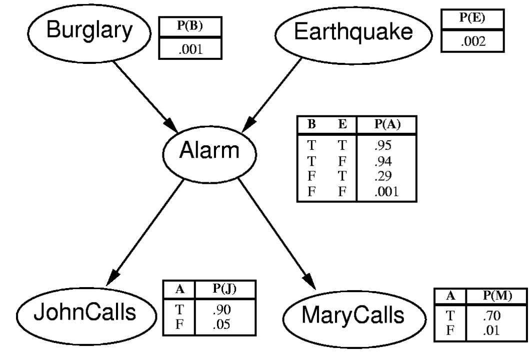
Read the joint pf from the graph:

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

Plug in the desired values:

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

= 0.001 * 0.998 * 0.94 * 0.9 * 0.3



Calculate P(J, ¬M, A, B, ¬E)

Read the joint pf from the graph:

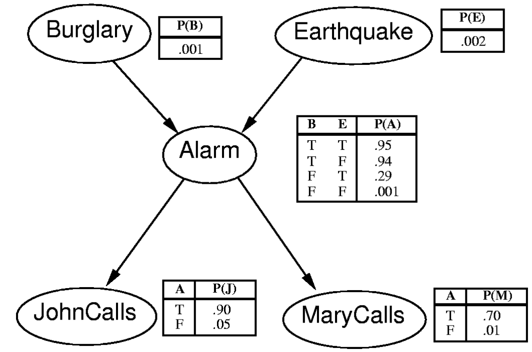
P(J, M, A, B, E) = P(B) P(E) P(A|B,E) P(J|A) P(M|A)

Plug in the desired values:

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

= 0.001 * 0.998 * 0.94 * 0.9 * 0.3

= 0.0002532924



Calculate P(J, ¬M, A, B, ¬E)

Read the joint pf from the graph:

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

Plug in the desired values:

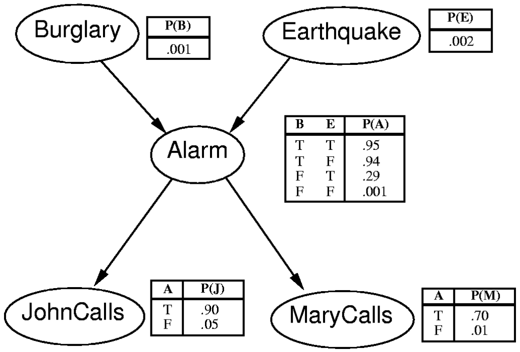
$$P(J, \neg M, A, B, \neg E) = P(B)\,P(\neg E)\,P(A|B,\neg E)\,P(J|A)\,P(\neg M|A)$$

$$= 0.001 * 0.998 * 0.94 * 0.9 * 0.3$$

$$= 0.0002532924$$

How about P(B | J, ¬M) ?

Calculate P(J, ¬M, A, B, ¬E)



Read the joint pf from the graph:

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

Plug in the desired values:

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

= 0.001 * 0.998 * 0.94 * 0.9 * 0.3

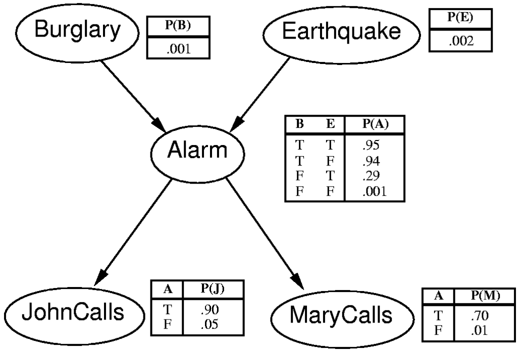
= 0.0002532924

How about P(B | J, ¬M) ?

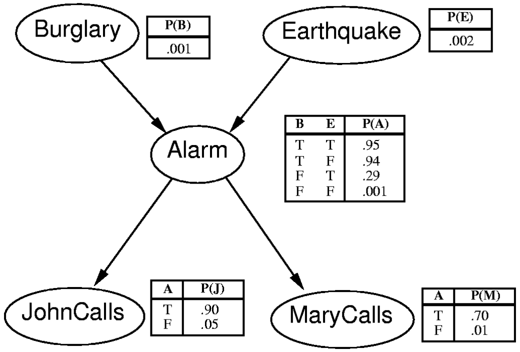
Remember, this means P(B=true | J=true, M=false)

Calculate P(B | J, ¬M)

$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M)}$$



Calculate P(B | J, ¬M)



$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M)}$$

By marginalization:

$$= \frac{\sum_i \sum_j P(J, \neg M, A_i, B, E_j)}{\sum_i \sum_j \sum_k P(J, \neg M, A_i, B_j, E_k)}$$

$$= \frac{\sum_i \sum_j P(B)P(E_j)P(A_i \mid B, E_j)P(J \mid A_i)P(\neg M \mid A_i)}{\sum_i \sum_j \sum_k P(B_j)P(E_k)P(A_i \mid B_j, E_k)P(J \mid A_i)P(\neg M \mid A_i)}$$

Quick checkpoint

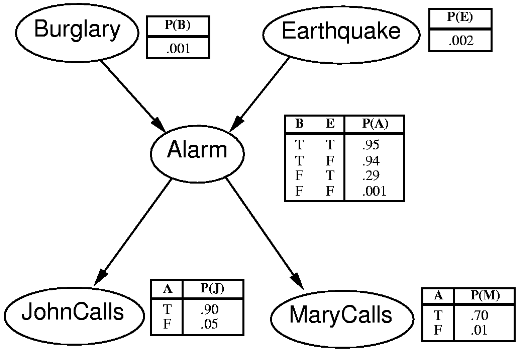
- Belief Net as a modelling tool
- By inspecting the cause-effect relationships, we can draw directed edges based on our domain knowledge
- The product of the CPTs give the joint distribution
  - We can calculate P(A | B) for any A and B
  - The factorization makes it computationally more tractable

Quick checkpoint

- Belief Net as a modelling tool
- By inspecting the cause-effect relationships, we can draw directed edges based on our domain knowledge
- The product of the CPTs give the joint distribution
  - We can calculate P(A | B) for any A and B
  - The factorization makes it computationally more tractable

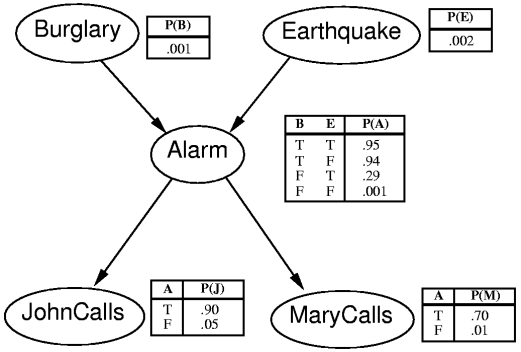
What else can we get?

Example: Conditional Independence



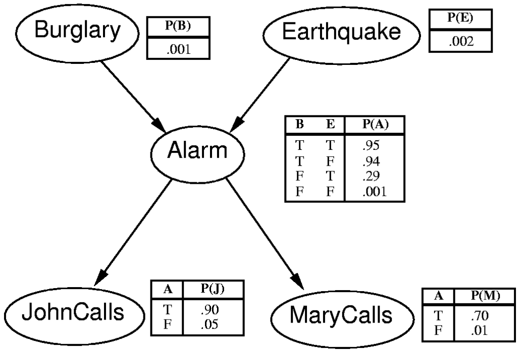
- Conditional independence is seen here
  - P(JohnCalls | MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls | Alarm)
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm

Example: Conditional Independence



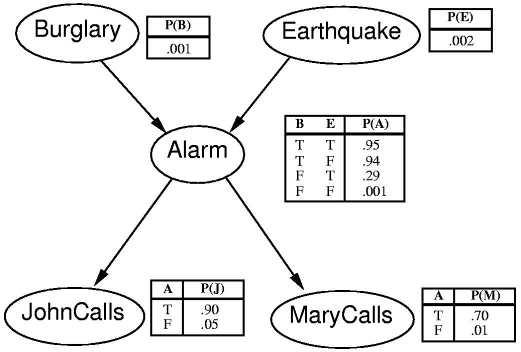
- Conditional independence is seen here
  - P(JohnCalls | MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls | Alarm)
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
- Does this mean that an earthquake or a burglary do not influence whether or not John calls?

Example: Conditional Independence



- Conditional independence is seen here
  - P(JohnCalls | MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls | Alarm)
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
- Does this mean that an earthquake or a burglary do not influence whether or not John calls?
  - No, but the influence is already accounted for in the Alarm variable
  - JohnCalls is <u>conditionally</u> independent of Earthquake, but not <u>absolutely</u> independent of it

Example: Conditional Independence



- Conditional independence is seen here
  - P(JohnCalls | MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls | Alarm)
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
- Does this mean that an earthquake or a burglary do not influence whether or not John calls?
  - No, but the influence is already accounted for in the Alarm variable
  - JohnCalls is <u>conditionally</u> independent of Earthquake, but not <u>absolutely</u> independent of it

***This conclusion is independent to values of CPTs!**

Question

If X and Y are independent, are they therefore independent given any variable(s)?

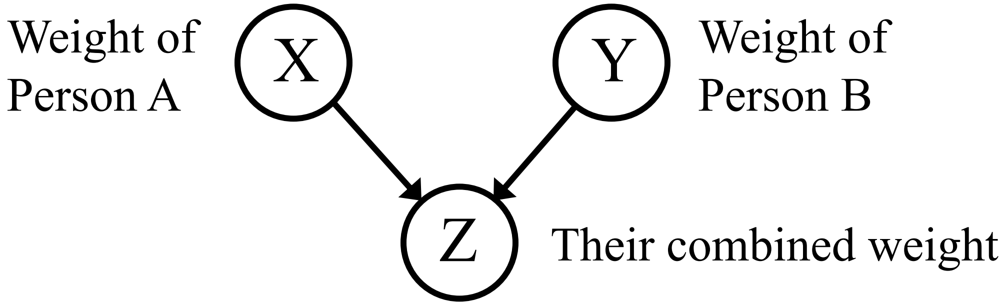
I.e., if $P(X, Y) = P(X)\, P(Y)$ [i.e., if $P(X|Y) = P(X)$], can we conclude that
$P(X \mid Y, Z) = P(X \mid Z)$?

Question

If X and Y are independent, are they therefore independent given any variable(s)?

I.e., if $P(X, Y) = P(X) P(Y)$ [i.e., if $P(X|Y) = P(X)$], can we conclude that
$P(X \mid Y, Z) = P(X \mid Z)$?

The answer is **no**, and here's a counter example:



$$P(X \mid Y) = P(X)$$
$$P(X \mid Y, Z) \neq P(X \mid Z)$$

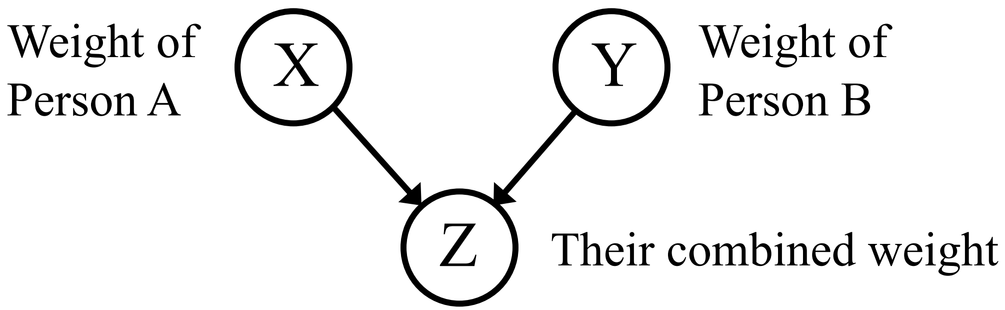
Note: Even though Z is a deterministic function of X and Y, it is still a random variable with a probability distribution

Question

If X and Y are independent, are they therefore independent given any variable(s)?

I.e., if $P(X, Y) = P(X)\,P(Y)$ [i.e., if $P(X|Y) = P(X)$], can we conclude that
$P(X \mid Y, Z) = P(X \mid Z)$?

The answer is **no**, and here's a counter example:



$P(X \mid Y) = P(X)$
$P(X \mid Y, Z) \neq P(X \mid Z)$

Note: Even though Z is a deterministic function of X and Y, it is still a random variable with a probability distribution

***Again: This conclusion is independent to values of CPTs!**

Key points of today's lecture

- Probability notations
  - Distinguish between events and random variables, apply rules of probabilities
- Representing a joint-distribution
  - number of parameters exponential in the number of variables
  - Calculating marginals and conditionals from the joint-distribution.
- Conditional independences and factorization of joint-distributions
  - Saves parameters, often exponential improvements
- Intro to Bayesian networks / directed graphical models.

Next lectures

- More Bayesian networks, directed graphical models.
- Read off conditional independences from the graph!
- More examples