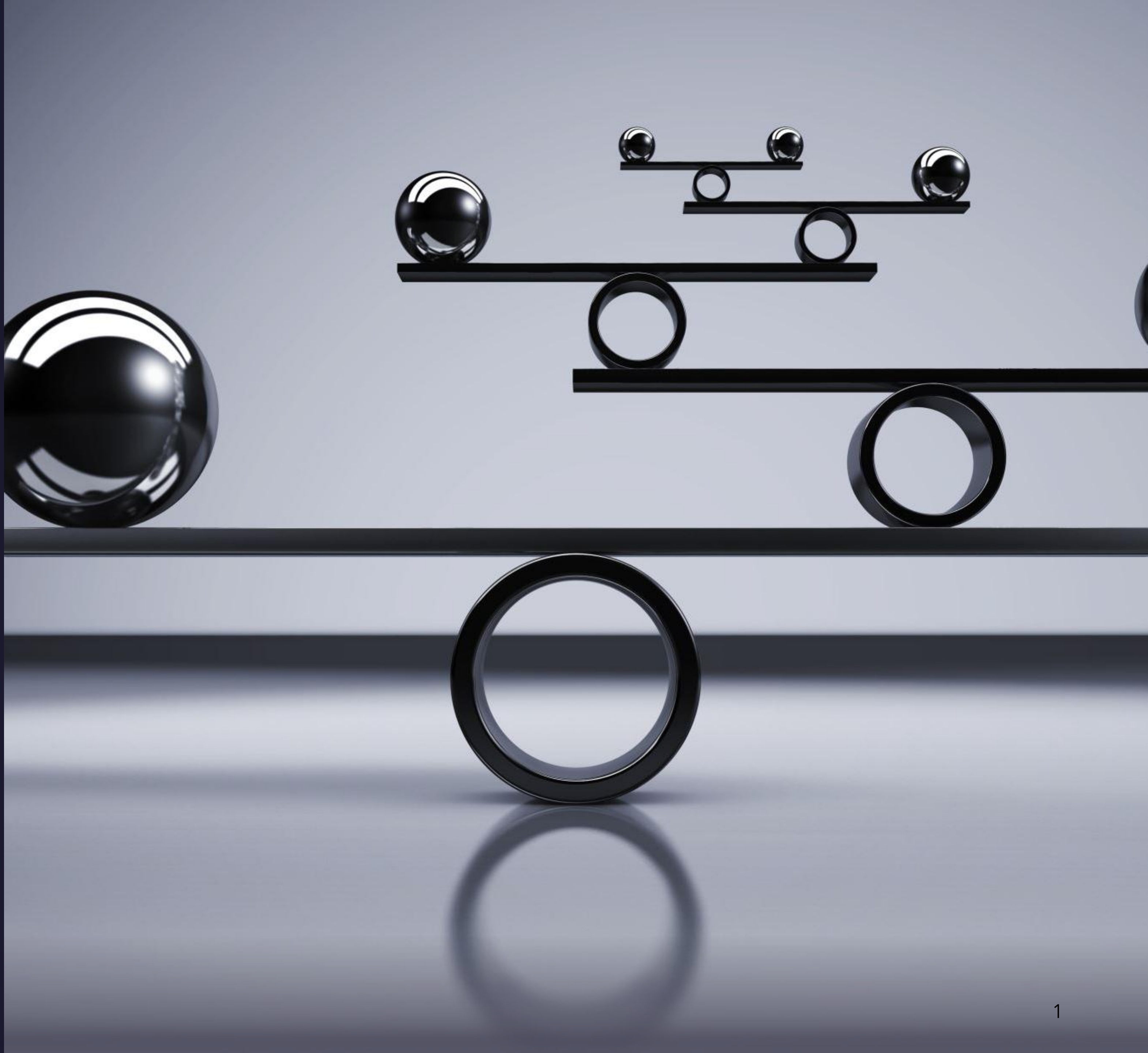
Chapter 14

Uncertain Knowledge & Reasoning



Bayesian Networks

- AKA belief network, probabilistic network, causal network, and knowledge map.
- **Bayesian network** is used to represent the dependencies among variables
- **Bayesian network:** is a directed graph in which each node is annotated with quantitative probability information

Bayesian Network

Definition:

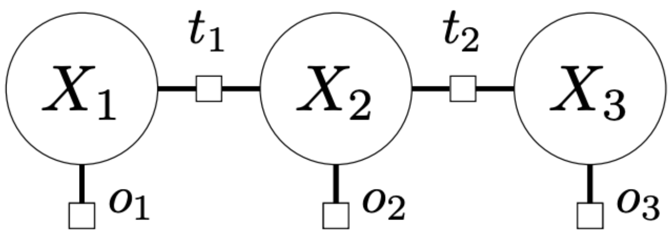
- Let $X = (X_1, \ldots, X_n)$ be random variables.
- A Bayesian network is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$P(X_1 = x_1, \ldots, X_n = x_n) \stackrel{\text{def}}{=} \prod_{i=1}^{n} p(x_i \mid x_{Parents(i)})$$

Markov networks vs. Bayesian network

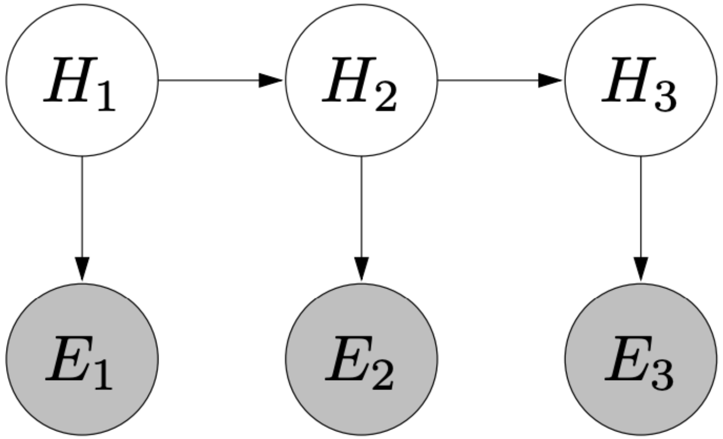
MARKOV NETWORKS

- factors can be arbitrary
- arbitrary set of preferences and constraints



BAYESIAN NETWORK

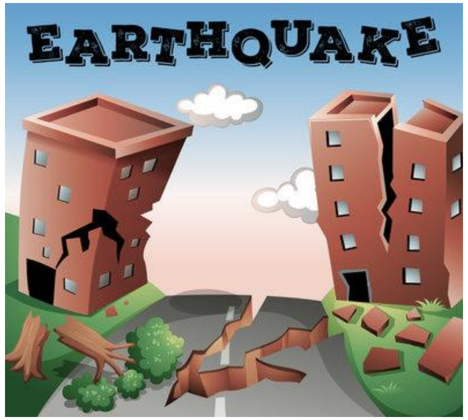
- factors are local conditional probabilities
- define a generative process represented by a directed graph



Applications

- Language modeling
- Document classification (Naïve Bayes)
- Topic modeling (Latent Dirichlet Allocation (LDA))
- Medical diagnosis
- Social network analysis

Example



Question:
Does hearing that there's an earthquake increase, decrease, or keep constant the probability of a burglary?

- $P(B = 1 \mid A = 1)$
- $P(B = 1 \mid A = 1, E = 1)$

Bayesian Network Components

- **Bayesian network:**

1. Each node corresponds to a random variable, which may be discrete or continuous.
2. A set of directed links connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y. The graph has no directed cycles (DAG).
3. Each node X_i has a conditional probability distribution $P(X_i | \text{Parents}(X_i))$ that quantifies the effect of the parents on the node.
4. A joint distribution which is produced by multiplying all the local conditional distributions together

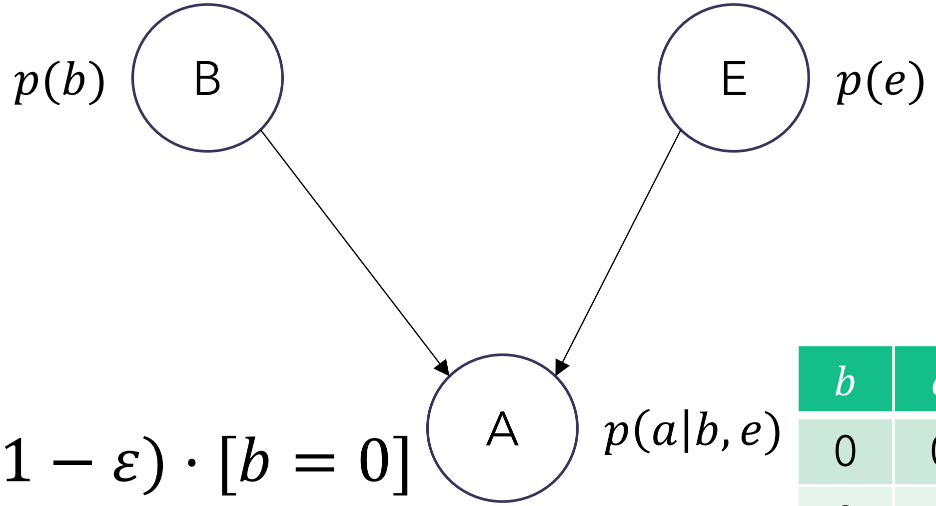
Bayesian Network Components

- **Bayesian network:**

1. Each node corresponds to a random variable, which may be discrete or continuous: Burglar, Earthquake, Alarm
2. A set of directed links connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y. The graph has no directed cycles (DAG): Burglars and earthquakes cause alarms
3. Each node X_i has a conditional probability distribution $P(X_i | \text{Parents}(X_i))$ that quantifies the effect of the parents on the node.
4. A joint distribution which is produced by multiplying all the local conditional distributions together

Joint Distribution

b	$p(b)$
0	$1-\varepsilon$
1	ε



e	$p(e)$
0	$1-\varepsilon$
1	ε

b	e	a	$p(a|b,e)$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

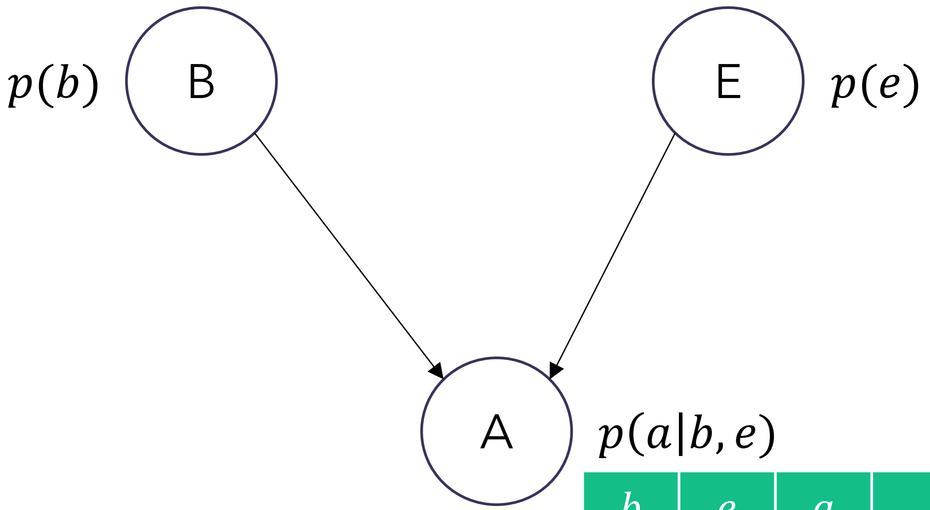
$$p(b) = \varepsilon \cdot [b = 1] + (1-\varepsilon) \cdot [b = 0]$$
$$p(e) = \varepsilon \cdot [e = 1] + (1-\varepsilon) \cdot [e = 0]$$
$$p(a\,|b, e) = [a = (b \vee e)]$$

The Joint Distribution is:

$$\mathbb{P}(B = b, E = e, A = a) \stackrel{\text{def}}{=} p(b)\, p(e)\, p(a|b,e)$$

Joint Distribution

b	$p(b)$
0	$1-\varepsilon$
1	ε



e	$p(e)$
0	$1-\varepsilon$
1	ε

b	e	a	$p(a\|b,e)$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

The Joint Distribution is:

$$\mathbb{P}(B=b,E=e,A=a) \overset{\text{def}}{=} p(b)\,p(e)\,p(a|b,e)$$

b	e	a	$p(b)$	$p(e)$	$p(a\|b,e)$	$\mathbb{P}(B=b,E=e,A=a)$
0	0	0	$1-\varepsilon$	$1-\varepsilon$	1	$(1-\varepsilon)^2$
0	0	1	$1-\varepsilon$	$1-\varepsilon$	0	0
0	1	0	$1-\varepsilon$	ε	0	0
0	1	1	$1-\varepsilon$	ε	1	$(1-\varepsilon)\varepsilon$
1	0	0	ε	$1-\varepsilon$	0	0
1	0	1	ε	$1-\varepsilon$	1	$(1-\varepsilon)\varepsilon$
1	1	0	ε	ε	0	0
1	1	1	ε	ε	1	ε^2

Probabilistic inference

- Probabilistic inference allows you to ask questions about the world
  - World is represented by the random variables X
- Given a Bayesian network $\mathbb{P}(X_1, \ldots, X_n)$ representing a probabilistic database:
  - a set of evidence variables E and values e, where $E = e$ and $E \subseteq X$
  - a set of query variables $Q \subseteq X$
- Result: Calculate the probability of the query variables, given the evidence, marginalize out all other variables: $\mathbb{P}(Q \mid E = e)$
  - $\mathbb{P}(Q = q \mid E = e)$ for all values q

What is the probability of burglary without any evidence?

The Joint Distribution is:

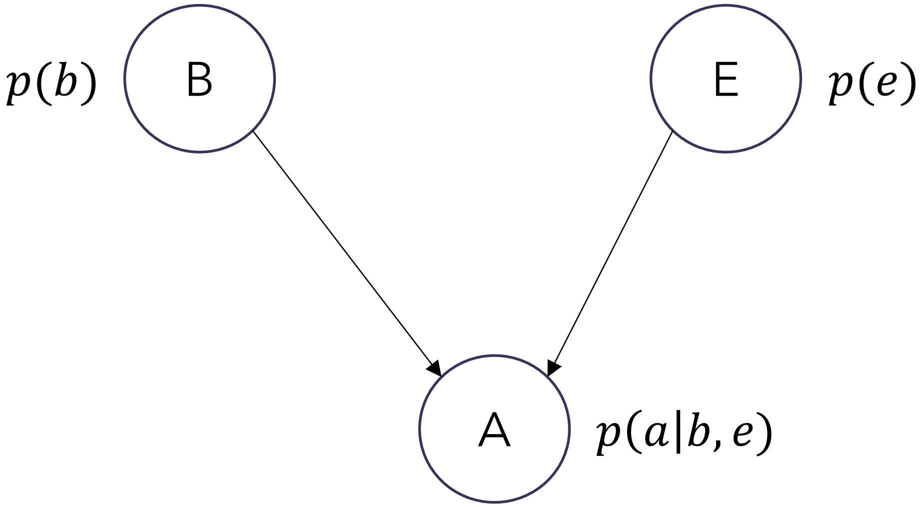
$$\mathbb{P}(B = b, E = e, A = a) \overset{\text{def}}{=} p(b)\, p(e)\, p(a|b, e)$$

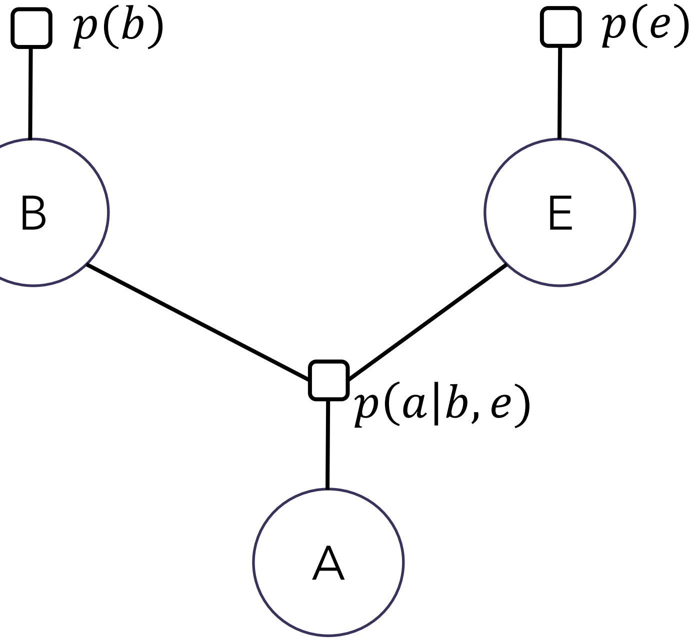
b	e	a	$\boldsymbol{p(b)}$	$\boldsymbol{p(e)}$	$\boldsymbol{p(a \mid b, e)}$	$\mathbb{P}(B = b, E = e, A = a)$
0	0	0	$1 - \varepsilon$	$1 - \varepsilon$	1	$(1 - \varepsilon)^2$
0	0	1	$1 - \varepsilon$	$1 - \varepsilon$	0	0
0	1	0	$1 - \varepsilon$	ε	0	0
0	1	1	$1 - \varepsilon$	ε	1	$(1 - \varepsilon)\varepsilon$
1	0	0	ε	$1 - \varepsilon$	0	0
1	0	1	ε	$1 - \varepsilon$	1	$(1 - \varepsilon)\varepsilon$
1	1	0	ε	ε	0	0
1	1	1	ε	ε	1	ε^2

$$P(B = 1) = \varepsilon(1 - \varepsilon) + \varepsilon^2 = \varepsilon$$

Inference via Reduction to Markov Networks

- The joint distribution is the product of all the local conditional distributions
- The local conditional distributions $p(a \mid b, e)$ are all non-negative, so they can be interpreted as simply factors in a factor graph





Inference via Reduction to Markov Networks

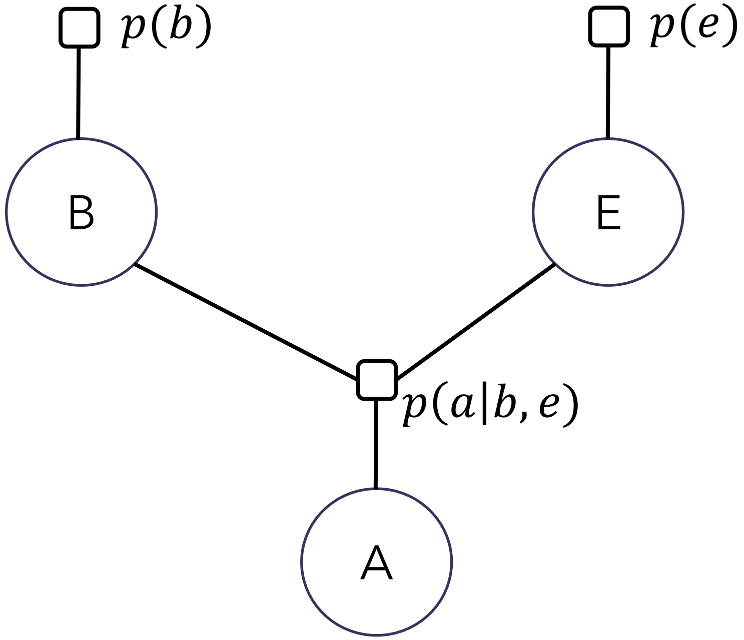
- **Markov networks** defines the joint distribution as the product of all the factors divided by some normalization constant Z:

$$\mathbb{P}(X = x) = \frac{Weight(x)}{\sum_x Weight(x)} = \frac{\prod_{j=1}^{m} f_j(x)}{Z}$$

- **Bayesian Networks** also define a probability distribution:

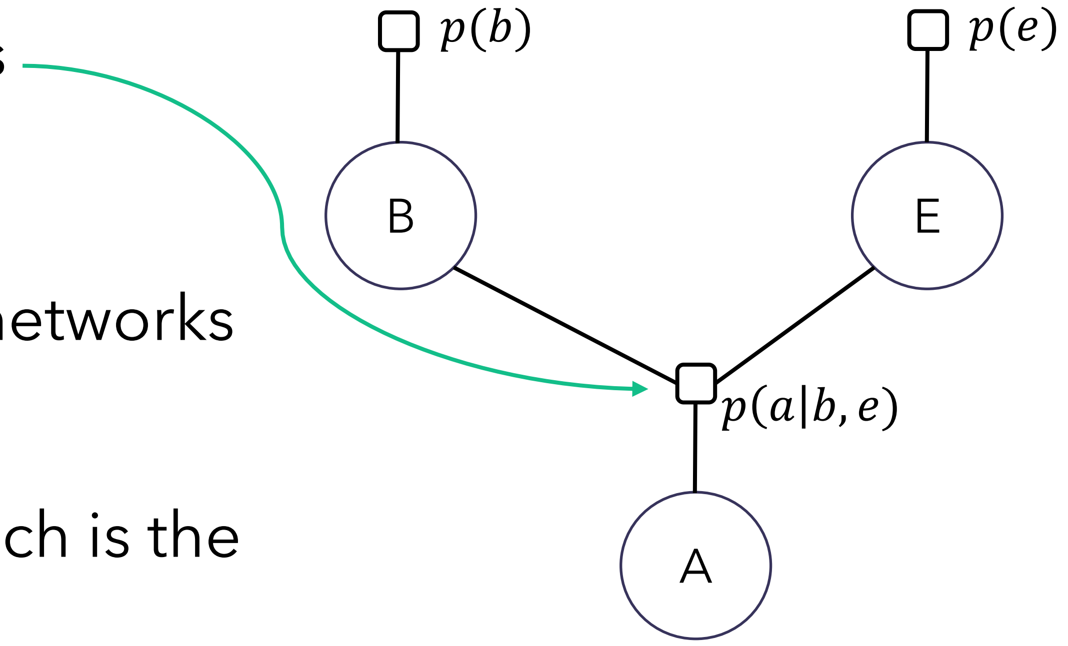
$$\mathbb{P}(X = x) = \prod_{i=1}^{n} p(x_i \mid x_{Parents(i)})$$

- Here, $Z = 1$ because the factors are local conditional distributions of a Bayesian network



Inference via Reduction to Markov Networks

- Single factor that connects all the parents
- NOT two factors, one per arrow!
- Run any inference algorithm for Markov networks (Gibbs sampling) $P(B = 1)$
- But there is something that's missing, which is the ability to condition on evidence



Conditioning on evidence

What is the probability of burglary given the alarm rang?

$$P(B=1|A=1) = \frac{\varepsilon(1-\varepsilon)+\varepsilon^2}{0+\varepsilon(1-\varepsilon)+\varepsilon^2+\varepsilon(1-\varepsilon)} = \frac{1}{2-\varepsilon}$$

b	e	a	$\boldsymbol{p(b)}$	$\boldsymbol{p(e)}$	$\boldsymbol{p(a\|b,e)}$	$\mathbb{P}(B=b,E=e,A=a)$
0	0	0	$1-\varepsilon$	$1-\varepsilon$	1	$(1-\varepsilon)^2$
0	0	1	$1-\varepsilon$	$1-\varepsilon$	0	0
0	1	0	$1-\varepsilon$	ε	0	0
0	1	1	$1-\varepsilon$	ε	1	$(1-\varepsilon)\varepsilon$
1	0	0	ε	$1-\varepsilon$	0	0
1	0	1	ε	$1-\varepsilon$	1	$(1-\varepsilon)\varepsilon$
1	1	0	ε	ε	0	0
1	1	1	ε	ε	1	ε^2

What is the probability of burglary given the alarm rang and there was an earthquake?

The Joint Distribution is:

$$\mathbb{P}(B = b, E = e, A = a) \overset{\text{def}}{=} p(b)\, p(e)\, p(a|b,e)$$

b	e	a	$\boldsymbol{p(b)}$	$\boldsymbol{p(e)}$	$\boldsymbol{p(a\|b,e)}$	$\mathbb{P}(B = b, E = e, A = a)$
0	0	0	$1-\varepsilon$	$1-\varepsilon$	1	$(1-\varepsilon)^2$
0	0	1	$1-\varepsilon$	$1-\varepsilon$	0	0
0	1	0	$1-\varepsilon$	ε	0	0
0	1	1	$1-\varepsilon$	ε	1	$(1-\varepsilon)\varepsilon$
1	0	0	ε	$1-\varepsilon$	0	0
1	0	1	ε	$1-\varepsilon$	1	$(1-\varepsilon)\varepsilon$
1	1	0	ε	ε	0	0
1	1	1	ε	ε	1	ε^2

$$P(B = 1|A = 1, E = 1) = \frac{\varepsilon^2}{\varepsilon^2 + \varepsilon(1-\varepsilon)} = \varepsilon$$

Question

$$P(B = 1|A = 1) = \frac{\varepsilon(1-\varepsilon) + \varepsilon^2}{0 + \varepsilon(1-\varepsilon) + \varepsilon^2 + \varepsilon(1-\varepsilon)} = \frac{1}{2-\varepsilon}$$

$$P(B = 1|A = 1, E = 1) = \frac{\varepsilon^2}{\varepsilon^2 + \varepsilon(1-\varepsilon)} = \varepsilon$$

- Does an earthquake decrease the probability of a burglary? No!

Key idea: explaining away!

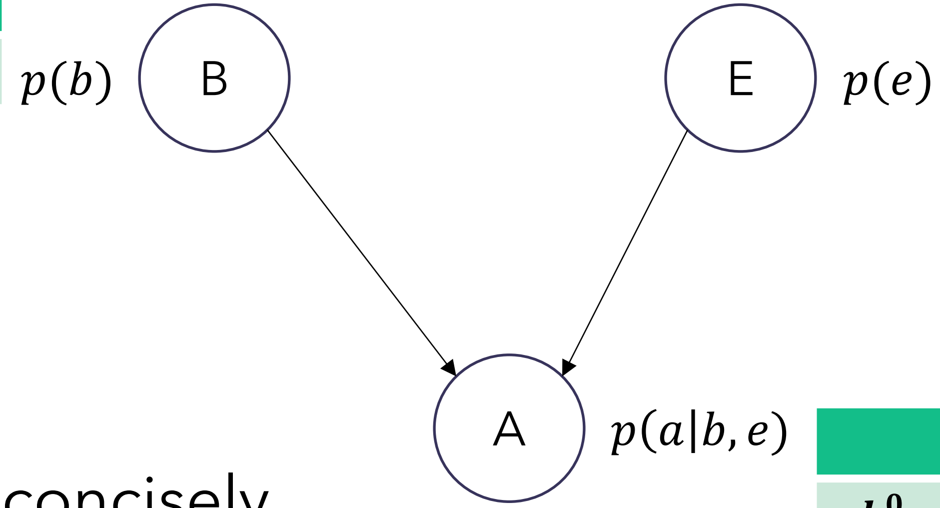
Suppose two causes (E,B) positively influence an effect (A). Conditioned on the effect, further conditioning on one cause reduces the probability of the other cause:

$$P(B = 1 \,|\, A = 1, E = 1) < P(B = 1 \,|\, A = 1)$$

Note: happens even if causes are independent!

Note

b^0	b^1
0.95	0.05



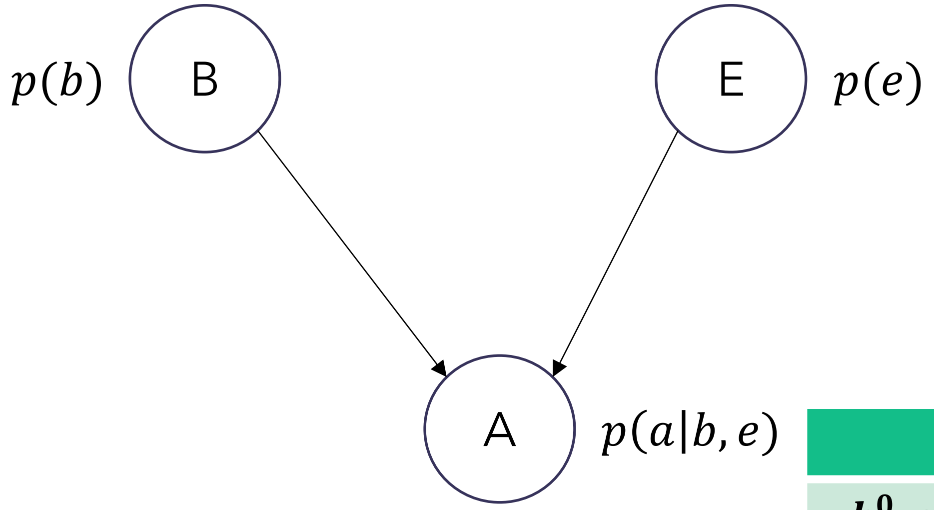
e^0	e^1
0.95	0.05

	a^0	a^1
b^0, e^0	1	0
b^0, e^1	0	1
b^1, e^0	0	1
b^1, e^1	0	1

- Probabilities can be written concisely
- Assume $\varepsilon = 0.05$

Example (1)

b^0	b^1
0.95	0.05



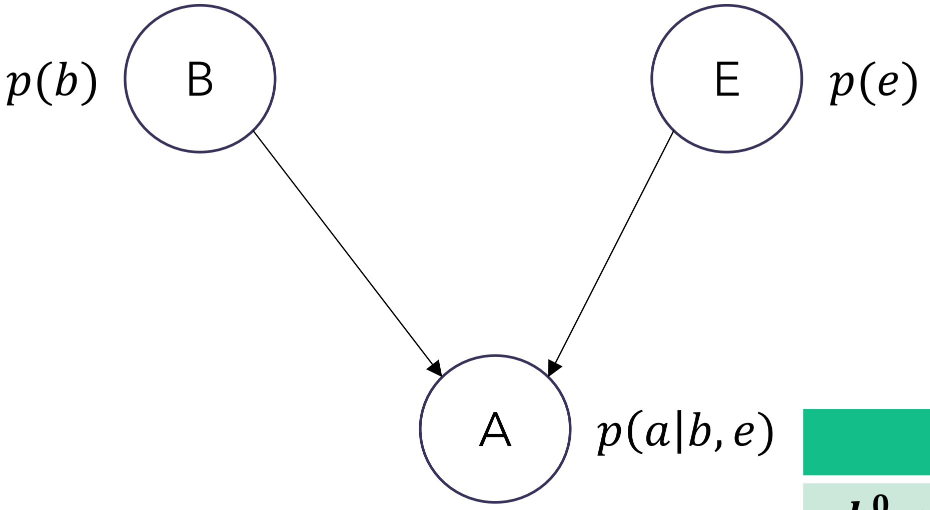
e^0	e^1
0.95	0.05

	a^0	a^1
b^0, e^0	1	0
b^0, e^1	0	1
b^1, e^0	0	1
b^1, e^1	0	1

- Assume $\varepsilon = 0.05$
- What is the probability of a burglary happening?
  - $\mathrm{P}(b = 1) = 0.05$
- What is the joint probability of a burglary, alarm, and no earthquake?
  - $\mathbb{P}(b = 1, e = 0, a = 1) = 0.05 * 0.95 * 1 = 0.0475$

Example (2)

b^0	b^1
0.95	0.05



e^0	e^1
0.95	0.05

	a^0	a^1
b^0, e^0	1	0
b^0, e^1	0	1
b^1, e^0	0	1
b^1, e^1	0	1

- Recall:
  - $P(a|b) = \frac{P(a \wedge b)}{P(b)} = \frac{1}{P(b)} P(a \wedge b) = \alpha P(a,b)$
- Given that the alarm rings, what is the probability of a burglary?
- A query can be answered using a Bayesian network by computing sums of products of conditional probabilities from the network
- $P(b|a) = \alpha P(a,b) = \alpha \sum_e P(a,b,e) = \alpha \sum_e P(b)P(e)P(a|b,e)$

$= \alpha P(b) \sum_e P(e)P(a|b,e)$

Given that the alarm rings, what is the probability of a burglary?
Note: we don't know anything about the earthquake.

Example (2)

$$P(b|a) = \alpha P(a,b) = \alpha \sum_e P(a,b,e) = \alpha \sum_e P(b)P(e)P(a|b,e)$$
$$= \alpha P(b) \sum_e P(e)P(a|b,e)$$

$$\mathbb{P}(b=1, e=0, a=1) = 0.05 * 0.95 * 1 = 0.0475$$

$$\mathbb{P}(b=1, e=1, a=1) = 0.05 * 0.05 * 1 = 0.0025$$

sum = 0.05

$$\mathbb{P}(b=0, e=0, a=1) = 0.95 * 0.95 * 0 = 0$$

$$\mathbb{P}(b=0, e=1, a=1) = 0.95 * 0.05 * 1 = 0.0475$$

sum = 0.0475

Normalize

b	$p(b)$
0	$\frac{0.0475}{0.05 + 0.0475} = 0.4871$
1	$\frac{0.05}{0.05 + 0.0475} = 0.5128$

- So, when the alarm goes off, the probability of a burglary increases!

➢ Since the value of ε is the same for earthquake, probabilities are the same when calculated

We can also check the answer from the joint distribution table ..

$$(B = 1|A = 1) = \frac{P(B = 1, A = 1)}{P(A = 1)}$$

$$= \frac{\epsilon(1-\epsilon) + \epsilon^2}{0 + (1-\epsilon)\epsilon + \epsilon(1-\epsilon) + \epsilon^2}$$

$$= \frac{\epsilon(1-\epsilon) + \epsilon^2}{2 * \epsilon(1-\epsilon) + \epsilon^2}$$

$$= \frac{0.05 * (1-0.05) + 0.05 * 0.05}{2 * 0.05 * (1-0.05) + 0.05 * 0.05} = 0.5128$$

b	e	a	$\boldsymbol{p(b)}$	$\boldsymbol{p(e)}$	$\boldsymbol{p(a\|b,e)}$	$\mathbb{P}(B = b, E = e, A = a)$
0	0	0	$1-\varepsilon$	$1-\varepsilon$	1	$(1-\varepsilon)^2$
0	0	1	$1-\varepsilon$	$1-\varepsilon$	0	0
0	1	0	$1-\varepsilon$	ε	0	0
0	1	1	$1-\varepsilon$	ε	1	$(1-\varepsilon)\varepsilon$
1	0	0	ε	$1-\varepsilon$	0	0
1	0	1	ε	$1-\varepsilon$	1	$(1-\varepsilon)\varepsilon$
1	1	0	ε	ε	0	0
1	1	1	ε	ε	1	ε^2

Example (3)

- Given that the alarm rings, what is the probability of a burglary if you know an earthquake happened?

$P(b|a=1,e=1) = \alpha P(a,b,e) = \alpha P(b)P(e)P(a|b,e)$

- $\mathbb{P}(b=1,e=1,a=1) = 0.05 * 0.95 * 1 = 0.0475$
- $\mathbb{P}(b=0,e=1,a=1) = 0.95 * 0.95 * 1 = 0.9025$

Normalize

b	$\boldsymbol{p}(b)$
0	$\frac{0.0475}{0.9025+0.0475} = 0.05$
1	$\frac{0.9025}{0.9025+0.0475} = 0.95$

- When the alarm goes off, but we know an earthquake happened, the probability of a burglary does not change!
- This is Explaining Away that people do: if the alarm rings and we know there is an earthquake, we discount the possibility of a burglary being the cause

Example (4)

Given that the alarm rings, what is the probability of both a burglary and an earthquake simultaneously?

- $P(b, e|a) = \alpha P(a, b, e) = \alpha\, P(b)P(e)P(a|b, e)$

$\mathbb{P}(b = 1, e = 1, a = 1) = 0.05 * 0.05 * 1 = 0.0025$ } both = 0.0025

$\mathbb{P}(b = 1, e = 0, a = 1) = 0.05 * 0.95 * 1 = 0.0475$

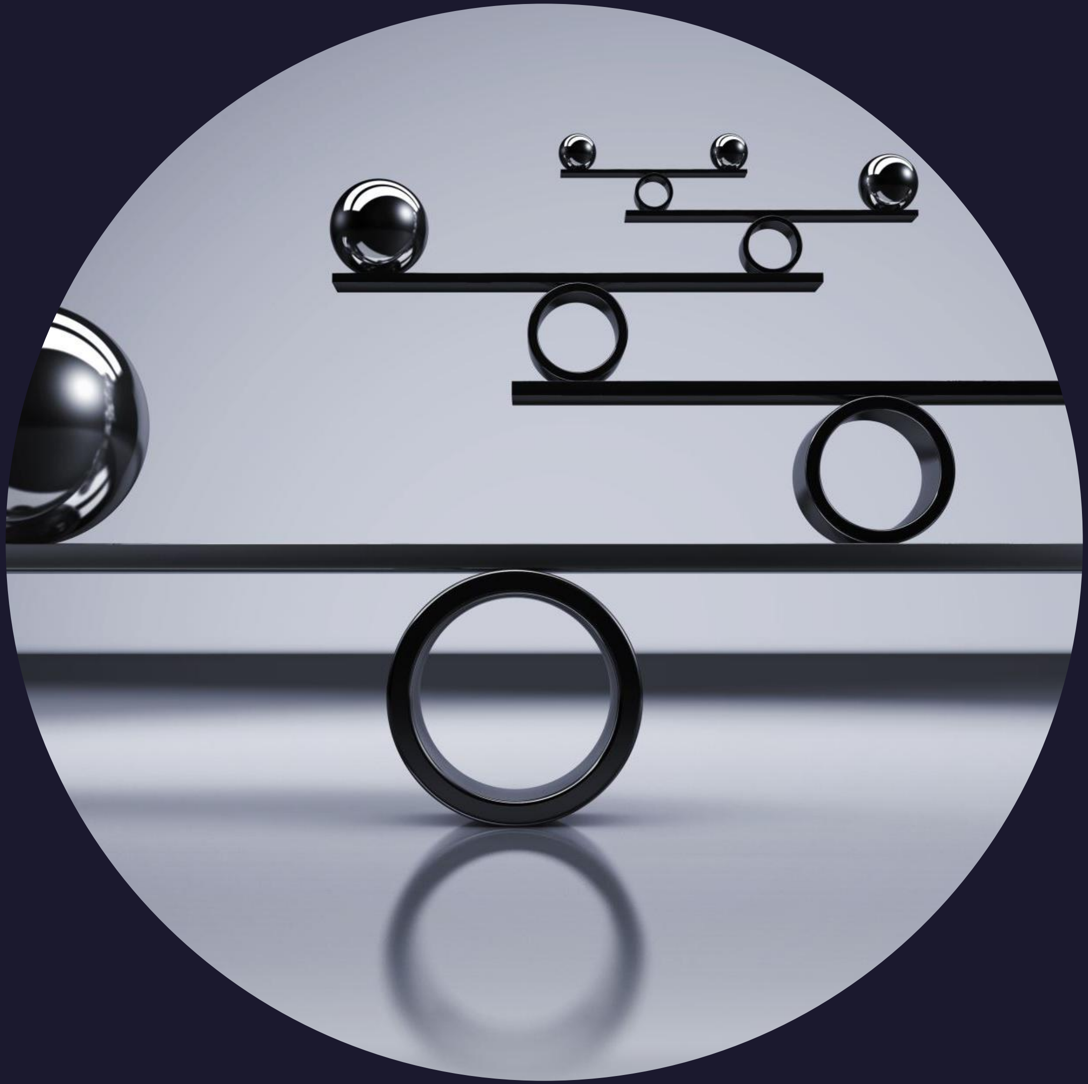
$\mathbb{P}(b = 0, e = 1, a = 1) = 0.95 * 0.05 * 1 = 0.0475$ } either one = 0.095

} Normalize

	$\boldsymbol{p}(b)$
both	$\frac{0.0025}{0.0975} = 0.0256$
either	$\frac{0.095}{0.0975} = 0.9744$

A Probabilistic Learning Algorithm

Naïve Bayes



Naïve Bayes

- Naïve Bayes is a very simple model which is often used for classification.
- Generative model
  - Generative models: how the input is generated from the output
  - Discriminative models: take the input and produce the output label
- Extremely easy and fast, just requires counting

Applying Bayes Rule $P(a|b) = \frac{P(b|a)P(a)}{P(b)}$

Diagnosis

Causation

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P\ (\text{effect})}$$

$$P(\text{disease}|\text{symptom}) = \frac{P(\text{symtom}|\text{disease})P(\text{disease})}{P\ (\text{symptom})}$$

- Example:
  - Meningitis causes the patient to have a stiff neck 70% of the time
  - The prior probability that a patient has meningitis is 1/50,000
  - The prior probability that any patient has a stiff neck is 1%

Patient has a stiff neck. What is the probability the patient has meningitis?

$$P(\text{meningitis}|\text{stiff neck}) = \frac{P(\text{stiff neck}|\text{meningitis})P(\text{meningitis})}{P\ (\text{stiff neck})} = \frac{0.7 * \left(\frac{1}{50000}\right)}{0.01}$$

Naïve Bayes Classifiers

- **Naïve Bayes assumptions**:

1. **Feature independence**: The features of the data are conditionally independent of each other, given the class label. $P(A,B) = P(A)P(B)$
2. **Continuous features are normally distributed**: If a feature is continuous, then it is assumed to be normally distributed within each class.
3. **Discrete features have multinomial distributions**: If a feature is discrete, then it is assumed to have a multinomial distribution within each class.
4. **Features are equally important**: All features are assumed to contribute equally to the prediction of the class label.
5. **No missing data**: The data should not contain any missing values.

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naïve Bayes Classifiers

- $P(y|x_1, \dots, x_n) = \frac{P(x_1|y) \times \dots \times P(x_n|y)P(y)}{P(x_1) \times \dots \times P(x_n)}$
- $P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1) \times \dots \times P(x_n)}$

1. $P(y) = \frac{9}{14}$

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naïve Bayes Classifiers

- $P(y|x_1,\dots,x_n) = \dfrac{P(x_1|y) \times \dots \times P(x_n|y)P(y)}{P(x_1) \times \dots \times P(x_n)}$
- $P(y|x_1,\dots,x_n) = \dfrac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1) \times \dots \times P(x_n)}$

1. $P(y) = \dfrac{9}{14}$
2. Calculate $P(x_i|y_i)$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	3	2	3/9	2/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Naïve Bayes Classifiers

Outlook

	Yes	No	P(yes)	P(no)
Sunny	3	2	3/9	2/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

- today = (Sunny, Hot, Normal, False)
- $P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1) \times \dots \times P(x_n)}$
- $P(\text{yes}|\text{S, H, N, F}) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$

$$= \frac{9}{14} \cdot \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \approx 0.02116$$

- $P(\text{no}|\text{S, H, N, F}) \propto \frac{9}{14} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \approx 0.0068$
- Normalize:

$$P(\text{yes}|\text{today}) = \frac{0.02116}{0.02116+0.0068} \approx 0.757$$

$$P(\text{no}|\text{today}) = \frac{0.0068}{0.02116+0.0068} \approx 0.243$$

Predict play golf