# CS 5100: Foundations of Artificial Intelligence

Probabilistic Inference

Prof. Amy Sliva

November 3, 2011

# Outline

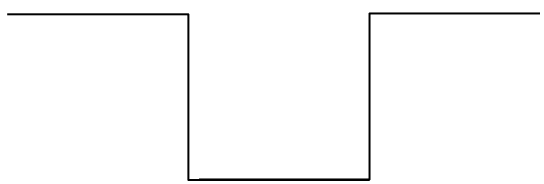
- Discuss Midterm
- Class presentations start next week!
- Reasoning under uncertainty
- Probability

# Group presentation guidelines

- Focus on the **main concepts**
  - Motivation, research challenges, techniques used, evaluation, outcomes, significance
  - Form you use for your reviews will help here!
- Identify the researcher/research group (e.g., Stanford AI Lab or Microsoft Research)
- Describe the project steps in broad outline—great detail not necessary!
  - Try to cover one element to describe in **moderate depth**
  - Be sure to explain "where's the AI" in this project
- Powerpoint—bring slides/laptop or email slides to me before class
  - Everyone in the group must have a turn to present!

# Presentation organization and timing

- Short technical talks (15 mins or less) are more difficult than longer talks
- Presenter(s) should rehearse to know if you can make your presentation in the allowed time
- Shape of a short talk—moderate depth of material in middle



- Recommended that each team meet in advance to rehearse and refine the presentation

# Presentation visuals (slides)

- Visuals should be *exhibits* that you talk about
  - Do not put lots of text on slides
  - Do not read your slides for your presentation
- Use interactivity, video, images to keep your audience awake ☺
- You may use figures from the paper, but cite your source on each slide

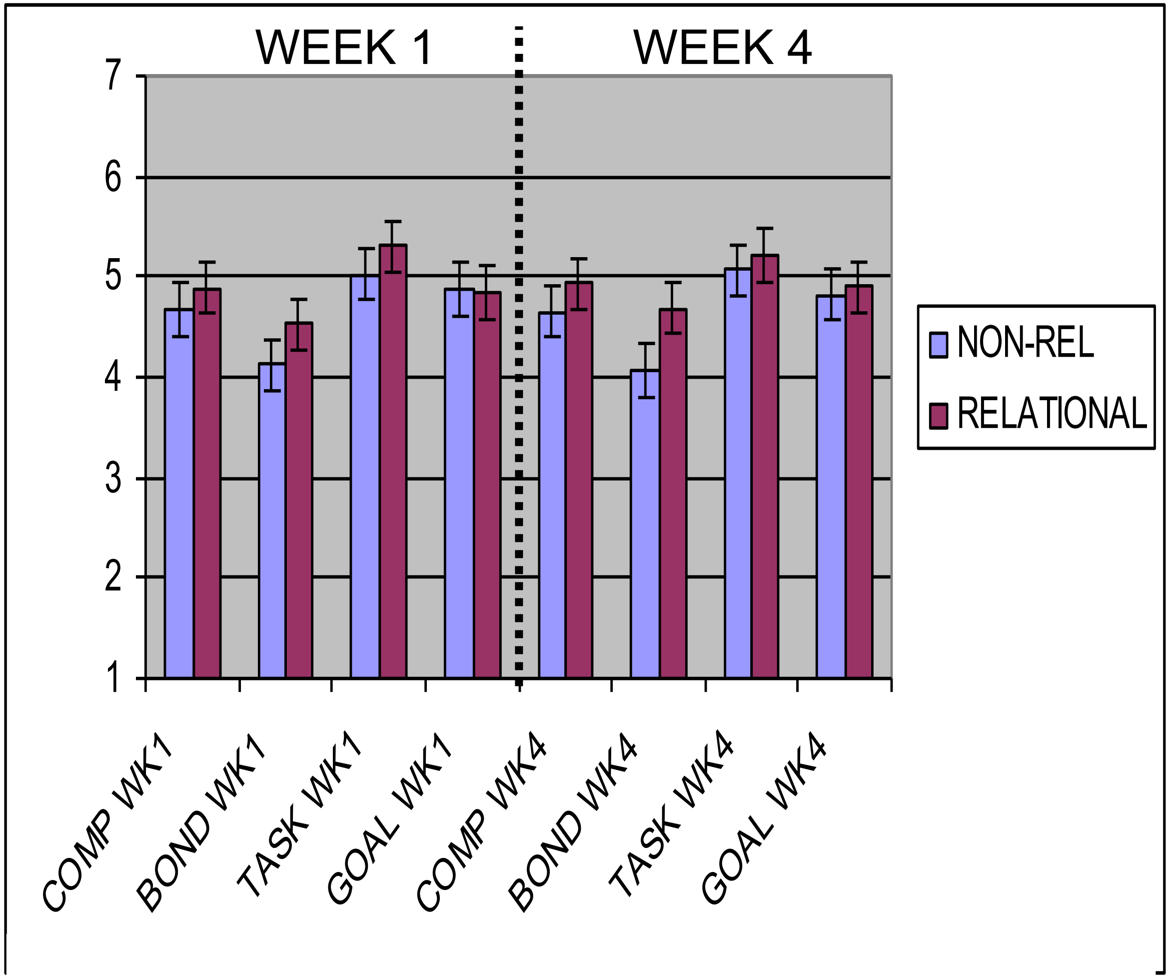
# Presentation visuals (cont.)

- Begin with title slide: title, author, institution, event, date
- 2nd slide should be an outline
- Don't use too many visuals—1 minute for a simple bullet list, 2-3 minutes for slides you need to explain or discuss

# Do not do this…

| Measure | Change From Day1 | To Day2 | ALL CONDS df | t | p | CONTROL df | t | p | NON-REL df | t | p | RELATIONL df | t | p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **WAI/COMP** | 7 | 27 | 54 | 0.205 | 0.838 | | | | 24 | 0.014 | 0.989 | 29 | 0.361 | 0.720 |
| **WAI/BOND** | 7 | 27 | 54 | 0.519 | 0.606 | | | | 24 | 0.376 | 0.710 | 29 | 1.489 | 0.147 |
| **WAI/TASK** | 7 | 27 | 54 | 0.134 | 0.894 | | | | 24 | 0.409 | 0.686 | 29 | 0.661 | 0.514 |
| **WAI/GOAL** | 7 | 27 | 54 | 0.155 | 0.877 | | | | 24 | 0.081 | 0.936 | 29 | 0.329 | 0.745 |
| **CONTINUE LAURA** | 30 | 44 | 54 | 0.868 | 0.389 | | | | 24 | 0.625 | 0.538 | 29 | 0.619 | 0.541 |
| **MIN/DAY** | -6-0 | 22-30 | 81 | 1.470 | 0.145 | 26 | 1.274 | 0.214 | 24 | 0.124 | 0.903 | 29 | 1.104 | 0.279 |
| | 1-7 | 22-30 | 81 | 0.691 | 0.492 | 26 | 0.758 | 0.456 | 24 | 0.109 | 0.914 | 29 | 0.358 | 0.723 |
| | 22-30 | 38-44 | 81 | 3.626 | **0.001** | 26 | 2.480 | **0.020** | 24 | 1.959 | 0.062 | 29 | 1.804 | 0.082 |
| **DAY/WK>30MIN** | -6-0 | 22-30 | 81 | 6.653 | **0.000** | 26 | 2.323 | **0.028** | 24 | 5.284 | **0.000** | 29 | 4.347 | **0.000** |
| | 1-7 | 22-30 | 81 | 6.272 | **0.000** | 26 | 2.401 | **0.024** | 24 | 3.818 | **0.001** | 29 | 4.597 | **0.000** |
| | 22-30 | 38-44 | 81 | 8.990 | **0.000** | 26 | 4.043 | **0.000** | 24 | 5.322 | **0.000** | 29 | 6.530 | **0.000** |
| **STEP/DAY** | 1-7 | 22-30 | 81 | 1.778 | 0.079 | 26 | 1.197 | 0.242 | 24 | 2.366 | **0.026** | 29 | 0.236 | 0.815 |
| **DAY/WK>10KSTEP** | 1-7 | 22-30 | 77 | 3.986 | **0.000** | 25 | 1.355 | 0.188 | 23 | 3.591 | **0.002** | 27 | 2.055 | **0.050** |
| **STAGE** | Intake | 30 | 81 | 6.988 | **0.000** | 26 | 3.403 | **0.002** | 24 | 4.000 | **0.001** | 29 | 4.738 | **0.000** |
| | 30 | 44 | 81 | 2.019 | **0.047** | 26 | 1.185 | 0.247 | 24 | 1.000 | 0.327 | 29 | 1.409 | 0.169 |
| **SELF-EFFICACY** | 1 | 29 | 81 | 4.782 | **0.000** | 26 | 0.872 | 0.391 | 24 | 3.314 | **0.003** | 29 | 4.750 | **0.000** |
| | 29 | 44 | 81 | 2.770 | **0.007** | 26 | 1.525 | 0.139 | 24 | 4.550 | **0.000** | 29 | 0.085 | 0.933 |
| **PROS** | 1 | 29 | 81 | 1.998 | **0.049** | 26 | 1.418 | 0.168 | 24 | 0.456 | 0.653 | 29 | 1.540 | 0.134 |
| | 29 | 44 | 81 | 0.393 | 0.695 | 26 | 1.147 | 0.262 | 24 | 0.225 | 0.824 | 29 | 0.308 | 0.760 |
| **CONS** | 1 | 29 | 81 | 0.902 | 0.370 | 26 | 1.124 | 0.271 | 24 | 0.499 | 0.622 | 29 | 0.823 | 0.417 |
| | 29 | 44 | 81 | 0.740 | 0.462 | 26 | 0.386 | 0.703 | 24 | 0.611 | 0.547 | 29 | 0.339 | 0.737 |
| **CONTINUE FT** | 30 | 44 | 81 | 1.520 | 0.133 | 26 | 1.442 | 0.161 | 24 | 1.163 | 0.256 | 29 | 0.000 | 1.000 |

# …but use figures whenever possible




Always include a title and brief caption

# Uncertain knowledge/beliefs arise from…

- **Chance (probability)**
- Incomplete knowledge
  - Defeasible (“default”) reasoning
  - Reasoning from ignorance
- Ambiguous or conflicting evidence
  - Evidential reasoning
- Vagueness
  - Fuzzy or open-textured concepts

# Examples of uncertain reasoning

- Chance
  - Will your first child be a boy or girl?
  - Will the coin come up heads or tails?

- Defeasible reasoning
  - If I believe Tweety is a bird, then I think he can fly
  - If I learn Tweety is a penguin, then I think he can’t fly
    - Assume the **normal case** unless or until you learn otherwise

- Reasoning from ignorance
  - I believe the President of the U.S. is alive

# More examples of uncertain reasoning

- Evidential reasoning—how **strongly** do I believe *P* based on evidence? (**confidence levels**)
  - Quantitative
    [0, 1], [-1, 1]
  - Qualitative
    {definite; very likely, likely, neutral, unlikely, very unlikely, definitely not}

- Fuzzy concepts
  - John in tall
  - My friend promises to return a book “soon”
    - Add “degree” to fuzzy assertions (between 0 and 1)

# Some problems with uncertain reasoning

- Using **probability**
  - Know $P(A)$ and $P(B)$—have methods for calculating $P(\neg A)$, $P(A \wedge B)$, $P(A \vee B)$, and $P(A \mid B)$
- Using evidential or fuzzy models
  - Problems with **consistency**—If “John is tall” (.8) and “John is smart” (.6), what can be said about “John is both tall and smart”?
- We will mainly focus on probabilistic reasoning models

# Syntax for probabilistic reasoning

- Basic element: **random variable**
  - Similar to propositional logic—possible worlds (**sample space**) defined by assignment of values to random variables

- **Boolean** random variables
  - E.g., *Cavity* (do I have a cavity?)
- **Discrete** random variables
  - E.g., *Weather* is one of <*sunny, rainy, cloudy, snow*>
- **Elementary proposition** constructed by assignment of a value to a single random variable
  - E.g., *Weather* = *sunny*, *Cavity* = *false* (abbreviated ¬*cavity*)
- **Complex propositions** formed from elementary propositions and standard logical connectives
  - E.g., *Weather* = *sunny* ∨ *Cavity* = *false*

# Syntax for probability (cont.)

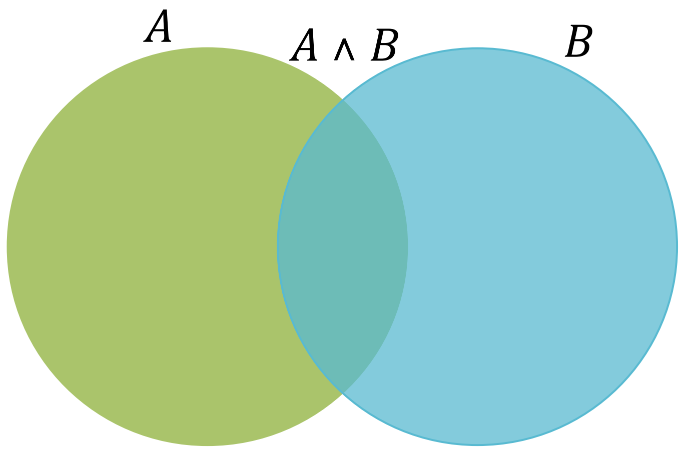
- **Atomic event**—A **complete** specification of the state of the world about which the agent is uncertain
  - E.g., If the world consists of two boolean variables *Cavity* and *Toothache*, then there are four distinct atomic events:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

- Atomic events are **mutually exclusive** and **exhaustive** (often called "outcomes")
- Events in general are **sets of atomic events**, such as Cavity = true

# Axioms of probability

- For any propositions $A$, $B$
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ **(Inclusion-exclusion principle)**



# Multivalued Random Variables

- Suppose $A$ can take on more than 2 values
- $A$ is a **random variable with arity *k*** if it can take on exactly one value out of the domain $\{v_1, v_2, \ldots, v_k\}$
- Then…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$$

# Easy fact about multivalued random variables

- Using the axioms of probability…

  $0 \le P(A) \le 1$, $P(true) = 1$, $P(false) = 0$

  $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

- And Assuming that $A$ obeys…

  $P(A = v_i \wedge A = v_j) = 0$ if $i \ne j$

  $P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$

- It is easy to prove that

  $$\mathbf{P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)}$$

- And thus we can prove

  $$\mathbf{\sum_{j=1}^{k} P(A = v_j) = 1}$$

# Prior probability

- **Prior** or **unconditional** probabilities of propositions—corresponds to belief prior to arrival of any (new) evidence
  - E.g., P(*Cavity* = *true*) = 0.1 and P(*Weather* = *sunny*) = 0.72
- **Probability distribution** gives values for all possible assignments
  - E.g., **P**(*Weather*) = <0.72, 0.1, 0.08, 0.1> (sums to 1)
- **Joint probability distribution** for a set of random variables gives probability of every atomic event on those random variables
  - E.g., **P**(*Weather*, *Cavity*) = a 4 ×2 matrix of values:

| *Weather =* | *sunny* | *rainy* | *cloudy* | *snow* |
|---|---|---|---|---|
| *Cavity = true* | 0.144 | 0.02 | 0.016 | 0.02 |
| *Cavity = false* | 0.576 | 0.08 | 0.064 | 0.08 |

- **Every probability question about a domain can be answered by the joint distribution**

# Conditional probability

- **Conditional** or **posterior** probabilities—based on known information
  - Eg., P(*cavity* | *toothache*) = 0.8
    Given that *toothache* is all I know

- **Notation** for conditional distributions (use boldface):
  - **P**(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors
- If we know more, e.g., *cavity* is also given, then we have P(*cavity* | *toothache, cavity*) = 1
- New evidence by be **irrelevant**, allowing simplification
  - E.g., cavity does not depend on weather:
    P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8

- **This kind of inference, sanctioned by domain knowledge, is crucial for probabilistic reasoning in AI!**

# More on conditional probability

- Definition of **conditional probability**:
  $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$ if $P(b) > 0$

- **Product rule** gives an alternative formulation:
  $P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$
- A general version holds for whole distributions
  - E.g., **P**(Weather, Cavity) = **P**(Weather | Cavity)**P**(Cavity)
    (View as a set of 4 × 2 equations, **NOT** matrix multiplication)

- **Chain rule** is derived by successive application of product rule

$$\begin{aligned} P(X_1,\ldots,X_n) &= P(X_1,\ldots,X_{n-1})P(X_n \mid X_1,\ldots,X_{n-1}) \\ &= P(X_1,\ldots,X_{n-2})P(X_{n-1} \mid X_1,\ldots,X_{n-2})\, P(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \ldots P(X_n \mid X_1,\ldots,X_{n-1}) \end{aligned}$$

**OR $\Pi_{i = 1 \text{ to } n}$ P(Xi | X1,…,Xi-1)**

# Inference by enumeration

- Start with the joint probability distribution

| | *toothache* | | *¬toothache* | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

- For any proposition ϕ, **sum the atomic events** where it is true:
  $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

# Inference by enumeration

- Start with the joint probability distribution

| | *toothache* | | *¬toothache* | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

- For any proposition ϕ, **sum the atomic events** where it is true:
  $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

  $P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

# Inference by enumeration

- Start with the joint probability distribution

| | *toothache* | | *¬toothache* | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.2$$

# In class exercise

- Given the joint distribution below and the definition $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$ :
  - What is P(Cavity = true)?
  - What is P(Weather = sunny)?
  - What is P(Cavity = True | Weather = sunny)?

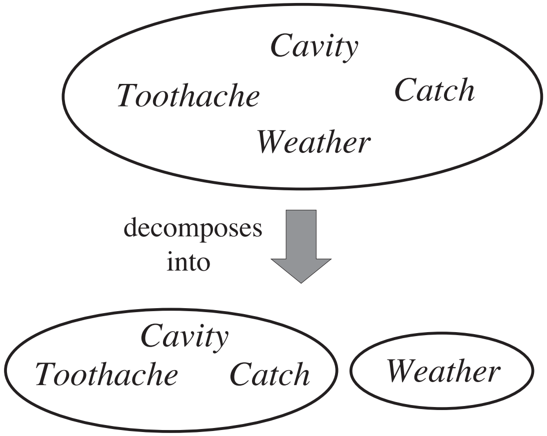
| *Weather =* | *sunny* | *rainy* | *cloudy* | *snow* |
|---|---|---|---|---|
| *Cavity = true* | 0.144 | 0.02 | 0.016 | 0.02 |
| *Cavity = false* | 0.576 | 0.08 | 0.064 | 0.08 |

- Given the meta-equation $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\mathbf{P}(Cavity)$
  - What are the 8 equations represented here?

# Independence

- $A$ and $B$ are independent iff
  $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$
  - E.g., $\mathbf{P}(\mathit{Toothache, Catch, Cavity, Weather})=$
    $\mathbf{P}(\mathit{Toothache, Catch, Cavity})\mathbf{P}(\mathit{Weather})$

    32 entries reduced to 12!




- **Absolute independence** powerful, but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do??

# Bayes Rule

- Product rule $P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$

  $\Rightarrow$ **Bayes rule:** $P(a \mid b) = \dfrac{P(b \mid a)P(a)}{P(b)}$

- Or in distribution form:
  $\mathbf{P}(Y \mid X) = \dfrac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha \mathbf{P}(X \mid Y)\, \mathbf{P}(Y)$

- Useful for assessing **diagnostic** probability from **causal** probability:
  - $P(\text{Cause} \mid \text{Effect}) = \dfrac{P(\text{Effect} \mid \text{Cause})\, P(\text{Cause})}{P(\text{Effect})}$
  - E.g., Let $M$ be meningitis, $S$ be stiff neck
    $P(M \mid S) = P(S \mid M)\, P(M) \,/\, P(S) = 0.8 \times 0.0001 \,/\, 0.1 = 0.0008$
  - Note: posterior probability of meningitis is still very small!

# Example: Expert system for medical diagnosis

- 100 diseases
- 20 symptoms

- # of parameters needed to calculate P($D$) when a patient provides his/her symptoms

- Strategy to reduce size—assume independence of symptoms

- Recalculate number of parameters needed

# Example: Expert system for medical diagnosis

- 100 diseases
- 20 symptoms

- # of parameters needed to calculate P($D$) when a patient provides his/her symptoms
  - **$2^{20}$ combinations of symptoms for each value (true or false) of each disease!**

- Strategy to reduce size—assume independence of symptoms

- Recalculate number of parameters needed
  - **20 single-variable distributions**

# More general forms of Bayes rule

- $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(B \mid A)\mathrm{P}(A)}{\mathrm{P}(B \mid A)\mathrm{P}(A) + \mathrm{P}(B \mid \neg A)\mathrm{P}(\neg A)}$

- $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(B \mid A, e)\mathrm{P}(A \mid e)}{\mathrm{P}(B \mid e)}$

- $\mathrm{P}(A = v_i \mid B) = \dfrac{\mathrm{P}(B \mid A = v_i)\mathrm{P}(A = v_i)}{\sum_{k=1}^{n} \mathrm{P}(B \mid A = v_k)\mathrm{P}(A = v_k)}$

# Conditional independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
  - If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
    - **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)
  - The same independence holds if I haven't got a cavity
    - **P**(*catch* | *toothache, ¬cavity*) = **P**(*catch* | *¬cavity*)
- Catch is **conditionally independent** of Toothache given Cavity: **P**(*Catch* | *Toothache, Cavity*) = **P**(*Catch* | *Cavity*)
  - Equivalent statements (from original definitions of independence)
    - **P**(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)
    - **P(*Toothache, Catch* | *Cavity*) = P(*Toothache* | *Cavity*)P(*Catch* | *Cavity*)**

# Conditional independence (cont.)

- Write out full joint distribution using chain rule:

  $\mathbf{P}(Toothache, Catch, Cavity)$
  $= \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch, Cavity)$
  $= \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)$
  $= \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)$

  i.e., 2 + 2+ 1 = 5 independent numbers

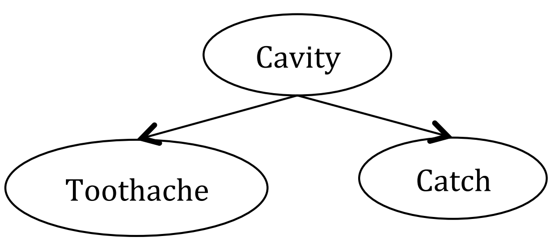
- In most cases, conditional independence reduces size of representation from **exponential** in *n* to **linear** in *n*
- Conditional independence is most basic and robust form of knowledge about uncertain environments

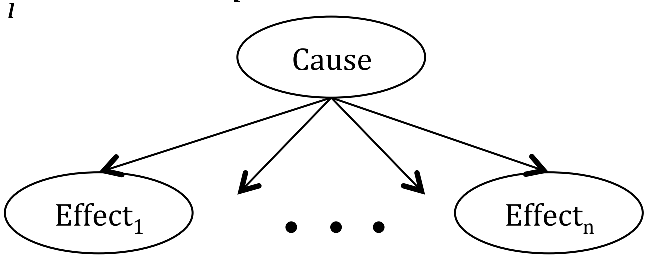
# Bayes rule and conditional independence

- $\mathbf{P}(Cavity \mid toothache \, catch)$
  $= \mathbf{P}(toothache \, catch \mid Cavity)\mathbf{P}(Cavity)$
  $= \mathbf{P}(toothache \mid Cavity)\mathbf{P}(catch \mid Cavity)\mathbf{P}(Cavity)$

- We say: "toothache and catch are independent, given cavity"
  - Cavity **separates** Toothache and Catch because it is a direct cause of both
  - Example of a **naïve Bayes** model

- $P(Cause, Effect_1, \ldots, Effect_n) = P(Cause)\prod_i P(Effect_i \mid Cause)$




- Total number of parameters is **linear** in n (number of systems)
- This is our first Bayesian inference net!

# Example of conditional independence

- *A*: Amy teaches the class
- *S*: It is sunny
- *L*: The lecturer arrives *slightly* late
- Assume lecturers are sometimes delayed by bad weather

- Start by writing down knowledge we're happy about:
  - $\mathrm{P}(S \mid A) = \mathrm{P}(S)$, $\mathrm{P}(S) = 0.3$, $P(A) = 0.6$
  - **Lateness is not independent of the weather and is not independent of the lecturer**
  - **Weather and lecturer are independent**

# Example of conditional independence

- $A$: Amy teaches the class
    - $S$: It is sunny
    - $L$: The lecturer arrives *slightly* late
- Assume lecturers are sometimes delayed by bad weather

- Start by writing down knowledge we're happy about:
    - $P(S \mid A) = P(S)$, $P(S) = 0.3$, $P(A) = 0.6$
    - **Lateness is not independent of the weather and is not independent of the lecturer**
    - **Weather and lecturer are independent**

- Know the joint probability of $S$ and $A$, so now need:
    - $\mathbf{P}(L \mid S, A)$ for the 4 cases where $S$ and $A$ are true/false

# Example of conditional independence

- *A*: Amy teaches the class
- *S*: It is sunny
- *L*: The lecturer arrives *slightly* late

- Assume lecturers are sometimes delayed by bad weather

$P(S \mid A) = P(S)$
$P(S) = 0.3$
$P(A) = 0.6$

$P(L \mid A \land S) = 0.05$
$P(L \mid A \land \neg S) = 0.1$
$P(L \mid \neg A \land S) = 0.1$
$P(L \mid \neg A \land \neg S) = 0.2$

- Now we can **derive** a full joint distribution with a “mere” six numbers instead of seven
  - NOTE: Savings are larger for larger numbers of variables