인공지능

22년 삼성 AI 전문가과정
6월 8일 수요일 2교시
장병탁

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang




Lecture Overview

인공지능

10차시 : Uncertainty, Probability, Information

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang

Introduction: Acting under Uncertainty

- Agents need to handle **uncertainty**, due to partial observability, non-determinism or adversaries.
- **In previous lectures,** the **logical agents** handled uncertainty by
  - Belief state—a representation of the set of all possible world states
  - Contingency plan that handles every possible eventuality
- Drawbacks of this **deterministic approach**
  - Every logically possible explanation lead to large belief-state representations
  - A correct contingent plan can grow arbitrarily large
  - There might be no plan to guarantee to achieve a goal
- **In this lecture,** we study methods for **handling uncertainty with degrees of belief.** This requires quantifying uncertainty and using probability theory and probabilistic reasoning methods.
  - **Decision-theoretic agents**
  - **Utility-based agents**

Uncertainty and Probability

Uncertainty

- Let action A_t = leave for airport t minutes before flight.
  Will A_t get me there on time?

$S = (A_{25}$ will get me there on time)

- Problems:
  - Partial observability (road state, other drivers' plans, etc.)
  - Noisy sensors (TBS traffic reports)
  - Uncertainty in action outcomes (flat tire, etc.)

Probability

$S = (A_{25}$ will get me there on time)

- Given the available evidence,
  A_{25} will get me there on time with probability 0.04

$P(S = true) = 0.04$

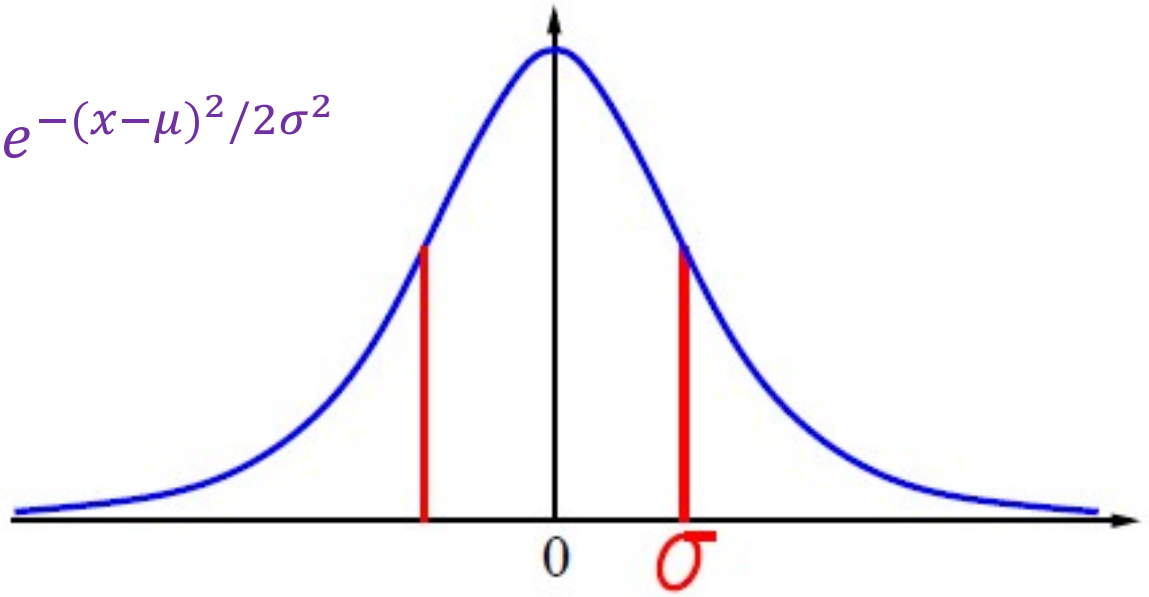
- Probabilistic assertions **summarize** effects of
  - Laziness: failure to enumerate exceptions, qualifications, etc.
  - Ignorance: lack of relevant facts, initial conditions, etc.

Probability Distribution

Probability for continuous variables

- Gaussian density

$$P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



Bayesian Probability

Subjective or Bayesian probability

- Probabilities relate propositions to one's own state of knowledge

  $P(A_{25}|\textit{no reported accidents}) = 0.06$

- Probabilities of propositions change with new evidence:

  $P(A_{25}|\textit{no reported accidents}, 5\ a.m.) = 0.15$

➢ Bayes' rule

- $P(a|b) = \frac{P(b|a)P(a)}{P(b)} = \alpha P(b|a)P(a)$

➢ Useful for assessing diagnostic probability from causal probability

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

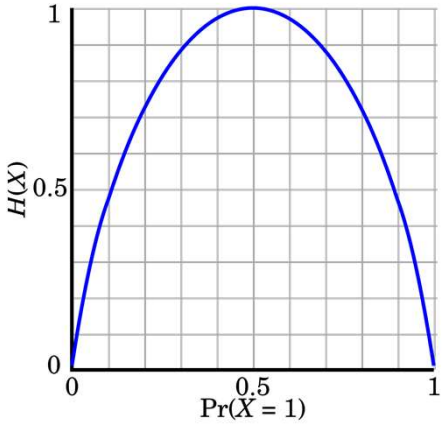
Information Theory

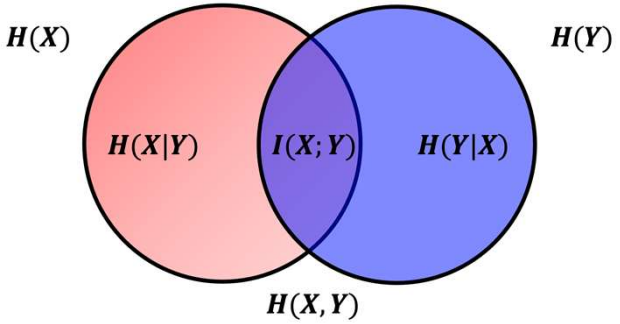
- Information: $I(x) = \log_2 \frac{1}{P_X(x)}$
- Entropy: $H(X) = \mathbb{E}_{x \in \mathcal{X}}[I(x)] = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)$
- Mutual information:

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log_2 \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}$$

- Relative entropy (KL divergence):

$$D_{\mathrm{KL}}(P_X || P_Y) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{P_Y(x)}$$





Outline (Lecture 10)

10.1 Acting under Uncertainty




10.1 Acting under Uncertainty (1/5)

1) Uncertainty

- Let action A_t = leave for airport t minutes before flight.
  Will A_t get me there on time?
- Problems:
  - Partial observability (road state, other drivers' plans, etc.)
  - Noisy sensors (TBS traffic reports)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Immense complexity of modelling and predicting traffic

10.1 Acting under Uncertainty (2/5)

2) Methods for handling uncertainty

- Default or nonmonotonic logic
  - Assume A_{25} works unless contradicted by evidence
- Rules with fudge factors
  - $A_{25} \mapsto_{0.3} AtAirportOnTime$
  - $Sprinkler \mapsto_{0.99} WetGrass$
  - $WetGrass \mapsto_{0.7} Rain$
- Probability

  $S = (A_{25}$ will get me there on time)
  - Given the available evidence,
    A_{25} will get me there on time with probability 0.04 $P(S = true) = 0.04$

10.1 Acting under Uncertainty (3/5)

3) Summarizing uncertainty

- Example: medical diagnosis

  $Toothache \Longrightarrow Cavity$

  $Toothache \Longrightarrow Cavity \vee GumProblem \vee Abscess \ldots$

  $Cavity \Longrightarrow Toothache$

- Logic fails to deal with this for three main reasons:
  - Laziness: hard to list a complete set of rules
  - Theoretical ignorance: no complete theory
  - Practical ignorance: not all necessary tests
- Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance, thereby solving the qualification problems.

10.1 Acting under Uncertainty (4/5)

4) Uncertainty and rational decisions

- **Utility theory**
  - An agent must first have preferences between the different possible outcomes of the various plans. We use utility theory to represent and reason with preferences.
  - utility being "the quality of being useful"
- **Decision theory**
  - Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called decision theory:

    Decision theory = probability theory + utility theory
- **Maximum expected utility (MEU)**
  - The fundamental idea of decision theory is that *an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action.* This is the principle of **maximum expected utility (MEU).**

10.1 Acting under Uncertainty (5/5)

A decision-theoretic agent that selects rational actions

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *belief_state*, probabilistic beliefs about the current state of the world
    action, the agent's action

  update *belief_state* based on *action* and *percept*
  calculate outcome probabilities for actions,
    given action descriptions and current *belief_state*
  select *action* with highest expected utility
    given probabilities of outcomes and utility information
  return *action*

Figure 12.1 A decision-theoretic agent that selects rational actions.

10.2 Basic Probability Notation

10.2 Basic Probability Notation (1/10)

1) Probability

- Probabilistic assertions **summarize** effects of
  - Laziness: failure to enumerate exceptions, qualifications, etc.
  - Ignorance: lack of relevant facts, initial conditions, etc.
- Subjective or Bayesian probability:
  - Probabilities relate propositions to one's own state of knowledge

$$P(A_{25}|\textit{no reported accidents}) = 0.06$$

  - Probabilities of propositions change with new evidence:

$$P(A_{25}|\textit{no reported accidents}, 5\ a.m.) = 0.15$$

10.2 Basic Probability Notation (2/10)

2) Probability basics

- Begin with a set Ω – the sample space
  - e.g., 6 possible rolls of a die.
  - $\omega \in \Omega$ is a sample point / possible world / atomic event
- A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ $s.t.$

$$0 \leq P(\omega) \leq 1, \quad \sum_{\omega} P(\omega) = 1$$

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

- An event A is any subset of Ω

10.2 Basic Probability Notation (3/10)

3) Random variables

- A random variable is a function from sample points to some range, e.g., the reals or Booleans
  - e.g., $Odd(1) = true$.
- P induces a probability distribution for any r.v. X:

$$P(X = x_i) = \sum_{\{\omega:X(\omega)=x_i\}} P(\omega)$$

  - e.g., $P(Odd = true) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

10.2 Basic Probability Notation (4/10)

4) Propositions

- Given Boolean random variables A and B:
  - event a = set of sample points where $A(\omega) = true$
  - event $\neg a$ = set of sample points where $A(\omega) = false$
  - event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$
- With Boolean variables, sample point = propositional logic model
  - e.g., $A(\omega) = true, B = false$, or $a \wedge \neg b$.
- **Proposition** = disjunction of atomic events in which it is true
  - e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$

    $\Longrightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

10.2 Basic Probability Notation (5/10)

5) Syntax for propositions

- Propositional or Boolean random variables
  - e.g., *Cavity* (do I have a cavity?)
  - *Cavity* = *true* is a proposition, also written *cavity*
- Discrete random variables (finite or infinite)
  - e.g., *Weather* is one of ⟨*sunny, rain, cloudy, snow*⟩
  - *Weather=rain* is a proposition
  - Values must be exhaustive and mutually exclusive
- Continuous random variables (bounded or unbounded)
  - e.g., $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.

10.2 Basic Probability Notation (6/10)

6) Prior probability

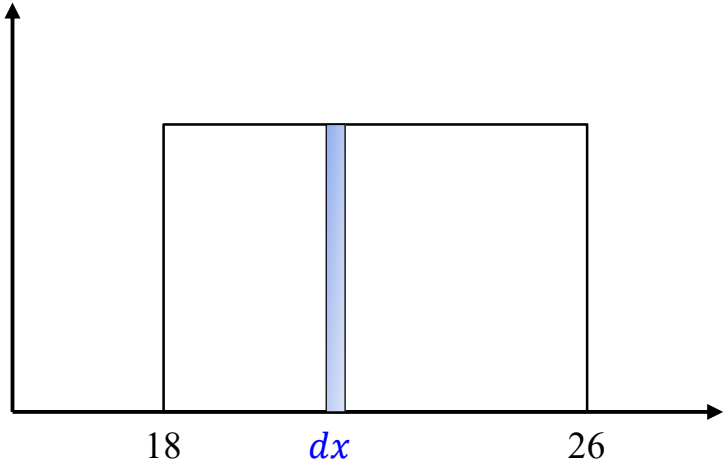
- Prior or unconditional probabilities of propositions
  - e.g., $P(Cavity = true) = 0.1$ and $P(Weather{=}sunny) = 0.72$
- Probability distribution gives values for all possible assignments:
  - e.g., $P(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)
- Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

$Weather =$	*sunny*	*rain*	*cloudy*	*snow*
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

10.2 Basic Probability Notation (7/10)

7) Probability for continuous variables

- Express distribution as a parameterized function of value:
  - $P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26



Here P is a *density*; integrates to 1.

$P(X = 20.5) = 0.125$ really means

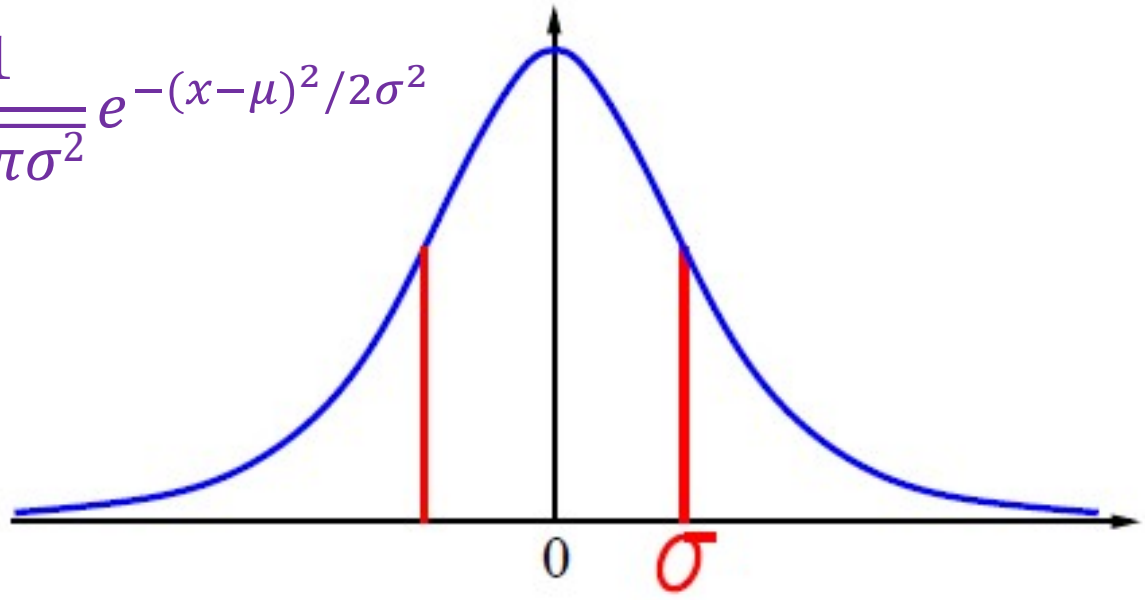
$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

10.2 Basic Probability Notation (8/10)

8) Probability for continuous variables, contd.

- Gaussian density

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



10.2 Basic Probability Notation (9/10)

9) Conditional probability

- Conditional or posterior probabilities
  - e.g., $P(cavity|toothache) = 0.8$
  - i.e., **given that** $toothache$ is all I know
  - **NOT** "if *toothache* then 80% chance of *cavity*"
  - If we know more, e.g., *cavity* is also given, then we have

$$P(cavity\ |toothcache,\ cavity) = 1$$

10.2 Basic Probability Notation (10/10)

10) Conditional probability, contd.

- Definition
  - $P(a|b) = \frac{P(a \wedge b)}{P(b)}$ if $P(b) \neq 0$
- Product rule
  - $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$
- Chain rule
  - $P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1})P(X_n|X_1, \ldots, X_{n-1})$

    $= P(X_1, \ldots, X_{n-2})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_n|X_1, \ldots, X_{n-1})$

    $= \cdots$

    $= \Pi_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$

10.3 Inference Using Full Joint Distributions

10.3 Inference Using Full Joint Distributions (1/3)

1) Inference by enumeration

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

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

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

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

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

10.3 Inference Using Full Joint Distributions (2/3)

2) Normalization

- Denominator can be viewed as a normalization constant α

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

$$P(cavity|toothache) = \alpha P(Cavity, toochache)$$
$$= \alpha[P(Cavity, toochache, catch) + P(Cavity, toothache, \neg catch)]$$
$$= \alpha[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\langle 0.12, 0.08\rangle + \langle 0.6, 0.4\rangle$$

10.3 Inference Using Full Joint Distributions (3/3)

3) Inference by enumeration, contd.

- Let **X** be all the variables. Typically, we want
  - the posterior joint distribution of the query variables **Y**
  - given specific values e for the evidence variables **E**
- Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
- Then the required summation of joint entries is done by summing out the hidden variables:
  - $\mathbf{P}(\mathbf{Y}|\,\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y}, \mathbf{e}) = \alpha \sum_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{e}, \mathbf{h})$
- The terms in the summation are joint entries because **Y**, **E**, and **H** together exhaust the set of random variables

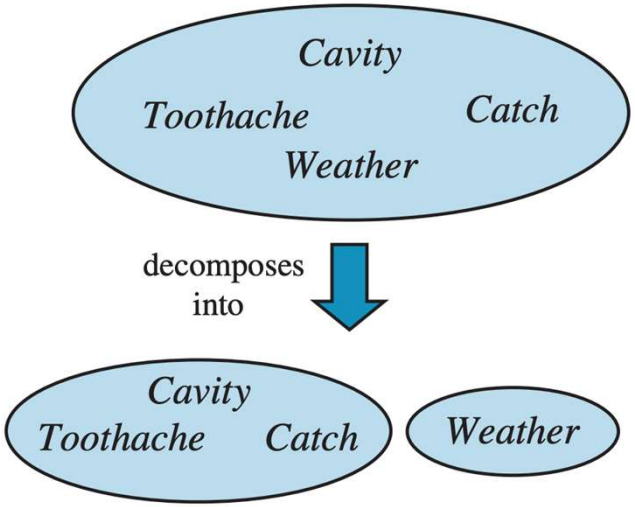
10.4 Independence and Bayes' Rule




10.4 Independence and Bayes' Rule (1/5)

1) Independence

- A and B are **independent** if
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)P(B)$



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

10.4 Independence and Bayes' Rule (2/5)

2) Conditional independence

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

10.4 Independence and Bayes' Rule (3/5)

3) Conditional independence, contd.

- Write out full joint distribution using chain rule:
  - $\mathbf{P}(Toothache, Catch, Cavity)$
    $= \mathbf{P}(Toothache | Catch, Cavity)\mathbf{P}(Catch, Cavity)$
    $= \mathbf{P}(Toothache | Catch, Cavity)\mathbf{P}(Catch | Cavity)\mathbf{P}(Cavity)$
    $= \mathbf{P}(Toothache | Cavity)\mathbf{P}(Catch | Cavity)\mathbf{P}(Cavity)$
  - i.e., 2 + 2 + 1 = 5 independent numbers (equations 1 and 2 remove 2)
  - In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.
  - **Conditional independence is our most basic and robust form of knowledge about uncertain environments.**

10.4 Independence and Bayes' Rule (4/5)

4) Bayes' rule

- Product rule
  - $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$
- Bayes' rule
  - $P(a|b) = \frac{P(b|a)P(a)}{P(b)}$ or in distribution form $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y)$
- Useful for assessing diagnostic probability from causal probability
  - $P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$

10.4 Independence and Bayes' Rule (5/5)

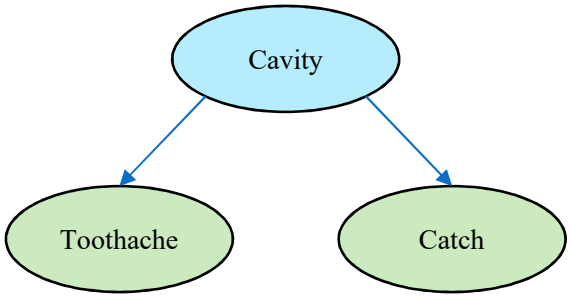
5) Bayes' rule and conditional independence

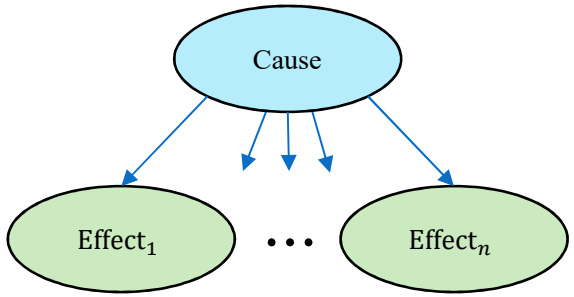
- $\mathbf{P}(Cavity|toothache \wedge catch)$

$$= \alpha\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$$

$$= \alpha\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$$

- An example of a **naive Bayes model**:
  - $\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\Pi_i\mathbf{P}(Effect_i|Cause)$





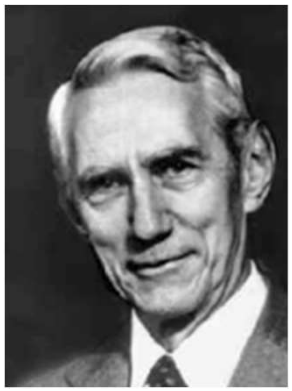
10.5 Information Theory

10.5 Information Theory (1/6)

- What is information? How to quantify the information?

Small information content

Large information content



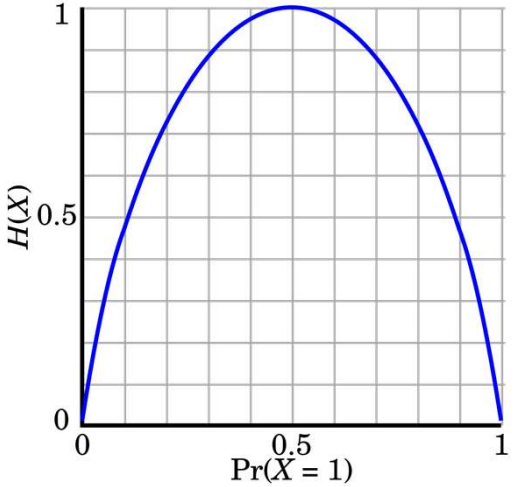
Claude Elwood Shannon
(1916-2001)

- Shannon's information theory deals with limits on data compression (source coding) and reliable data transmission (channel coding)
  - How much can data can be compressed?
  - How fast can data be reliably transmitted over a noisy channel?

<출처> ECE587, Information Theory, Duke University

10.5 Information Theory (2/6)

- Notation
  - Random variable: X
  - Sample value of a random variable: x
  - Set of possible sample values x of X: $\mathcal{X}$
  - Probability mass function (PMF) of discrete X: $P_X(x)$
  - Probability density function (PDF) of continuous X: $p_X(x)$



- **Information and Entropy**
  - Information (Uncertainty): $I(x) = \log_2 \frac{1}{P_X(x)}$
  - Entropy: $H(X) = \mathbb{E}_{x \in \mathcal{X}}[I(x)] = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)$ bits
    - A measure of the average uncertainty associated with a random variable

      → always non-negative!

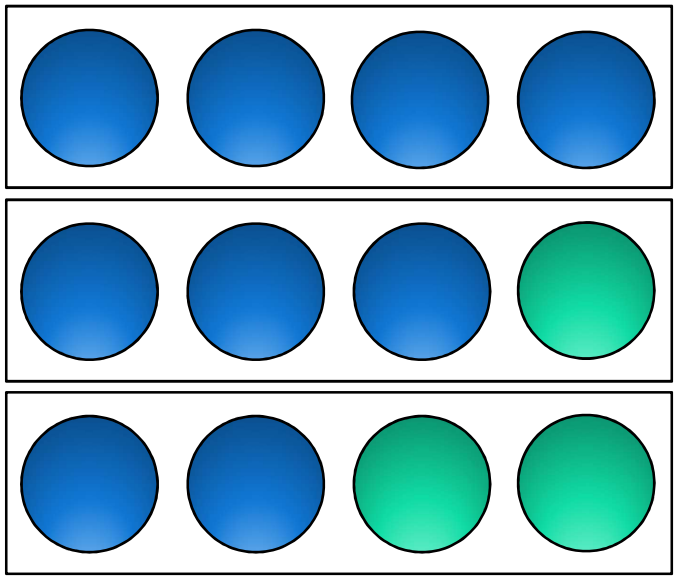
Entropy is a lower bound on the number of bits need to represent a random variable

10.5 Information Theory (3/6)

➢ A simple example of entropy

High entropy means high disorder and low energy! $H(X) = \mathbb{E}_{x \in \mathcal{X}}[I(x)] = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)$

When a random variable has a uniform distribution, the system has the highest entropy!



P(blue): 1, P(green): 0
Entropy: $H(X) = (-\log_2 1) \times 4 = 0$

P(blue): 0.75, P(green): 0.25
Entropy: $H(X) = (-\log_2 0.75) \times 3 - \log_2 0.25 = 3.245$

P(blue): 0.5, P(green): 0.5
Entropy: $H(X) = (-\log_2 0.5) \times 4 = 4$

10.5 Information Theory (4/6)

➢ Joint entropy

- The entropy of two discrete random variables X, Y with joint PMF $P_{X,Y}(x, y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_{X,Y}(x,y)$$

- Property: $H(X,Y) \leq H(X) + H(Y)$

➢ Conditional entropy

- Entropy of a random variable given another random variable
- There are various ways of writing this equation

$$H(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_{Y|X}(y|x)$$

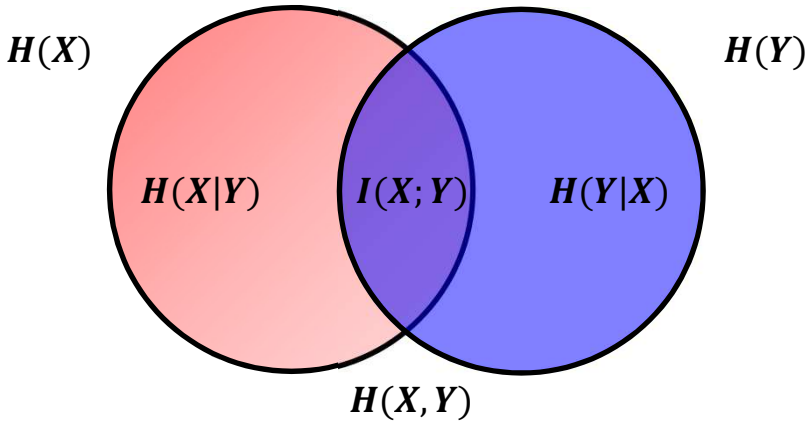
10.5 Information Theory (5/6)

➢ Mutual Information

- Measure of the amount of information that one random variable contains about another random variable

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log_2 \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}$$

- Useful expression: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$



10.5 Information Theory (6/6)

➢ Cross Entropy

- Measure of the distribution P_Y relative to a distribution P_X over a given set

$$H(P_X, P_Y) = \mathbb{E}_{x \in \mathcal{X}}[\log_2 P_Y(x)] = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_Y(x)$$

➢ Relative Entropy (Kullback-Leibler divergence)

- Measure of divergence between two distributions $P_X(x)$ and $P_Y(y)$

$$D_{\mathrm{KL}}(P_X||P_Y) = \mathbb{E}_{x \in \mathcal{X}}\left[\log_2 \frac{P_X(x)}{P_Y(x)}\right] = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{P_Y(x)}$$

- Property 1. Non-negativity: $D_{\mathrm{KL}}(P_X||P_Y) \geq 0$ (if $P_X = P_Y$, $D_{\mathrm{KL}}(P_X||P_Y) = 0$)
- Property 2. Non-symmetricity: $D_{\mathrm{KL}}(P_X||P_Y) \neq D_{\mathrm{KL}}(P_Y||P_X)$

Additional Materials




Probability Distribution & Statistical Mechanics

p_i : probability of occurrence of state i of a stochastic system

$p_i \geq 0$ (for all i) and $\sum_i p_i = 1$

E_i : energy of the system when it is in state i

In thermal equilibrium, the probability of state i is

(Canonical distribution / Gibbs distribution)

$$p_i = \frac{1}{Z}\exp\left(-\frac{E_i}{k_B T}\right)$$

$$Z = \sum_i \exp\left(-\frac{E_i}{k_B T}\right)$$

$\exp\left(-E/k_B T\right)$: Boltzmann factor

Z : sum over states (partition function)

We set $k_B = 1$ and view $-\log p_i$ as "energy"

1. States of low energy have a higher probability of occurrence than the states of high energy.
2. As the temperature T is reduced, the probability is concentrated on a smaller subset of low-energy states.

Markov Chain Monte Carlo (MCMC), Metropolis Algorithm

Metropolis Algorithm

A stochastic algorithm for simulating the evolution of a physical system to thermal equilibrium. A modified Monte Carlo method.

Markov Chain Monte Carlo (MCMC) method

Algorithm Metropolis

1. $X_n = x_i$. Randomly generate a new state x_j.

2. $\Delta E = E(x_j) - E(x_i)$

3. If $\Delta E < 0$, then $X_{n+1} = x_j$

   else if $\Delta E \geq 0$, then

   { Select a random number $\xi \sim$ U[0,1].

   If $\xi < \exp(-\Delta E / T)$, then $X_{n+1} = x_j$, (*accept*)

   else $X_{n+1} = x_i$. (*reject*)

   }

Markov Chain Monte Carlo (MCMC), Gibbs Sampling

Gibbs sampling

An iterative adaptive scheme that generates a single value for the conditional distribution for each component of the random vector X, rather than all values of the variables at the same time.

$\mathbf{X} = X_1, X_2, ..., X_K$: a random vector of K components

Assume we know $P(X_k \mid \mathbf{X}_{-k})$, where $\mathbf{X}_{-k} = X_1, X_2, ..., X_{k-1} X_{k+1}, ..., X_K$

Gibbs sampling algorithm (Gibbs sampler)

1. Initialize $x_1(0), x_2(0), ..., x_K(0)$.
2. $i \leftarrow 1$

   $x_1(1) \sim P(X_1 \mid x_2(0), x_3(0), x_4(0), ..., x_K(0))$

   $x_2(1) \sim P(X_2 \mid x_1(1), x_3(0), x_4(0), ..., x_K(0))$

   $x_3(1) \sim P(X_3 \mid x_1(1), x_2(1), x_3(0), ..., x_K(0))$

   …

   $x_k(1) \sim P(X_k \mid x_1(1), x_2(1), ..., x_{k-1}(1), x_{k+1}(0), x_K(0))$

   …

   $x_K(1) \sim P(X_K \mid x_1(1), x_2(1), ..., x_{K-1}(1))$
3. If (termination condition not met), then $i \leftarrow i+1$ and go to step 2.

Summary

1. **Uncertainty** arises because of both laziness and ignorance. It is inescapable in nondeterministic, or partially observable environments.
2. **Probabilities** express the agent's inability to reach a definite decision regarding the truth of a sentence. Probabilities summarize the agent's beliefs relative to the evidence.
3. Basic probability statements include **prior probabilities** and **conditional probabilities** over simple and complex propositions.
4. The **full joint probability distribution** specifies the probability of each complete assignment of values to random variables.
5. **Bayes' rule** allows unknown probabilities to be computed from known conditional probabilities, usually in the causal direction.
6. **Conditional independence** brought about by direct causal relationships in the domain might allow the full joint distribution to be factored into smaller, conditional distributions.
7. **Information theory:** information, entropy, joint entropy, conditional entropy, mutual information, cross, entropy, relative entropy (KL divergence)