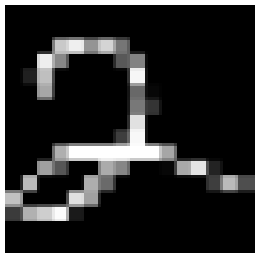
课程大纲

- 第一部分：人工智能概述/Introduction and Agents (chapters 1,2)
- 第二部分：问题求解/Search (chapters 3,4,5,6)
- 第三部分：知识与推理/Logic (chapters 7,8,9)
- 第四部分：不确定知识与推理/Uncertainty (chapters 13,14)
  - 概率101
  - 概率模型
- 第五部分：学习/Learning (chapters 20,21)
  - 监督学习（分类）
  - 非监督学习（聚类，降维）

手写数字识别

分类问题

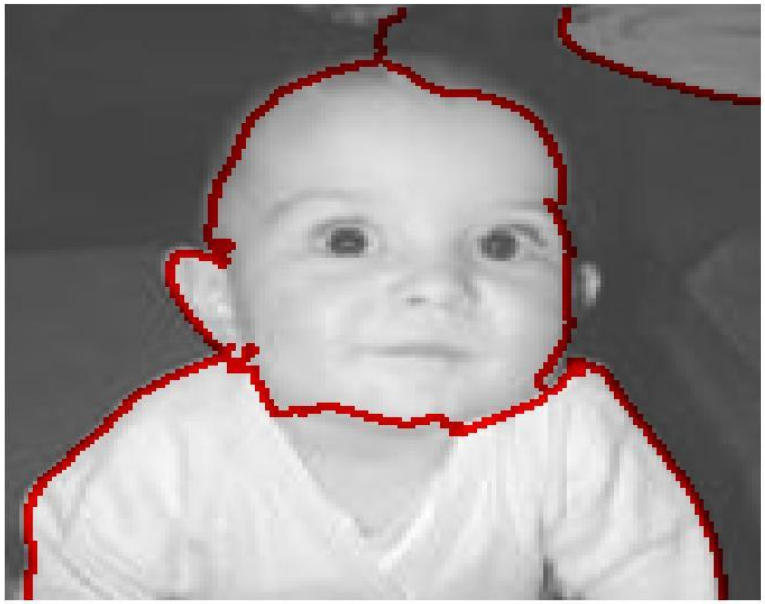


{2,3} or {+1,-1}

文本分类

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

图像分割



Uncertainty
不确定性

Chapter 13

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes‘ Rule

  — 独立性及贝叶斯法则

Uncertainty

- An agent can never be completely certain about the state of the world/domain since there is often ambiguity and uncertainty

- Plausible/probabilistic inference
  - I've got this evidence; what's the chance that this conclusion is true?
    - I've got a sore neck; how likely am I to have meningitis（脑膜炎）?

Uncertainty

- Say we have a rule:

  if toothache then problem is cavity

- But not all patients have toothaches due to cavities, so we could set up rules like:

  if toothache and ¬gum-disease（牙龈疾病） and ¬filling（补牙） and ... then problem = cavity

- This gets complicated; better method:

  if toothache then problem is cavity with 0.8 probability

  or *P(cavity | toothache) = 0.8*

  - *the probability of cavity is 0.8 given toothache is observed*

Uncertainty

Let action A_t = leave for airport t minutes before flight

Will A_t get me there on time?

Problems:

1. partial observability/部分可观察性 (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood（错误风险）: "A_{25} will get me there on time", or
2. leads to conclusions that are too weak for decision making:

   "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Uncertainty in the world and our models

- True uncertainty: rules are probabilistic in nature
  - rolling dice, flipping a coin
- Laziness: too hard to determine exception-less rules
  - takes too much work to determine all of the relevant factors
  - too hard to use the enormous rules that result
- Theoretical ignorance: don't know all the rules
  - problem domain has no complete, consistent theory (e.g., medical diagnosis)
- Practical ignorance: do know all the rules BUT
  - haven't collected all relevant information for a particular case

Method for handling uncertainty

Probability theory serves as a formal means for

- Representing and reasoning with uncertain knowledge
- Model degrees of belief（信度）in a proposition (event, conclusion, diagnosis, etc.)
- Given the available evidence,

  A_{25} will get me there on time with probability 0.04

Probability is the language of uncertainty

- Central pillar of modern AI

Probability

概率理论提供了一种方法以概括来自我们的惰性和无知的不确定性。

Probabilistic assertions summarize effects of

- Laziness（惰性）: failure to enumerate exceptions（例外）, qualifications（条件）, etc.
- Ignorance（无知）: lack of relevant facts, initial conditions, etc.

Subjective probability（主观概率）:

- Probabilities relate propositions（命题） to agent's own state of knowledge
  - e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions（断言） about the world

Probabilities of propositions change with new evidence:

- e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time | …) = 0.04
$P(A_{90}$ gets me there on time | …) = 0.70
$P(A_{120}$ gets me there on time | …) = 0.95
$P(A_{1440}$ gets me there on time | …) = 0.9999

Which action to choose?

Depends on my preferences（偏好） for missing flight vs. time spent waiting, etc.

Utility theory（效用理论） is used to represent and infer preferences

Decision theory = probability theory + utility theory
决策理论=概率理论+效用理论

Syntax

Basic element: random variable（随机变量）

- A random variable is some aspect of the world about which we (may) have uncertainty
- Are capitalized (usually) e.g., Cavity, Weather, Temperature

Similar to propositional logic: possible worlds defined by assignment of values to random variables.

Boolean random variables（布尔随机变量）

e.g., *Cavity（牙洞）* (do I have a cavity?)

Discrete random variables（离散随机变量）

e.g., *Weather* is one of <*sunny, rainy, cloudy, snow*>

Domain values must be exhaustive（穷尽的） and mutually exclusive（互斥的）

Continuous random variables（连续随机变量）

e.g., Temp=21.6; also allow, e.g., Temp < 22.0

Syntax

Elementary proposition（命题） constructed by assignment of a value to a random variable: e.g., *Weather* = *sunny*, *Cavity* = *false*
(abbreviated as ¬*cavity*)

Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather* = *sunny* ∨ *Cavity* = *false*

Syntax

Atomic event: A complete specification of the state of the world about which the agent is uncertain

原子事件：对智能体无法确定的世界状态的一个完整的详细描述。

E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

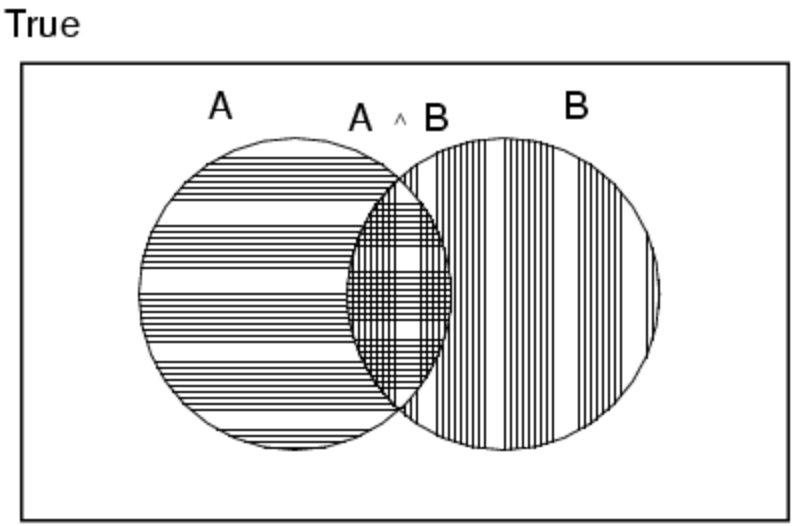
$Cavity = false \wedge Toothache = false$
$Cavity = false \wedge Toothache = true$
$Cavity = true \wedge Toothache = false$
$Cavity = true \wedge Toothache = true$

Atomic events are mutually exclusive and exhaustive

Axioms（公理） of probability

For any propositions *A*, *B*

- $0 \leq P(A) \leq 1$
- $P(true) = 1$ and $P(false) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Prior probability（先验概率）

Prior or unconditional probabilities（无条件概率） of propositions

在没有任何其它信息存在的情况下关于命题的信度

e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72

correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

概率分布给出一个随机变量所有可能取值的概率

P(*Weather*) = <0.72,0.1,0.08,0.1> (normalized（归一化的）, i.e., sums to 1)

Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables (i.e., every sample point)

联合概率分布给出一个随机变量集的值的全部组合的概率

P(*Weather*,*Cavity*) = a 4 × 2 matrix of values:

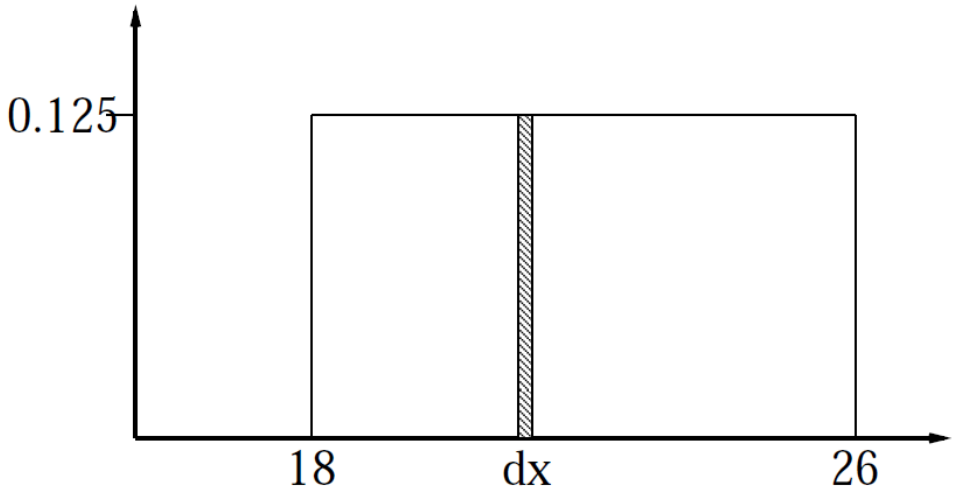
Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Probability for continuous variables

Express distribution as a parameterized（参数化的） function of value:

P(X =x) = U[18, 26](x) = uniform（均匀分布） density between 18 and 26



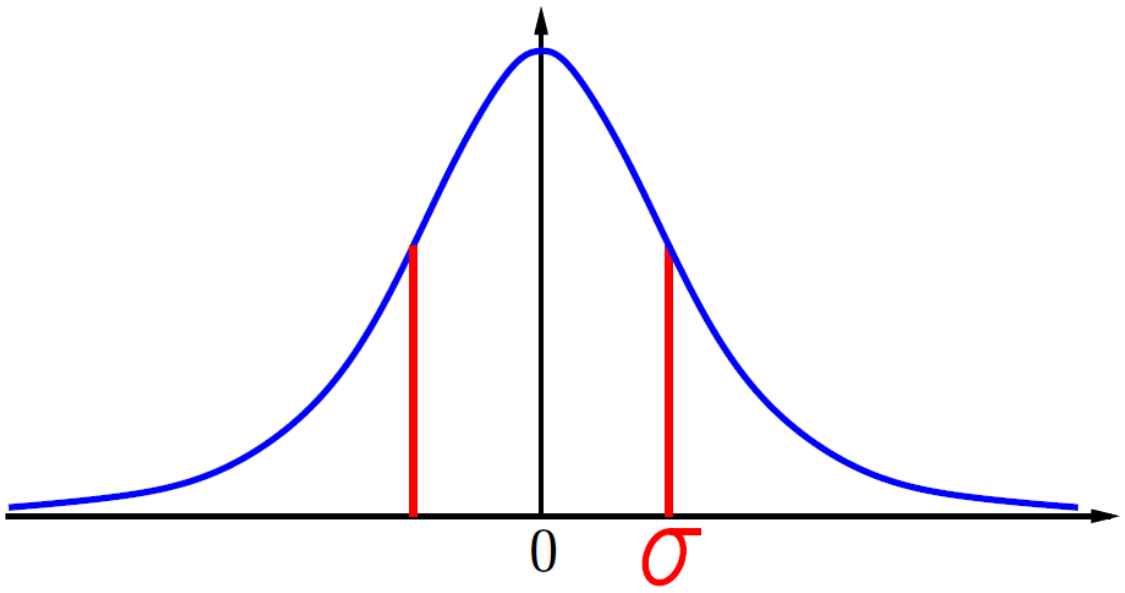
Here P is a density; integrates to 1.

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

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

Probability for continuous variables

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



Marginal Distributions （边缘概率分布）

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t, s)$$

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional probability（条件概率）

Conditional or posterior probabilities（后验概率）P(a|b)

- formalizes the process of accumulating evidence and updating probabilities based on new evidence
- Specifies the belief in a proposition (event, conclusion, diagnosis, etc.) that is conditioned on a proposition (evidence, feature, symptom, etc.) being true

e.g., P(*cavity* | *toothache*) = 0.8

i.e., given that *toothache* is all I know

- (Notation for conditional distributions（条件概率分布）:
  - P(*cavity* | *toothache*) = a single number
  - **P**(Cavity, Toothache) = 2x2 table summing to 1
  - **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 may be irrelevant, allowing simplification, e.g.,

  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

Definition of conditional probability:

$P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$

Product rule（乘法规则） 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.,

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

(View as a set of 4×2 equations, not matrix multiplication)

Chain rule（链式法则） is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n | X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n_1} | X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n | X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \textstyle\prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1})\end{aligned}$$

Conditional probability

Conditional probabilities behave exactly like standard probabilities, for example:

$0 <= P(a \mid e) <= 1$

conditional probabilities are between 0 and 1 inclusive

$P(a_1 \mid e) + P(a_2 \mid e) + \ldots + P(a_k \mid e) = 1$

conditional probabilities sum to 1 where $a_1, \ldots, a_k$ are all values in the domain of random variable A

$P(\neg a \mid e) = 1 - P(a \mid e)$

negation for conditional probabilities

Inference by enumeration

Start with the joint probability distribution（全联合概率分布）:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

一个命题的概率等于所有当它为真时的原子事件的概率和

$$P(\Phi) = \sum_{\omega:\omega \models \Phi} P(\omega)$$

Inference by enumeration

Start with the joint probability distribution（全联合概率分布）:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

一个命题的概率等于所有当它为真时的原子事件的概率和

$$P(\Phi) = \sum_{\omega:\omega \models \Phi} P(\omega)$$

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

Inference by enumeration

Start with the joint probability distribution（全联合概率分布）:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

For any proposition ϕ, sum the atomic events where it is true:

一个命题的概率等于所有当它为真时的原子事件的概率和

$$P(\Phi) = \sum_{\omega:\omega \models \Phi} P(\omega)$$

P(*cavity* ∨ *toothache*)

= 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28

Inference by enumeration

Start with the joint probability distribution（全联合概率分布）:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

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

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

Normalization（归一化）

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Denominator（分母） can be viewed as a normalization constant α

P(*Cavity* | *toothache*) = α **P**(*Cavity, toothache*)
= α [**P**(*Cavity, toothache, catch*) + **P**(*Cavity, toothache,* ¬ *catch*)]
= α [<0.108,0.016> + <0.012,0.064>]
= α <0.12,0.08> = <0.6,0.4>

General idea: compute distribution on query variable by fixing evidence variables（证据变量） and summing over hidden variables（未观测变量）

Inference by enumeration, contd.

Typically, we are interested in

the posterior joint distribution of the query variables（查询变量） **Y**

given specific values **e** for the evidence variables（证据变量） **E**

Let the hidden variables（未观测变量） be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$

The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables（ **Y**, **E**, **H**构成了域中所有变量的完整集合）

Obvious problems:

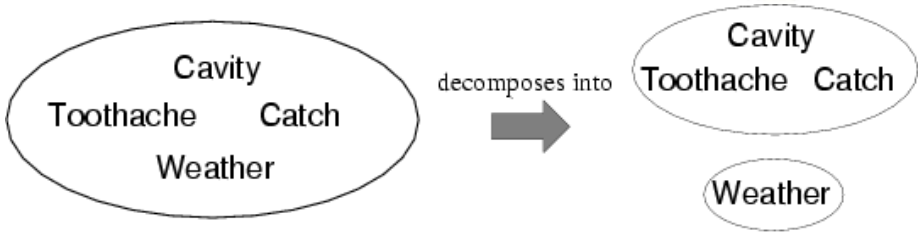
1. Worst-case time complexity $O(d^n)$ where d is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries?

Independence（独立性）

A and *B* are independent iff

$\mathbf{P}(A \mid B) = \mathbf{P}(A)$ or $\mathbf{P}(B \mid A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$

E.g: roll of 2 die: P({1},{3}) = 1/6*1/6 = 1/36



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

32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$

Absolute independence powerful but rare

Dentistry（牙科领域）is a large field with hundreds of variables, none of which are independent. What to do?

Independence misused

- An innocent old math joke

A famous statistician would never travel by airplane, because he had studied air travel and estimated that the probability of there being a bomb on any given flight was one in a million, and he was not prepared to accept these odds.

One day, a colleague met him at a conference far from home. "How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of two bombs are (1/1,000,000) x (1/1,000,000). This is a very, very small probability, which I can accept. So now I bring my own bomb along!"

Conditional independence
条件独立性

- Random variables can be dependent, but **conditionally independent**

- Example: Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary don't talk to each other

- Is *JohnCall independent of MaryCall?*
  - **No** – If John called, it is likely the alarm went off, which increases the probability of Mary calling
  - $P(MaryCall \mid JohnCall) \neq P(MaryCall)$

Conditional independence

- But, if we *know* the status of the alarm, JohnCall will *not* affect whether or not Mary calls

  $P(MaryCall \mid Alarm, JohnCall) = P(MaryCall \mid Alarm)$

- We say *JohnCall* and *MaryCall* are conditionally independent given Alarm

- In general, "*A and B are conditionally independent given C*" *means:*

  $P(A \mid B, C) = P(A \mid C)$

  $P(B \mid A, C) = P(B \mid C)$

  $P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$

Conditional independence

P(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries

Domain knowledge: Cavity directly causes toothache and probe-catches. If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)

The same independence holds if I haven't got a cavity:

(2) **P**(*catch* | *toothache*, ¬*cavity*) = **P**(*catch* | ¬*cavity*)

Catch is conditionally independent of *Toothache* given *Cavity*:

P(*Catch* | *Toothache, Cavity*) = **P**(*Catch* | *Cavity*)

Equivalent statements:

P(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)

P(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Conditional independence contd.

Write out full joint distribution using chain rule:

P(*Toothache, Catch, Cavity*)

= **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch, Cavity*)

= **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch* | *Cavity*) **P**(*Cavity*)

= **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity)

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

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*.

在大多数情况下，使用条件独立性能将全联合概率的表示由n的指数关系减为n的线性关系。

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule（贝叶斯法则）

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Bayes' Rule（贝叶斯法则）

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

$\Rightarrow$ Bayes' rule: $P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$

or in distribution form

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

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems (e.g. ASR, MT)

- In the running for most important AI equation!

Bayes' Rule

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

Useful for assessing diagnostic probability（诊断概率） from causal probability（因果概率）:

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis（脑膜炎）, S be stiff neck（脖子僵硬）:

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!
Note: you should still get stiff necks checked out! Why?

Bayes' Rule in Practice

Probabilistic Inference Using Bayes' Rule: I

H = "having a headache"

F = "coming down with Flu"

- P(H)=1/10
- P(F)=1/40
- P(H|F)=1/2

One day you wake up with a headache. You come with the following reasoning: "since 50% of flues are associated with headaches, so I must have a 50-50 chance of coming down with flu"

Is this reasoning correct?

Probabilistic Inference Using Bayes' Rule: I

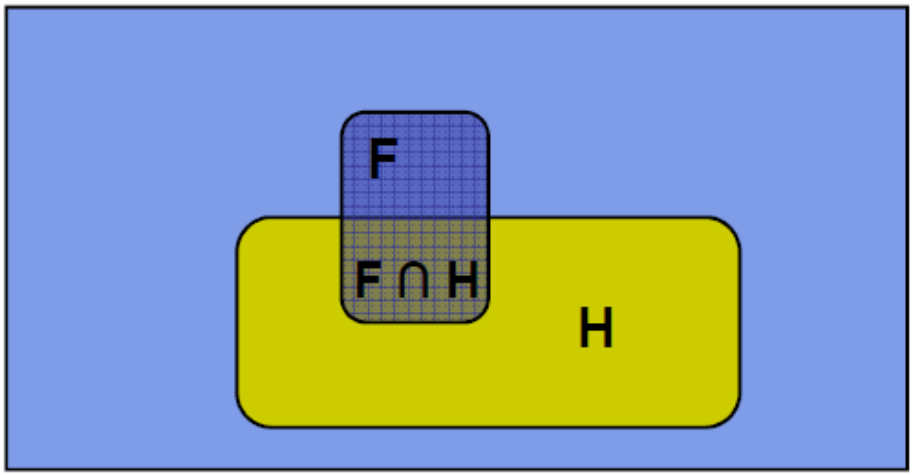
H = "having a headache"

F = "coming down with Flu"

- P(H)=1/10
- P(F)=1/40
- P(H|F)=1/2

The Problem:

P(F|H) = ?



Probabilistic Inference Using Bayes' Rule: I

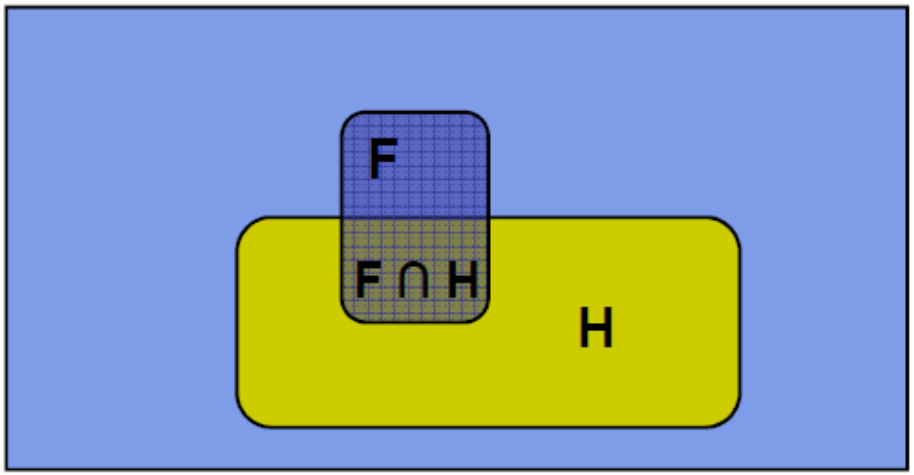
H = "having a headache"

F = "coming down with Flu"

- P(H)=1/10
- P(F)=1/40
- P(H|F)=1/2

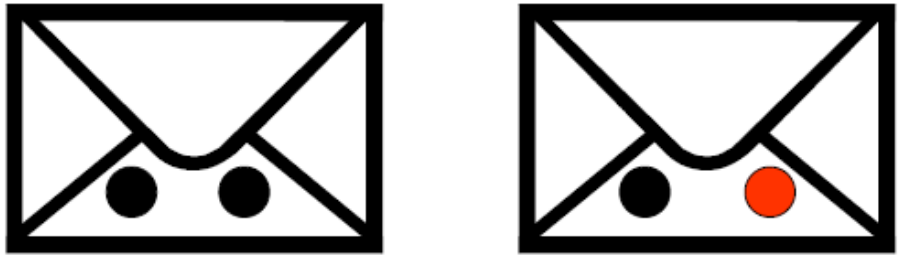
The Problem:

$$P(F|H) = P(H|F)\,P(F)/P(H)$$
$$= 1/8 \neq P(H|F)$$



Probabilistic Inference Using Bayes' Rule: II

- In a bag there are two envelopes
  - one has a red ball (worth $100) and a black ball
  - one has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it's black
- At this point you're given the option to switch the envelope. To switch or not to switch?

Probabilistic Inference Using Bayes' Rule: II

E: envelope, 1=(R,B), 2=(B,B)

B: the event of drawing a black ball

$P(E|B) = P(B|E)*P(E) / P(B)$

We want to compare $P(E=1|B)$ *vs.* $P(E=2|B)$

$P(B|E=1) = 0.5, P(B|E=2) = 1$

$P(E=1) = P(E=2) = 0.5$

$P(B) = P(B|E=1)P(E=1) + P(B|E=2)P(E=2) = (.5)(.5) + (1)(.5) = .75$

$P(E=1|B) = P(B|E=1)P(E=1)/P(B) = (.5)(.5)/(.75) = 1/3$

$P(E=2|B) = P(B|E=2)P(E=2)/P(B) = (1)(.5)/(.75) = 2/3$

After seeing a black ball, the posterior probability of this envelope being 1 (thus worth $100) is *smaller than it being 2*

Thus you should switch

Quiz

- A doctor performs a test that has 99% reliability, i.e., 99% of people who are sick test positive, and 99% of people who are healthy test negative. The doctor estimates that 1% of the population is sick.

- Question: A patient tests positive. What is the chance that the patient is sick?
- 0-25%, 25-75%, 75-95%, or 95-100%?

Quiz

- A doctor performs a test that has 99% reliability, i.e., 99% of people who are sick test positive, and 99% of people who are healthy test negative. The doctor estimates that 1% of the population is sick.

- Question: A patient tests positive. What is the chance that the patient is sick?
- 0-25%, 25-75%, 75-95%, or 95-100%?

- Intuitive answer: 99%; Correct answer: 50%

Bayes' rule with multiple evidence and conditional independence

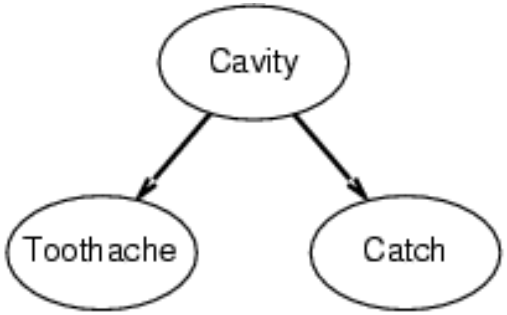
P(*Cavity* | *toothache* ∧ *catch*)

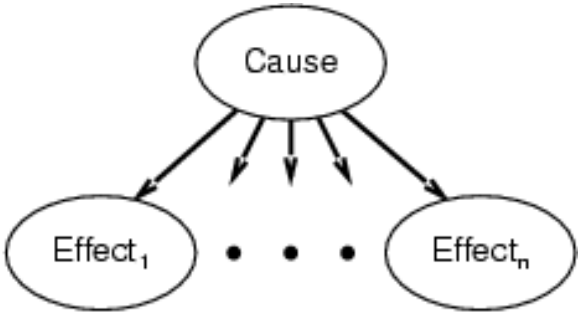
= α**P**(*toothache* ∧ *catch* | *Cavity*) **P**(*Cavity*)

= α**P**(*toothache* | *Cavity*) **P**(*catch* | *Cavity*) **P**(*Cavity*)

This is an example of a naïve Bayes model（朴素贝叶斯模型）:

$$\mathbf{P}(Cause, Effect_1, ..., Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$





Total number of parameters（参数） is linear in *n*

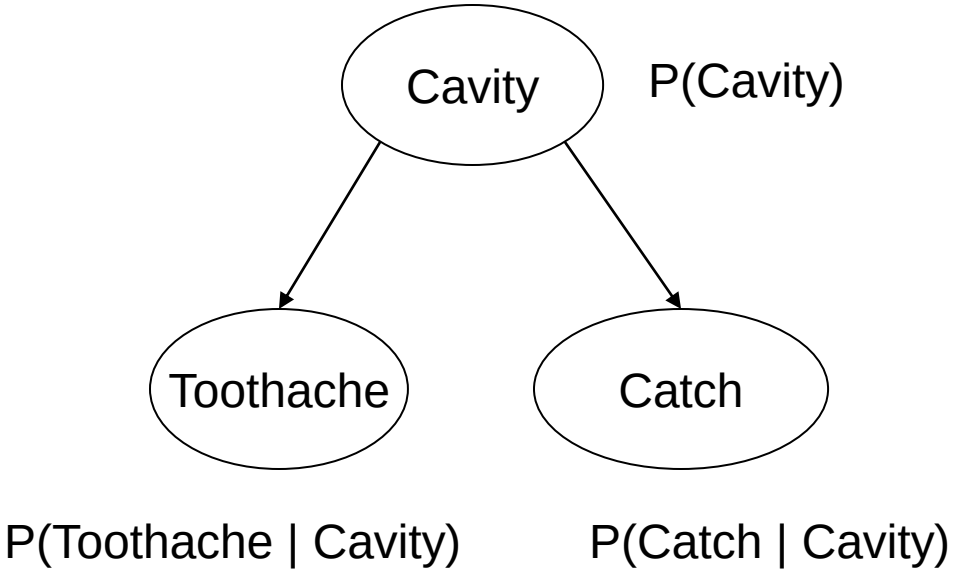
The Chain Rule

- Write out full joint distribution using chain rule:
  - P(Toothache, Catch, Cavity)

    = P(Toothache | Catch, Cavity) P(Catch, Cavity)

    = P(Toothache | Catch, Cavity) P(Catch | Cavity) P(Cavity)

    = P(Toothache | Cavity) P(Catch | Cavity) P(Cavity)



Graphical model notation:

- *Each variable is a node*
- *The parents of a node are the other variables which the decomposed joint conditions on*
- *MUCH more on this to come!*

Where do probability distributions come from?

- Idea One: Human, Domain Experts
  - Harder than it sounds
  - E.g. what's P(raining | cold)?

- Idea Two: Simpler probability facts and some algebra

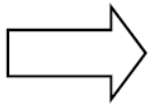
e.g., P(F)

P(B)

P(H|¬F,B)

P(H|F,¬B)

…



¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

- Use chain rule and independence assumptions to compute joint distribution

Where do probability distributions come from?

- Idea Three: Learn them from data!
  - A good chunk of machine learning research is essentially about various ways of learning various forms of them!

Estimation

- How to estimate the a distribution of a random variable X?

- *Maximum likelihood* （最大似然）:
  - Collect observations from the world
  - For each value x, look at the empirical rate of that value:

$$\hat{P}(x) = \frac{\text{count}(x)}{\text{total samples}}$$

  - This estimate is the one which maximizes the likelihood of the data

Estimation

- Problems with maximum likelihood estimates:
  - If I flip a coin once, and it's heads, what's the estimate for P(heads)?
  - What if I flip it 50 times with 27 heads?
  - What if I flip 10M times with 8M heads?

- Basic idea:
  - We have some prior expectation about parameters (here, the probability of heads)
  - Given little evidence, we should skew towards our prior
  - Given a lot of evidence, we should listen to the data

Summary of important rules

- Conditional Probability: $P(A|B) = P(A,B)/P(B)$
- Product rule: $P(A,B) = P(A|B)P(B)$
- Chain rule: $P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(C|D)P(D)$
- Conditionalized version of Chain rule:

  $$P(A,B|C) = P(A|B,C)P(B|C)$$

- Bayes' rule: $P(A|B) = P(B|A)P(A)/P(B)$
- Conditionalized version of Bayes' rule:

  $$P(A|B,C) = P(B|A,C)P(A|C)/P(B|C)$$

- Addition / Conditioning rule: $P(A) = P(A,B) + P(A,\neg B)$

  $$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

Summary

Probability is a rigorous formalism for uncertain knowledge
概率是对不确定知识一种严密的形式化方法

Joint probability distribution specifies probability of every atomic event
全联合概率分布指定了对随机变量的每种完全赋值，即每个原子事件的概率

Queries can be answered by summing over atomic events
可以通过把对应于查询命题的原子事件的条目相加的方式来回答查询

For nontrivial domains, we must find a way to reduce the joint size

Independence and conditional independence provide the tools

作业

- 13.8，13.11，13.18（第二版）=13.15，13.18，13.22（第三版）