

# Uncertainty and Bayesian Networks

吉建民

USTC

[jianmin@ustc.edu.cn](mailto:jianmin@ustc.edu.cn)

2024 年 4 月 14 日

## Used Materials

Disclaimer: 本课件采用了 S. Russell and P. Norvig's Artificial Intelligence –A modern approach slides, 徐林莉老师课件和其他网络课程课件, 也采用了 GitHub 中开源代码, 以及部分网络博客内容

# 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, 10, 11, 12)
- ▶ 第四部分：不确定知识与推理 / Uncertainty (chapters 13, 14, 15, 16, 17)
- ▶ 第五部分：学习 / Learning (chapters 18, 19, 20, 21)
- ▶ 第六部分：应用 / Application (chapters 22, 23, 24, 25)

# Table of Contents

Uncertainty

Probability

Syntax and Semantics

Inference

Independence and Bayes' Rule

Bayesian network

Graphical models

Bayesian networks

Inference in Bayesian networks

# Uncertainty

Let action  $A_t =$  “leave for airport  $t$  minutes before flight”

Will  $A_t$  get me there on time?

Problems:

- ▶ partial observability (部分可观察性, e.g., road state, other drivers' plans, etc.)
- ▶ noisy sensors (e.g., traffic reports)
- ▶ uncertainty in action outcomes (e.g., flat tire, etc.)
- ▶ immense complexity of modeling and predicting traffic

# Uncertainty

Hence a purely logical approach either:

- ▶ risk falsehood (错误风险): “ $A_{25}$  will get me there on time”,  
or
- ▶ 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 ...)

# Method for handling uncertainty

## Probability

- ▶ Model agent's degree of belief (信度)
- ▶ Given the available evidence,  
 $A_{25}$  will get me there in time with probability 0.04

# Table of Contents

Uncertainty

**Probability**

Syntax and Semantics

Inference

Independence and Bayes' Rule

Bayesian network

Graphical models

Bayesian networks

Inference in Bayesian networks



# Probability

概率理论提供了一种方法来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 (理论的无知)** : 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.,  $P(A_{25} | \text{no reported accidents}) = 0.06$

These are **not** assertions (断言) about the world

Probabilities of propositions change with new evidence:

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

# Making decisions under uncertainty

Suppose I believe the following:

- ▶  $P(A_{25} \text{ gets me there on time} \mid \dots) = 0.04$
- ▶  $P(A_{90} \text{ gets me there on time} \mid \dots) = 0.70$
- ▶  $P(A_{120} \text{ gets me there on time} \mid \dots) = 0.95$
- ▶  $P(A_{1440} \text{ gets me there on time} \mid \dots) = 0.9999$

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

# Table of Contents

Uncertainty

Probability

**Syntax and Semantics**

Inference

Independence and Bayes' Rule

Bayesian network

Graphical models

Bayesian networks

Inference in Bayesian networks

# Syntax

Basic element: **random variable (随机变量)**

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  $\neg$  cavity)

Complex propositions formed from elementary propositions and standard logical connectives.

e.g., Weather = sunny  $\vee$  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:

- ▶  $\text{Cavity} = \text{false} \wedge \text{Toothache} = \text{false}$
- ▶  $\text{Cavity} = \text{false} \wedge \text{Toothache} = \text{true}$
- ▶  $\text{Cavity} = \text{true} \wedge \text{Toothache} = \text{false}$
- ▶  $\text{Cavity} = \text{true} \wedge \text{Toothache} = \text{true}$

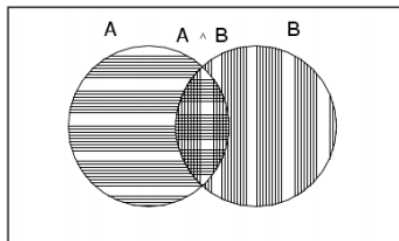
Atomic events are mutually exclusive (**互斥**) and exhaustive (**穷尽的**)

# Axioms (公理) of probability

For any propositions  $A$ ,  $B$

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

True



## 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.,  $P(\text{Cavity} = \text{true}) = 0.1$  and  $P(\text{Weather} = \text{sunny}) = 0.72$   
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

e.g.,  $P(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$  (**normalized (归一化的)**, i.e., sums to 1)



## Prior probability (先验概率)

**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

e.g.,  $P(\text{Weather}, \text{Cavity}) =$  a  $4 \times 2$  matrix of values:

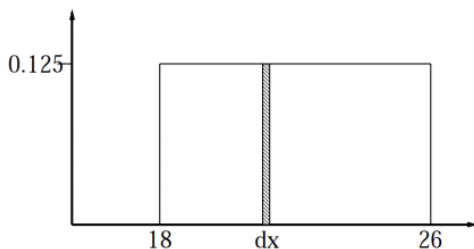
Weather =	sunny	rainy	doudy	snow
Cavity = true	0.144	0.02	0.016	0,02
Cavity = false	0.576	0.08	0.064	0.08

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:

$P(X = x) = U[18, 26](x) = \text{uniform (均匀分布) density between 18 and 26}$



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

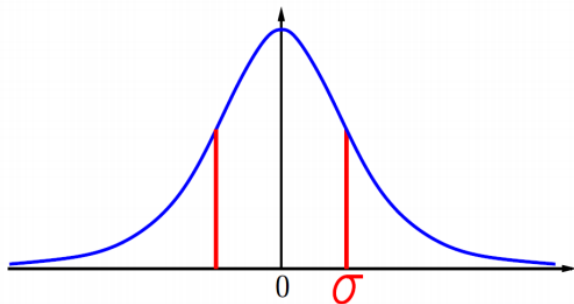
$P(X = 20.5) = 0.125$  means

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

# Probability for continuous variables

Normal distribution:

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



# Conditional probability (条件概率)

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

*e.g.*,  $P(\text{cavity}|\text{toothache}) = 0.8$

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

Notation for conditional distributions (条件概率分布) :

$P(\text{Cavity}|\text{Toothache}) =$  a  $2 \times 2$  matrix of values

If we know more, *e.g.*, cavity is also given, then we have

$P(\text{cavity}|\text{toothache}, \text{cavity}) = 1$

New evidence may be irrelevant, allowing simplification, *e.g.*,

$P(\text{cavity}|\text{toothache}, \text{sunny}) = P(\text{cavity}|\text{toothache}) = 0.8$

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

# Conditional probability

Definition of conditional probability:

$$P(a|b) = P(a \wedge b) / P(b) \quad \text{if } P(b) > 0$$

**Product rule (乘法规则)** gives an alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

A general version holds for whole distributions, e.g.,

$$P(\text{Weather}, \text{Cavity}) = P(\text{Weather} | \text{Cavity}) P(\text{Cavity})$$

(View as a set of  $4 \times 2$  equations, not matrix multiplication)

**Chain rule (链式法则)** is derived by successive application of product rule:  $P(X_1, \dots, X_n) = P(X_1, \dots, X_{n-1})P(X_n | X_1, \dots, X_{n-1})$

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

= ...

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

# Table of Contents

Uncertainty

Probability

Syntax and Semantics

**Inference**

Independence and Bayes' Rule

Bayesian network

Graphical models

Bayesian networks

Inference in Bayesian networks

## Inference by enumeration

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

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

For any proposition  $\phi$ , 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) :

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

For any proposition  $\phi$ , 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

$$P(\phi) = \sum_{\omega(\omega \models \phi)} P(\omega)$$

$$P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$



## Inference by enumeration

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

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

For any proposition  $\phi$ , 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

$$P(\phi) = \sum_{\omega(\omega \models \phi)} P(\omega)$$

$$P(\text{cavity} \vee \text{toothache}) =$$

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

## Inference by enumeration

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

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

Can also compute conditional probabilities:

$$\begin{aligned} P(\neg \text{cavity} | \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \end{aligned}$$

## Normalization (归一化)

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	.072	.008
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	.144	.576

Denominator (分母) can be viewed as a **normalization constant**  $\alpha$

$$P(\text{Cavity}|\text{toothache}) = \alpha P(\text{Cavity}, \text{toothache})$$

$$\begin{aligned} &= \alpha [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\ &= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] \\ &= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle \end{aligned}$$

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  $H = X - Y - E$

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

$$P(Y|E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)$$

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)

## Inference by enumeration, contd.

Obvious problems:

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

# Table of Contents

Uncertainty

Probability

Syntax and Semantics

Inference

**Independence and Bayes' Rule**

Bayesian network

Graphical models


Bayesian networks

Inference in Bayesian networks

# Independence (独立性)

A and B are independent iff

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B) \text{ or } P(A, B) = P(A)P(B)$$

e.g. roll of 2 die:  $P(1,3) = 1/6 * 1/6 = 1/36$  



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

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?

## Conditional independence (条件独立性)

$P(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$  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:

$$\text{— } P(\textit{catch}|\textit{toothache}, \textit{cavity}) = P(\textit{catch}|\textit{cavity})$$

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

$$\text{— } P(\textit{catch}|\textit{toothache}, \neg\textit{cavity}) = P(\textit{catch}|\neg\textit{cavity})$$

Catch is conditionally independent of Toothache given Cavity:

$$\text{— } P(\textit{Catch}|\textit{Toothache}, \textit{Cavity}) = P(\textit{Catch}|\textit{Cavity})$$

Equivalent statements:

$$P(\textit{Toothache}|\textit{Catch}, \textit{Cavity}) = P(\textit{Toothache}|\textit{Cavity})$$

$$P(\textit{Toothache}, \textit{Catch}|\textit{Cavity}) =$$

$$P(\textit{Toothache}|\textit{Cavity})P(\textit{Catch}|\textit{Cavity})$$



## Conditional independence contd.

Write out full joint distribution using chain rule:

$$\begin{aligned}P(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) &= P(\textit{Toothache}|\textit{Catch}, \textit{Cavity})P(\textit{Catch}, \textit{Cavity}) \\ &= P(\textit{Toothache}|\textit{Catch}, \textit{Cavity})P(\textit{Catch}|\textit{Cavity})P(\textit{Cavity}) \\ &= P(\textit{Toothache}|\textit{Cavity})P(\textit{Catch}|\textit{Cavity})P(\textit{Cavity})\end{aligned}$$

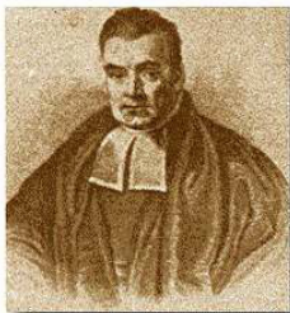
*i.e.*  $2 + 2 + 1 = 5$  independent numbers

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  $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$  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 (诊断概率)** from **causal probability (因果概率)** :

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$

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

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 * 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

# Probabilistic Inference Using Bayes' Rule

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% chance of coming down with flu”

Is this reasoning correct?

# Probabilistic Inference Using Bayes' Rule

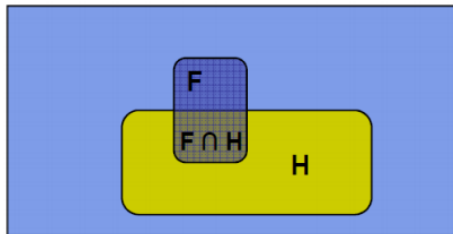
H = "having a headache"

F = "coming down with Flu"

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

The Problem:

$P(F|H) = ?$



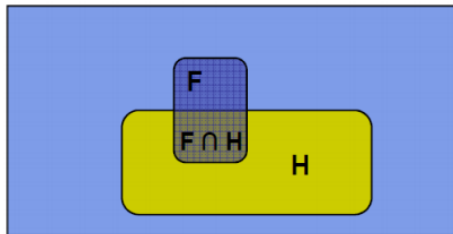
# Probabilistic Inference Using Bayes' Rule

H = "having a headache" F = "coming down with Flu"

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

The Problem:

$$\begin{aligned}P(F|H) &= P(H|F)P(F)/P(H) \\ &= 1/8 \\ &\neq P(H|F)\end{aligned}$$



## Bayes' Rule and conditional independence

$$\begin{aligned}P(\text{Cavity}|\text{toothache} \wedge \text{catch}) \\ &= \alpha P(\text{toothache} \wedge \text{catch}|\text{Cavity})P(\text{Cavity}) \\ &= \alpha P(\text{toothache}|\text{Cavity})P(\text{catch}|\text{Cavity})P(\text{Cavity})\end{aligned}$$

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

$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i|\text{Cause})$$



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

# Where do probability distributions come from?

- ▶ Idea One: Human, Domain Experts
- ▶ Idea Two: Simpler probability facts and some algebra

e.g.,  $P(F)$

$P(B)$

$P(H|\neg F, B)$

$P(H|F, \neg B)$

...



$\neg F$	$\neg B$	$\neg H$	0.4	
$\neg F$	$\neg B$	H	0.1	
$\neg F$	B	$\neg H$	0.17	
$\neg F$	B	H	0.2	
F	$\neg B$	$\neg H$	0.05	
F	$\neg B$	H	0.05	
F	B	$\neg H$	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!

# Summary of Uncertainty

- ▶ 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

- ▶ 第三版：13.15, 13.18, 13.21, 13.22

说明：第二版、第三版对应的题目内容不同，不过都能起到相似的训练目的，区别不大

# Table of Contents

Uncertainty

Probability

Syntax and Semantics

Inference

Independence and Bayes' Rule

**Bayesian network**

Graphical models

Bayesian networks

Inference in Bayesian networks

# Frequentist vs. Bayesian

## 客观 vs. 主观

**Frequentist (频率主义者)**: probability is the long-run expected frequency of occurrence.  $P(A) = n/N$ , where  $n$  is the number of times event  $A$  occurs in  $N$  opportunities.

“某事发生的概率是 0.1”意味着 0.1 是在无穷多样本的极限条件下能够被观察到的比例

But 在许多情景下不可能进行重复试验

e.g., 发生第三次世界大战的概率是多少?

**Bayesian**: degree of belief. It is a measure of the plausibility (似然性) of an event given incomplete knowledge.

# Probability

- ▶ 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

# Independence/Conditional Independence

A and B are **independent** iff

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B) \text{ or } P(A, B) = P(A)P(B)$$

A is **conditionally independent** of B given C:

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

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 environment.**

# Probability Theory

Probability theory can be expressed in terms of two simple equations

- ▶ Sum Rule (加法规则)

- ▶ probability of a variable is obtained by marginalizing (边缘化) or summing out other variables

$$p(a) = \sum_b p(a, b)$$

- ▶ Product Rule (乘法规则)

- ▶ joint probability expressed in terms of conditionals

$$p(a, b) = p(b|a)p(a)$$

All probabilistic inference and learning amounts to repeated application of sum and product rule



# Outline

- ▶ Graphical models (概率图模型)
- ▶ Bayesian networks
  - ▶ Syntax (语法)
  - ▶ Semantics (语义)
- ▶ Inference (推导) in Bayesian networks

# What are Graphical Models?

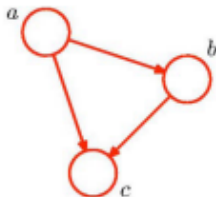
They are diagrammatic (图表的) representations of probability distributions

—marriage between probability theory and graph theory

- ▶ Also called probabilistic graphical models
- ▶ They augment analysis instead of using pure algebra (代数)

# What is a Graph?

- ▶ Consists of nodes (also called vertices) and links (also called edges or arcs)



- ▶ In a probabilistic graphical model
  - ▶ each node represents a random variable (or group of random variables)
  - ▶ Links express probabilistic relationships between variables

# Graphical Models in CS

- ▶ Natural tool for handling uncertainty (不确定性) and complexity (复杂性)
  - which occur throughout applied mathematics and engineering
- ▶ Fundamental to the idea of a graphical model is the notion of modularity (模块性)
  - a complex system is built by combining simpler parts.

# Why are Graphical Models useful

- ▶ Probability theory provides the glue whereby
  - ▶ the parts are combined, ensuring that the system as a whole is consistent
  - ▶ providing ways to interface models to data.
- ▶ Graph theoretic side provides:
  - ▶ Intuitively appealing interface
    - by which humans can model highly-interacting sets of variables
  - ▶ Data structure
    - that lends itself naturally to designing efficient general-purpose (通用的) algorithms

# Graphical models: Unifying Framework

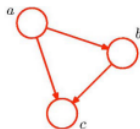
- ▶ View classical multivariate (多变量的) probabilistic systems as instances of a common underlying formalism (形式)
  - ▶ mixture models (混合模型), factor analysis (因子分析), hidden Markov models, Kalman filters (卡尔曼滤波器), etc.
  - ▶ Encountered in systems engineering, information theory, pattern recognition and statistical mechanics
- ▶ Advantages of View:
  - ▶ Specialized techniques in one field can be transferred between communities and exploited
  - ▶ Provides natural framework for designing new systems

# Role of Graphical Models in Machine Learning

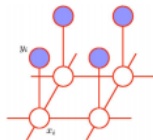
- ▶ Simple way to visualize (形象化)  
structure of probabilistic model
- ▶ Insights into properties of model  
Conditional independence properties by inspecting graph
- ▶ Complex computations  
required to perform inference and learning expressed as  
graphical manipulations

# Graph Directionality

- ▶ Directed graphical models
  - Directionality associated with arrows
- ▶ Bayesian networks
  - Express causal relationships (因果关系) between random variables
- ▶ More popular in AI and statistics



- ▶ Undirected graphical models
  - links without arrows
- ▶ Markov random fields (马尔科夫随机场)
  - Better suited to express soft constraints between variables
- ▶ More popular in Vision and physics





# Bayesian networks

一种简单的，图形化的数据结构，用于表示变量之间的依赖关系（条件独立性），为任何全联合概率分布提供一种简明的规范。

## Syntax:

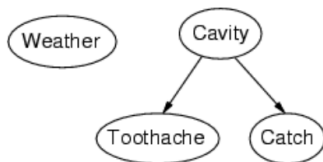
- ▶ a set of nodes, one per variable
- ▶ a directed (有向), acyclic (无环) graph (link  $\approx$  "direct influences")
- ▶ a conditional distribution for each node given its parents:

$P(X_i | Parents(X_i))$ —量化其父节点对该节点的影响

In the simplest case, conditional distribution represented as a **conditional probability table 条件概率表 (CPT)** giving the distribution over  $X_i$  for each combination of parent values

## Example

Topology (拓扑结构) of network encodes conditional independence assertions:



Weather is independent of the other variables

Toothache and Catch are conditionally independent given Cavity

## Example

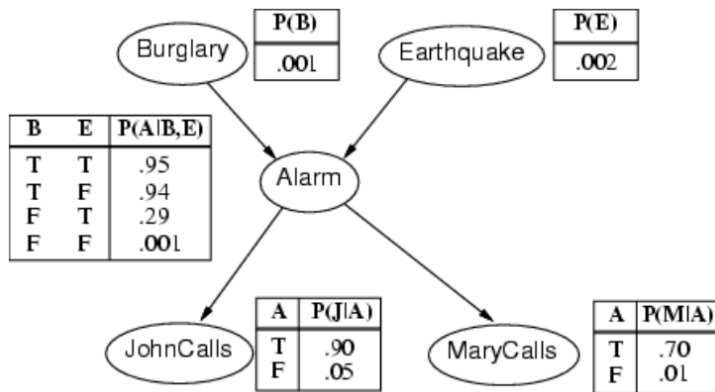
I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar (夜贼) ?

Variables: Burglary (入室行窃) , Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects “causal (因果) ” knowledge:

- ▶ A burglar can set the alarm off
- ▶ An earthquake can set the alarm off
- ▶ The alarm can cause Mary to call
- ▶ The alarm can cause John to call

## Example contd.

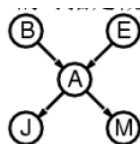


## Compactness (紧致性)

A CPT for Boolean  $X_i$  with  $k$  Boolean parents has  $2^k$  rows for the combinations of parent values

一个具有  $k$  个布尔父节点的布尔变量的条件概率表中有  $2^k$  个独立的可指定概率

Each row requires one number  $p$  for  $X_i = \text{true}$   
(the number for  $X_i = \text{false}$  is just  $1-p$ )



If each variable has no more than  $k$  parents, the complete network requires  $O(n \cdot 2^k)$  numbers

*i.e.* grows linearly with  $n$ , vs.  $O(2^n)$  for the full joint distribution

For burglary net,  $1 + 1 + 4 + 2 + 2 = 10$  numbers (vs.  $2^5 - 1 = 31$ )

# Global semantics (全局语义)

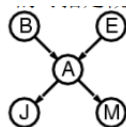
The full joint distribution is defined as the product of the local conditional distributions:

全联合概率分布可以表示为贝叶斯网络中的条件概率分布的乘积

“Global” semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$

e.g.,  $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$



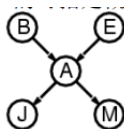
## Global semantics (全局语义)

The full joint distribution is defined as the product of the local conditional distributions:

全联合概率分布可以表示为贝叶斯网络中的条件概率分布的乘积

“Global” semantics defines the full joint distribution as the product of the local conditional distributions:

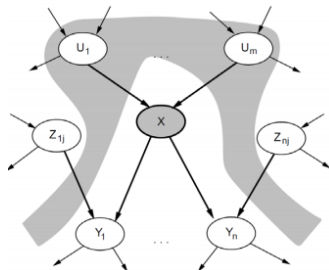
$$\begin{aligned} P(x_1, \dots, x_n) &= \prod_{i=1}^n P(x_i | \text{parents}(X_i)) \\ \text{e.g., } P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 * 0.7 * 0.001 * 0.999 * 0.998 \\ &\approx 0.00063 \end{aligned}$$



# Local semantics

Local semantics: each node is conditionally independent of its nondescendants (非后代) given its parents

给定父节点，一个节点与它的非后代节点是条件独立的



Theorem: Local semantics  $\Leftrightarrow$  global semantics



## Constructing Bayesian networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics  
需要一种方法使得局部的条件独立关系能够保证全局语义得以成立

1. Choose an ordering of variables  $X_1, \dots, X_n$
2. For  $i = 1$  to  $n$   
    add  $X_i$  to the network

    select parents from  $X_1, \dots, X_{i-1}$  such that

$$P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$$

This choice of parents guarantees the global semantics:

$$\begin{aligned} P(X_1, \dots, X_n) &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) (\text{chainrule}) \\ &= \prod_{i=1}^n P(X_i | \text{Parents}(X_i)) (\text{byconstruction}) \end{aligned} \tag{1}$$

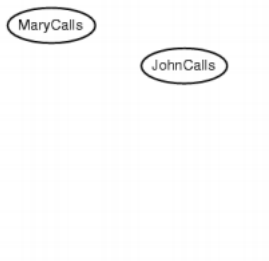
# Constructing Bayesian networks

要求网络的拓扑结构确实反映了合适的父节点集对每个变量的那些直接影响。

添加节点的正确次序是首先添加“根本原因”节点，然后加入受它们直接影响的变量，以此类推。

## Example

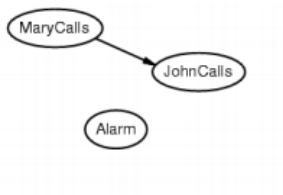
Suppose we choose the ordering M, J, A, B, E



- ▶  $P(J|M) = P(J)$ ?

## Example

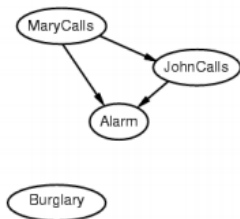
Suppose we choose the ordering  $M, J, A, B, E$



- ▶  $P(J|M) = P(J)$ ? **No**
- ▶  $P(A|J, M) = P(A|J)$ ?  $P(A|J, M) = P(A)$ ?

## Example

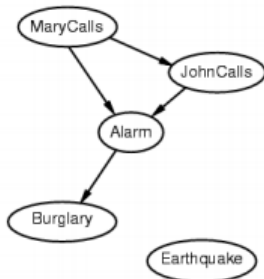
Suppose we choose the ordering  $M, J, A, B, E$



- ▶  $P(J|M) = P(J)$ ? **No**
- ▶  $P(A|J, M) = P(A|J)$ ?  $P(A|J, M) = P(A)$ ? **No**
- ▶  $P(B|A, J, M) = P(B|A)$ ?
- ▶  $P(B|A, J, M) = P(B)$ ?

## Example

Suppose we choose the ordering  $M, J, A, B, E$



- ▶  $P(J|M) = P(J)$ ? **No**
- ▶  $P(A|J, M) = P(A|J)$ ?  $P(A|J, M) = P(A)$ ? **No**
- ▶  $P(B|A, J, M) = P(B|A)$ ? **Yes**
- ▶  $P(B|A, J, M) = P(B)$ ? **No**
- ▶  $P(E|B, A, J, M) = P(E|A)$ ?
- ▶  $P(E|B, A, J, M) = P(E|A, B)$ ?

## Example

Suppose we choose the ordering  $M, J, A, B, E$



- ▶  $P(J|M) = P(J)$ ? **No**
- ▶  $P(A|J, M) = P(A|J)$ ?  $P(A|J, M) = P(A)$ ? **No**
- ▶  $P(B|A, J, M) = P(B|A)$ ? **Yes**
- ▶  $P(B|A, J, M) = P(B)$ ? **No**
- ▶  $P(E|B, A, J, M) = P(E|A)$ ? **No**
- ▶  $P(E|B, A, J, M) = P(E|A, B)$ ? **Yes**

## Example contd.



Deciding conditional independence is hard in noncausal (非因果) directions

(Causal models and conditional independence seem hardwired for humans!)

Network is less compact:  $1 + 2 + 4 + 2 + 4 = 13$  numbers needed



# Inference tasks

**Simple queries:** compute posterior probability  $P(X_i|E = e)$   
e.g.,  $P(\text{NoGas}|\text{Gauge} = \text{empty}, \text{Lights} = \text{on}, \text{Starts} = \text{false})$

Conjunctive queries (联合查询) :

$$P(X_i, X_j|E = e) = P(X_i|E = e)P(X_j|X_i, E = e)$$

Optimal decisions: decision networks include utility information;  
probabilistic inference required for  
 $P(\text{outcome}|\text{action}, \text{evidence})$

## Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

$$P(X|e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)$$

注：在贝叶斯网络中可以将全联合分布写成条件概率乘积的行驶：

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i))$$

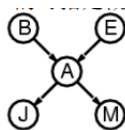
在贝叶斯网络中可以通过计算条件概率的乘积并求和来回答查询。

# Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation.

Simple query on the burglary network:

$$\begin{aligned} &P(B|j, m) \\ &= P(B, j, m) / P(j, m) \\ &= \alpha P(B, j, m) = \alpha \sum_e \sum_a P(B, e, a, j, m) \end{aligned}$$

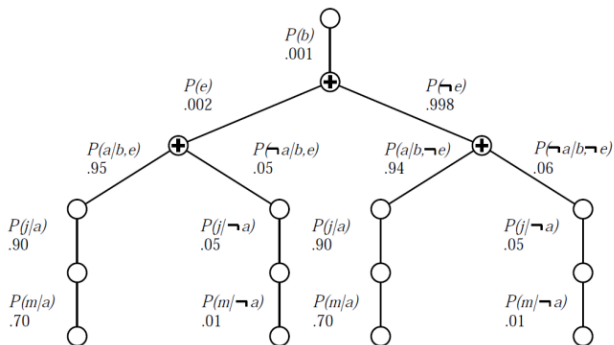


Rewrite full joint entries using product of CPT entries:

$$\begin{aligned} &P(B|j, m) \\ &= \alpha \sum_e \sum_a P(B)P(e)P(a|B, e)P(j|a)P(m|a) \\ &= \alpha P(B) \sum_e P(e) \sum_a P(a|B, e)P(j|a)P(m|a) \end{aligned}$$

Recursive depth-first enumeration:  $O(n)$  space,  $O(d^n)$  time

# Evaluation tree



Enumeration is inefficient: repeated computations  
e.g., computes  $P(j|a)P(m|a)$  for each value of  $e$

## Inference by variable elimination

Variable elimination (变量消元) : carry out summations right-to-left, storing intermediate results (factors: 因子) to avoid recomputation

$$\begin{aligned} \mathbf{P}(B|j, m) &= \alpha \underbrace{\mathbf{P}(B)}_B \sum_e \underbrace{P(e)}_E \sum_a \underbrace{\mathbf{P}(a|B, e)}_A \underbrace{P(j|a)}_J \underbrace{P(m|a)}_M \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B, e) P(j|a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B, e) f_J(a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a f_A(a, B, e) f_J(a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b, e) \text{ (sum out } A) \\ &= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E) \\ &= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b) \end{aligned}$$

## Complexity of exact inference

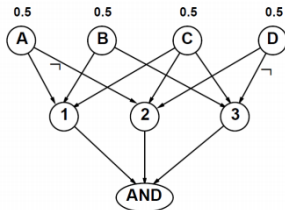
Singly connected networks 单联通网络 (or polytrees 多树, 对应的无向图是树):

- ▶ any two nodes are connected by at most one (undirected) path
- ▶ time and space cost of variable elimination are  $O(d^k n)$   
多树上的变量消元的时间和空间复杂度都与网络规模呈线性关系。

Multiply connected networks 多联通网络:

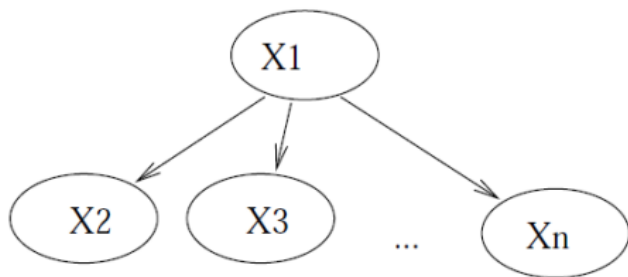
- ▶ can reduce 3SAT to exact inference  $\Rightarrow$  NP-hard
- ▶ equivalent to counting 3SAT models  $\Rightarrow$  #P-complete

1.  $A \vee B \vee C$
2.  $C \vee D \vee \neg A$
3.  $B \vee C \vee \neg D$



## Example: Naïve Bayes model

There is a single parent variable and a collection of child variables whose values are conditionally independent from one another given the parent.



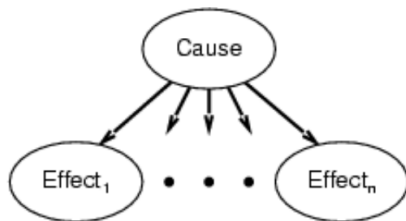
$$\begin{aligned} &P(X_1 = x_1, \dots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1) \dots P(X_n = x_n|X_1 = x_1) \end{aligned}$$

# Naïve Bayes model

$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})$$

$$P(\text{Cause} | \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Effects}, \text{Cause}) / P(\text{Effects})$$

$$= \alpha P(\text{Cause}, \text{Effects}) = \alpha P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})$$



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



## Example: Spam detection

Imagine the problem of trying to automatically detect spam e-mail messages (垃圾邮件) . A simple approach to get started is to look only at the “Subject:” headers in the e-mail messages and attempt to recognize spam by checking some simple computable features (特征) . The two simple features we will consider are:

- ▶ **Caps:** Whether the subject header is entirely capitalized
- ▶ **Free:** Whether the subject header contains the word ‘free’, either in upper case or lower case

e.g., a message with the subject header “NEW MORTGAGE RATE” is likely to be spam. Similarly, for “Money for Free”, “FREE lunch”, etc.

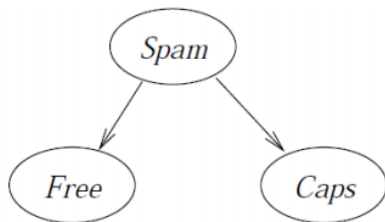
## Example: Spam detection

The model is based on the following three random variables, Caps, Free and Spam, each of which take on the values Y (for Yes) or N (for No)

- ▶ **Caps** = Y if and only if the subject of the message does not contain lowercase letters
- ▶ **Free** = Y if and only if the word 'free' appears in the subject (letter case is ignored)
- ▶ **Spam** = Y if and only if the message is spam

$$P(\text{Free}, \text{Caps}, \text{Spam}) = P(\text{Spam}) P(\text{Caps}|\text{Spam}) P(\text{Free}|\text{Spam})$$

## Example: Spam detection



$$P(\text{Free}, \text{Caps}, \text{Spam}) = P(\text{Spam})P(\text{Caps}|\text{Spam})P(\text{Free}|\text{Spam})$$

# Example: Spam detection

<i>Free</i>	<i>Caps</i>	<i>Spam</i>	# messages
Y	Y	Y	20
Y	Y	N	1
Y	N	Y	5
Y	N	N	0
N	Y	Y	20
N	Y	N	3
N	N	Y	2
N	N	N	49
Total:			100

<i>Spam</i>	$P(\text{Spam})$
Y	$\frac{20+5+20+2}{100} = 0.47$
N	$\frac{1+0+3+49}{100} = 0.53$

<i>Caps</i>	<i>Spam</i>	$P(\text{Caps} \text{Spam})$
Y	Y	$\frac{20+20}{20+5+20+2} \approx 0.8511$
Y	N	$\frac{1+3}{1+0+3+49} \approx 0.0755$
N	Y	$\frac{5+2}{20+5+20+2} \approx 0.1489$
N	N	$\frac{0+49}{1+0+3+49} \approx 0.9245$

<i>Free</i>	<i>Spam</i>	$P(\text{Free} \text{Spam})$
Y	Y	$\frac{20+5}{20+5+20+2} \approx 0.5319$
Y	N	$\frac{1+0}{1+0+3+49} \approx 0.0189$
N	Y	$\frac{20+2}{20+5+20+2} \approx 0.4681$
N	N	$\frac{3+49}{1+0+3+49} \approx 0.9811$

## Example: Spam detection

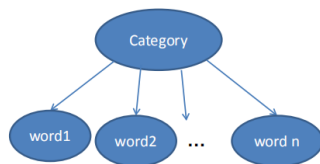
$$\begin{aligned} &P(\text{Free} = Y, \text{Caps} = N, \text{Spam} = N) \\ &= P(\text{Spam} = N)P(\text{Caps} = N|\text{Spam} = N)P(\text{Free} = Y|\text{Spam} = N) \\ &\approx 0.53 * 0.9245 * 0.0189 \\ &\approx 0.0093 \end{aligned}$$

## Example: Learning to classify text documents

文本分类是在文档所包含的文本基础上，把给定的文档分配到固定类别集合中某一个类别的任务。这个任务中常常用到朴素贝叶斯模型。在这些模型中，查询变量是文档类别，“结果”变量则是语言中每个词是否出现。我们假设文档中的词的出现都是独立的，其出现频率由文档类别确定。

- ▶ 准确地解释当给定一组类别已经确定的文档作为“训练数据”时，这样的模型是如何构造的。
- ▶ 准确地解释如何对新文档进行分类。
- ▶ 这里独立性假设合理吗？请讨论。

## Example: Learning to classify text documents



The model consists of the prior probability  $\mathbf{P}(\mathbf{Category})$  and the conditional probabilities  $\mathbf{P}(\mathbf{word\ }i|\mathbf{Category})$

- ▶  $P(\mathbf{Category}=c)$  is estimated as the fraction of all documents that are of category  $c$
- ▶  $P(\mathbf{word\ }i = \mathbf{true}|\mathbf{Category}=c)$  is estimated as the fraction of documents of category  $c$  that contain word  $i$

# Twenty Newsgroups

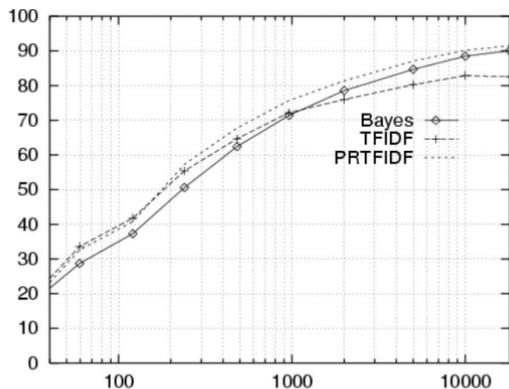
Given 1000 training documents from each group. Learn to classify new documents according to which newsgroup it came from

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

Naïve Bayes: 89% classification accuracy



# Learning Curve for 20 Newsgroups



Accuracy vs. Training set size (1/3 withheld for test)

# TFIDF

- ▶ TFIDF (tf-idf)

- ▶ Term Frequency:  $TF_w = \frac{\text{在某一类中词条 } w \text{ 出现的次数}}{\text{该类中所有的词条数目}}$

- ▶ Inverse Document Frequency:

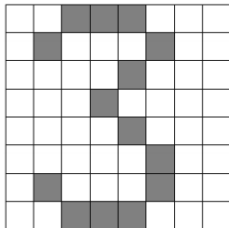
$$IDF_w = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条 } w \text{ 的文档数} + 1}\right)$$

- ▶  $TFIDF_w = TF_w \times IDF_w$ , 某一特定文件内的高词语频率, 以及该词语在整个文件集合中的低文件频率, 可以产生出高权重的 TFIDF

- ▶ PRTFIDF (A Probabilistic Classifier Derived from TFIDF)

# Example: A Digit Recognizer

- ▶ Input: pixel grids



- ▶ output: a digit 0-9

# Naïve Bayes for Digits

Simple version:

- ▶ One feature  $F_{ij}$  for each grid position  $\langle i, j \rangle$
- ▶ Possible feature values are on / off, based on whether intensity is more or less than 0.5 in underlying image
- ▶ Each input maps to a feature vector e.g.,

$$1 \rightarrow (F_{0,0} = 0, F_{0,1} = 0, F_{0,2} = 1, \dots, F_{15,15} = 0)$$

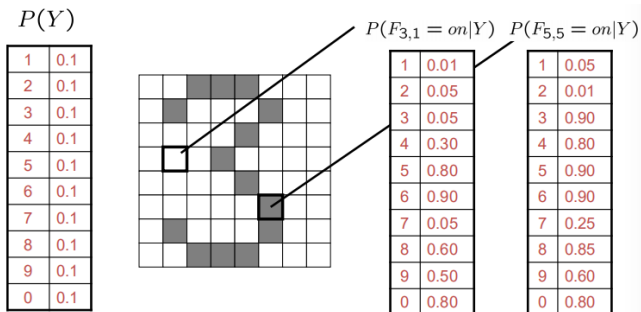
- ▶ Here: lots of features, each is binary

Naïve Bayes model:

$$P(Y|F_{0,0}, \dots, F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

What do we need to learn?

# Examples: CPTs



# Comments on Naïve Bayes

- ▶ Makes probabilistic inference tractable by making a strong assumption of conditional independence.
- ▶ Tends to work fairly well despite this strong assumption.
- ▶ Experiments show it to be quite competitive with other classification methods on standard datasets.
- ▶ Particularly popular for text categorization, e.g., spam filtering.

# Summary

- ▶ Bayesian networks provide a natural representation for (causally induced) conditional independence
- ▶ Topology + CPTs = compact representation of joint distribution
- ▶ Generally easy for domain experts to construct
- ▶ Exact inference by variable elimination:
  - ▶ polytime on polytrees, NP-hard on general graphs
  - ▶ space = time, very sensitive to topology
- ▶ Naïve Bayes model

# 作业

- ▶ 第三版：14.12, 14.13