概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(新疆大学)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

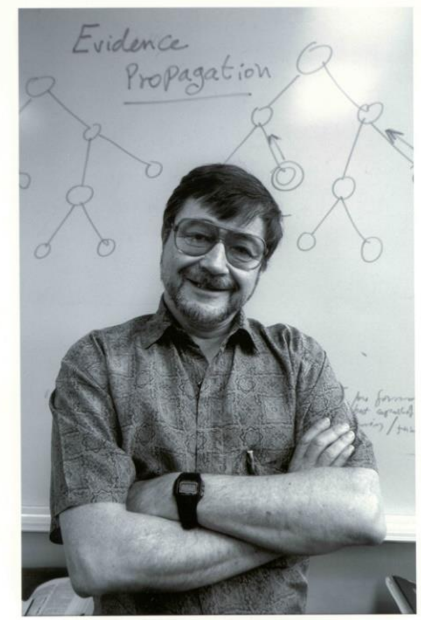
Tel: 62796193

Email: ozj@tsinghua.edu.cn

引言

- 概率建模，推理和学习
  - 人工智能
  - 处理不确定性是智能的一种重要表现
  - 不确定性是客观世界中的一种真实、广泛的存在
  - 2012年图灵奖: UCLA的Judea Pearl教授

开创性的工作—贝叶斯网络和消息传递, revolutionized AI



Uncertainty appears to be an inescapable aspect of most real-world applications. —— Koller & Friedman, p.2

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. —— Albert Einstein, 1956.

就认为数学的法则是源于现实而言，它们就不是确定性的；就它们是确定性的而言，它们就不是源于现实的。—— 爱因斯坦, 1956.

2018图灵奖

The technical achievements of this year's Turing Laureates, which have led to significant breakthroughs in AI technologies include, but are not limited to, the following:

- **Geoffrey Hinton**
  - Backpropagation
  - Boltzmann Machines
  - Improvements to convolutional neural networks

- **Yoshua Bengio**
  - Probabilistic models of sequences
  - High-dimensional word embeddings and attention
  - Generative adversarial networks (GANs)

- **Yann LeCun** Energy-based models
  - Convolutional neural networks
  - Improving backpropagation algorithms
  - Broadening the vision of neural networks

引言

- (概率)图模型
  - 概率论与图论相结合的产物
  - 为统计推理和学习提供了一个统一的灵活**框架**
  - **统一**了目前广泛应用的许多统计模型和方法
    - 如: 多元高斯模型、主成分分析（PCA）、因子分析（FA）、马尔可夫随机场（MRF）、条件随机场（CRF）、隐马尔科夫模型（HMM）、Kalman滤波、粒子滤波、变分推理、以及Turbo-codes、LDPC-codes 等

课程内容

- 图模型理论
  - 图论相关知识
  - 有向图模型（贝叶斯网络）
  - 无向图模型（马尔可夫随机场）
  - 图模型的推理理论（精确推理、采样近似、变分近似）
  - 图模型的学习理论（参数学习、结构学习）

（有向图模型与无向图模型：表示理论）

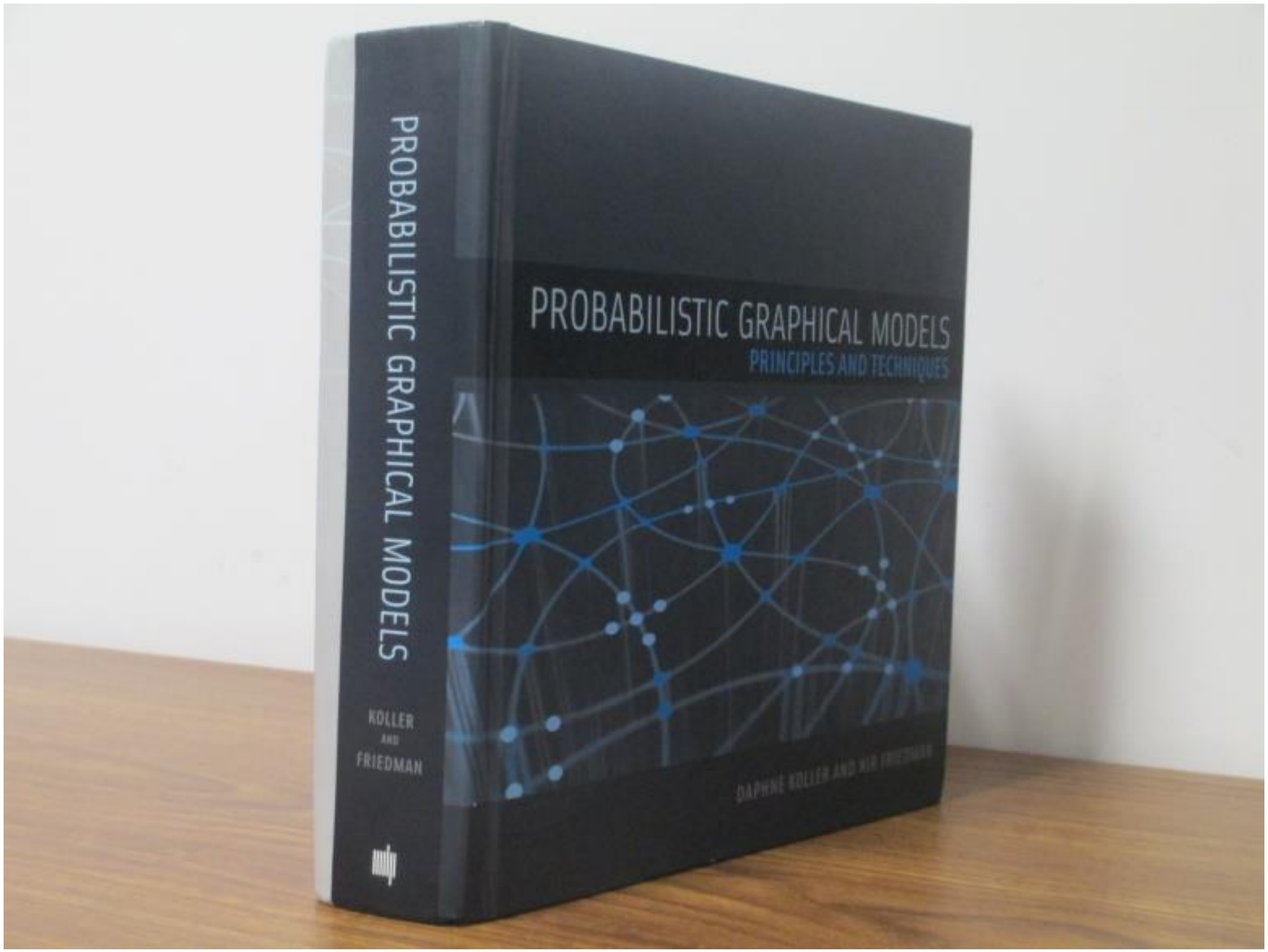
- 图模型应用
  - 语音识别
  - 文本处理/NLP
  - 图像处理/计算机视觉/CV
  - 通信信道编码

课程章节

- 第一章 图模型的表示理论（2）
  - Semantics (DGM, UGM)
  - HMM, CRF
- 第二章 图模型的推理理论（4）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第三章 图模型的学习理论（2）
  - 参数学习：maxlikelihoodEstimate，RFLearning, BayesEstimate
  - 结构学习：StructureLearning

	pgm-1 semantics		pgm-2 hmm-crf pgm-3 exact	pgm-4 kalman pgm-5 sampling
pgm-6 variational pgm-7 ML	pgm-8 Bayesian			

参考书


PROBABILISTIC GRAPHICAL MODELS
KOLLER
AND
FRIEDMAN
PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES

参考书

- Daphne Koller, Nir Friedman. “**Probabilistic graphical models : principles and techniques**”. MIT Press, c2009. （O212.8 FK81） KF书
  - Detailed, 1231 pages.
  - Representation, Inference, Learning
- R. G. Cowell, A. P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter. **"Probabilistic Networks and Expert Systems".** Springer-Verlag. 1999. （TP182 FP96） CDLS书
  - One of the best book available, although the treatment is restricted to exact inference.
- J. Pearl. **"Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference".** Morgan Kaufmann. 1988. （TP18 FP35）
  - The book that got it all started! A very insightful book, still relevant today.

- S. Lauritzen. **"Graphical Models".** Oxford. 1996.（国家图书馆2-97\O21\L38）
  - The definitive mathematical exposition of the theory of graphical models.
- M. I. Jordan (ed). **"Learning in Graphical Models".** MIT Press. 1999.（O157.5 FL43）
  - Loose collection of papers on machine learning, many related to graphical models. One of the few books to discuss *approximate* inference.
- Christopher M. Bishop. **"Pattern Recognition and Machine Learning".** Springer 2006. （电子书） Bishop书
  - Comprehensive, good reference.
- D.J. MacKay. **"Information Theory, Inference, and Learning Algorithms".** Cambridge Univ. Press, 2003. （电子书）
  - Information theory, coding.
- Kevin Patrick Murphy. "**Machine Learning: a Probabilistic Perspective**". MIT Press, 2012. Murhpy书
  - Newest. 1098 pages.
  - PMTK Matlab toolkit
  - More on UGMs than Bishop book

相关课程网站

- Probabilistic Graphical Models, CMU, Spring 2014, Eric Xing，29 lectures
  http://www.cs.cmu.edu/~epxing/Class/10708-14/lecture.html
- Statistical Learning Theory, Michael Jordan
  http://www.cs.berkeley.edu/~jordan/courses.html
- Probabilistic Models for Artificial Intelligence, Daphne Koller
  http://robotics.stanford.edu/~koller/courses.html
- Probabilistic Graphical Models, Carlos Guestrin
  http://www.cs.cmu.edu/~guestrin/teaching.html
- Graphical Models, Jeff Bilmes
  http://ssli.ee.washington.edu/people/bilmes/teaching-frame.html
- Probabilistic Inference Algorithms and Machine Learning, Brendan Frey
  http://www.psi.toronto.edu/~frey/apm/index.html
- Probabilistic graphical models, Kevin Murphy
  http://www.cs.ubc.ca/~murphyk/
- Machine Learning, Tommi Jaakkola
  http://people.csail.mit.edu/tommi/courses.html

Lectures

不安全 | oa.ee.tsinghua.edu.cn/~ouzhijian/pgm/pgm.htm

Zhijian Ou @ Tsinghua Univers

不安全 | oa.ee.tsinghua.edu.cn/~ouzhijian/teaching.htm

Zhijian Ou - Teaching

- Home
- **Teaching**
- Research
- Paper
- Software
- News

Undergraduate courses

- **Probability and Stochastic Processes (1)**
  2012 Spring, 2013 Spring, 2014 Spring,
  2016 Spring, 2017 Spring, 2018 Spring, 2019 Spring.
- **Probability and Stochastic Processes (2)**
  2012 Fall, 2013 Fall,
  2015 Fall, 2016 Fall, 2017 Fall, 2018 Fall.
- **Mathematics and engineering applications**
  2005 Spring, 2006 Spring, 2007 Spring, 2008 Spring, 2009 Spring, 2010 Spring, 2011 Spring.

Graduate courses

- **Theory and Applications of Probabilistic Graphical Models** [Course Material]
  2004 Fall, 2005 Fall, 2006 Fall, 2007 Fall, 2008 Fall, 2009 Fall, 2010 Fall, 2011 Fall, 2012 Fall,
  2017 Spring, 2018 Spring, 2019 Spring.

Short courses I hosted

- Y0230321: **Machine Learning for Speech Processing**
  12/17/2012 - 12/21/2012
  By Dr. Shinji Watanabe, Mitsubishi Electric Research Laboratories (MERL), USA.
- **Advanced Monte Carlo methods**
  2:00 - 3:30 PM, Aug. 20 [Mon], 22 [Wed], 24 [Fri], 27 [Mon], 29 [Wed], 31 [Fri], 2018.
  By Prof. Zhiqiang Tan, Rutgers University, USA.

预备知识

- 概率论
  - 随机变量

  - Capital letters X, Y, Z, X_i : discrete or continuous random variables (r.v.), uni-variate or multi-variate
  - Lower case letters x, y, z, x_i : their particular values (in general, vectors in a vector space)
  - A, B, C : sets of integers, e.g. $A = \{1,2,3\} = 1{:}3$
  - X_A : a set of r.v. indexed by A , e.g. $X_A = \{X_1, X_2, X_3\} = X_{1:3}$

预备知识(续)

- 概率论
  - 随机变量
  - 概率分布

  - 离散随机变量的概率分布函数 (probability mass function, pmf)
  - 连续随机变量的概率密度函数 (probability density function, pdf)

$$p(x) \triangleq p_X(x)$$

  - Let $X_{1:n}=\{X_1, \ldots, X_n\}$ be a random vector.

$$p(x_{1:n}) \triangleq p_{X_{1:n}}(x_{1:n})$$

预备知识(续)

- 概率论
  - 随机变量
  - 概率分布
  - 边缘分布，条件分布，贝叶斯公式

  - x, y, z 的联合概率分布: $p(x, y, z)$
  - x 的边缘概率分布: $p(x) = \Sigma_y \Sigma_z p(x, y, z)$ ← marginalization
  - $p(x, y, z) = p(x) \times p(y \mid x) \times p(z \mid x, y)$ ← chain rule (factorization)
  - 贝叶斯公式:

$$p(x \mid y) = \frac{p(x)\, p(y \mid x)}{p(y)}$$

预备知识(续)

- 概率论
  - 随机变量
  - 概率分布
  - 全概率，条件概率，贝叶斯公式
  - 独立性，条件独立性

  - Two r.v. X and Y are independent (written $X \perp Y$) if and only if

$$p(x, y) = p(x) \times p(y)$$

$$p(x \mid y) = p(x), \ \ p(y \mid x) = p(y)$$

  - $X \perp Y \mid Z$：

$$p(x, y \mid z) = p(x \mid z) \times p(y \mid z)$$

预备知识(续)

- 概率论
  - 随机变量
  - 概率分布
  - 全概率，条件概率，贝叶斯公式
  - 独立性，条件独立性
  - 数字特征 (期望，协方差矩阵，自相关矩阵)
  - 随机变量 X 的数学期望 (expectation)

$$E[X]=\begin{cases}\int x\cdot p(x)dx & \text{if } X \text{ is continuous}\\ \sum_x x\cdot p(x) & \text{if } X \text{ is discrete}\end{cases}$$

  - 协方差矩阵 (covariance matrix)

$$Cov[X]=E\left[(X-E[X])(X-E[X])^T\right]=\underline{E\left[XX^T\right]}-E[X]E[X]^T$$

第一章 引言

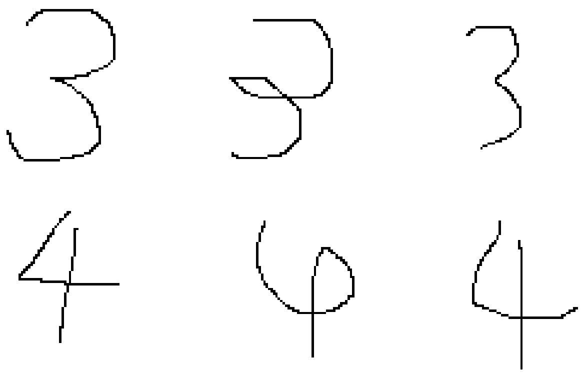
1.1 统计推理和学习的概念 { 模式识别 / 信号估计

1.2 简单图论知识

1.3 图模型入门

模式识别

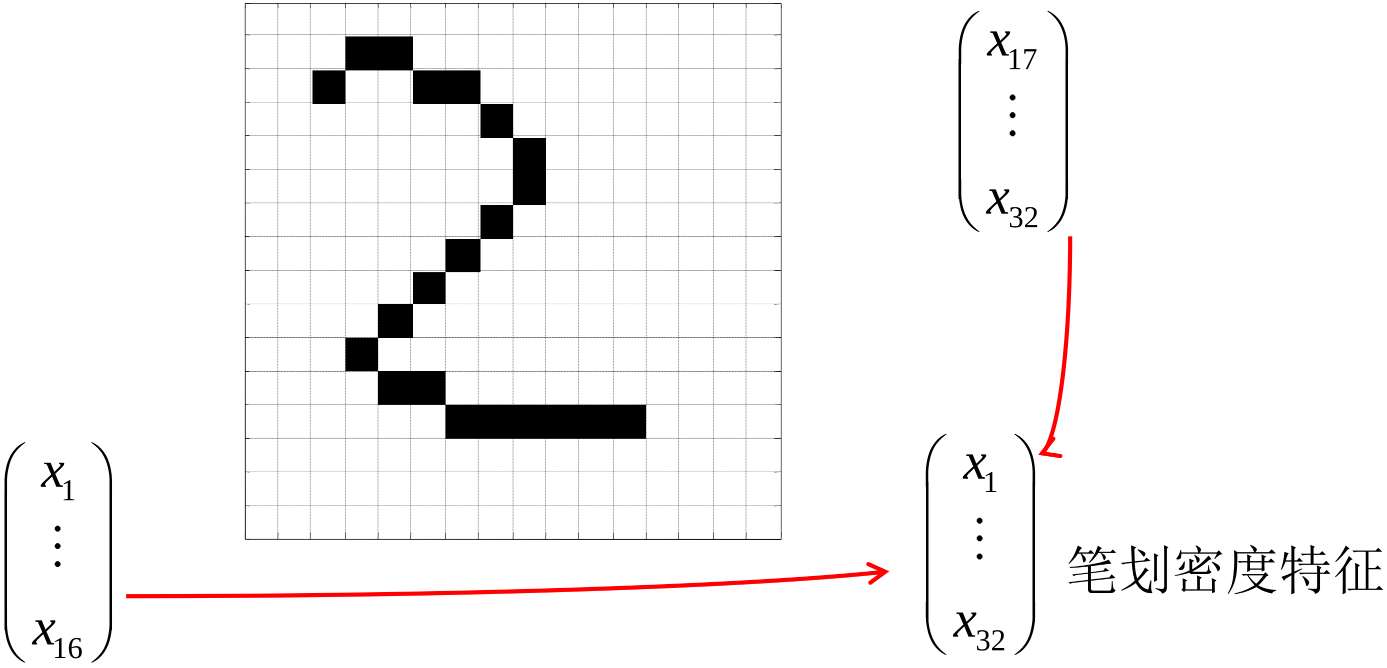
- 人们为了认识客观事物，把事物按相似的程度组成不同的类别



- 让计算机面对某一具体事物时能将其正确地归入某一类别？

模式识别

- 模式识别是把物体归入某一类别的过程
  - 物体类别未知 → W: discrete *class* variable, $W \in \{1,\ldots,K\}$
  - 对物体的观测（特征） → X: *observation* variable
- 建立概率模型: 一对随机变量 (W, X)

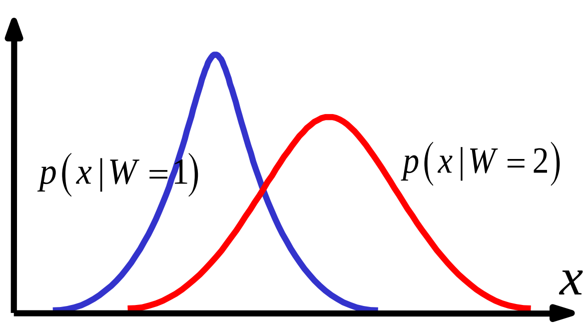


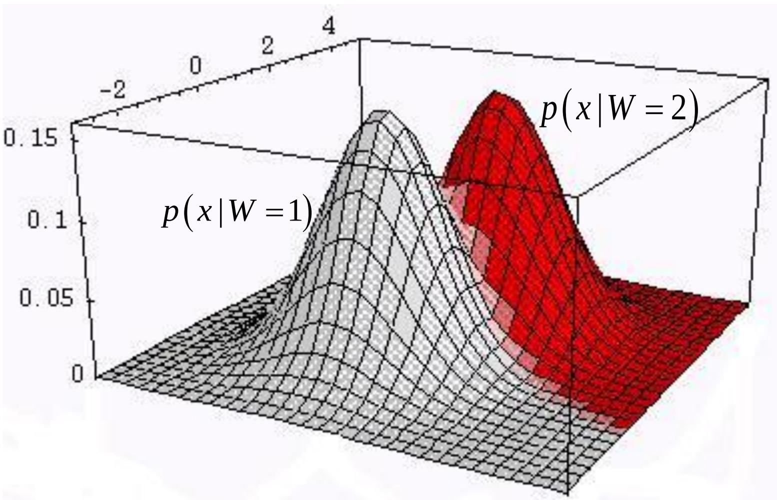
模式识别的概率模型

- 设有概率模型 $p(W, X) = p(W)\,p(X \mid W)$

$p(W)$ 类先验概率（e.g. 字频）

$p(X \mid W)$ 类条件分布：属于类别 $W=k$ 的物体的观测值的分布





模式识别的概率模型

- 设有概率模型 $p(W,X)=p(W)\,p(X\,|\,W)$
- 求: 观测到一个特定的X=x, 应该把 x 分到哪一类?
  - 猜猜看，尽可能猜中（分类错误率小）
  - 做一个决策 d , 将观测值 x 分到第 $d(x)$ 类
  - 错误率：$\min\limits_{d} P_e(d)=P\big(d(X)\neq W\big)$
- 解: $$d(x)=\arg\max_{k=1,\cdots,K} p\big(W=k\,|\,X=x\big)$$

对$\forall x$，选择使后验概率 $p(W=k\,|\,X=x)$ 最大的 k 作为分类结果 $d(x)$

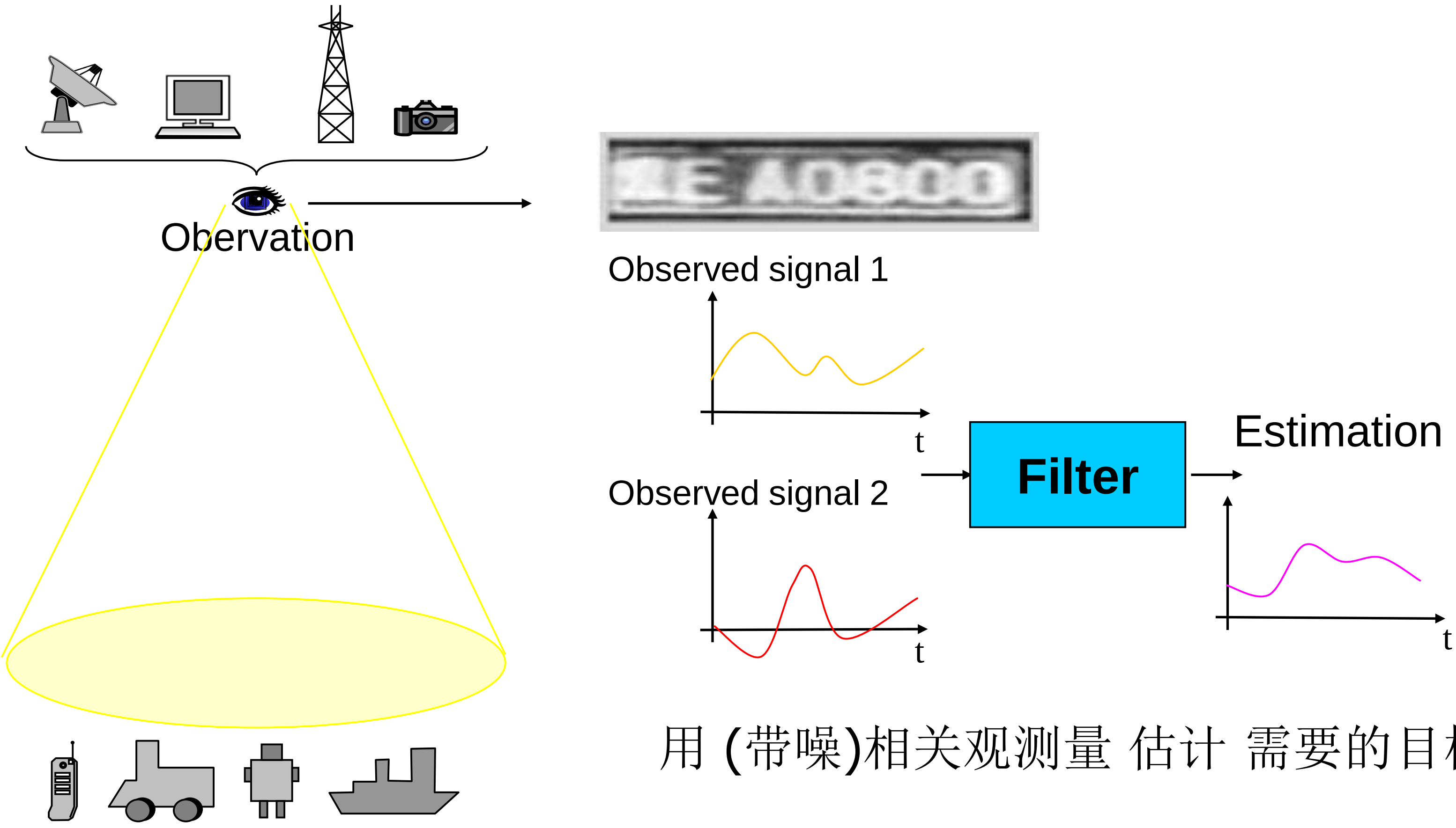
最大后验（MAP，Maximum A Posteriori）判决

MAP判决：广泛应用

max p(类别 | 观测值)

- 信道译码： p(发送信元 | 接收端波形)
- 语音识别： p(词序列 | 语音)
- 人脸图像识别： p(身份id | 人脸图像)
- 机器翻译： p(英文语句 | 中文语句)
- 网页分类： p(网页类别 | 网页)
- …

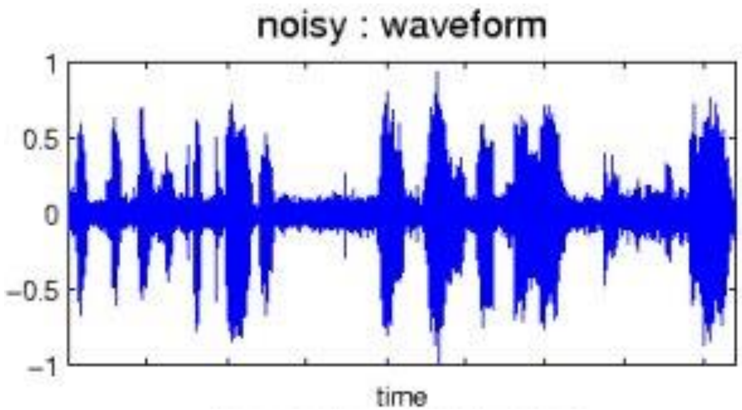
信号估计/滤波

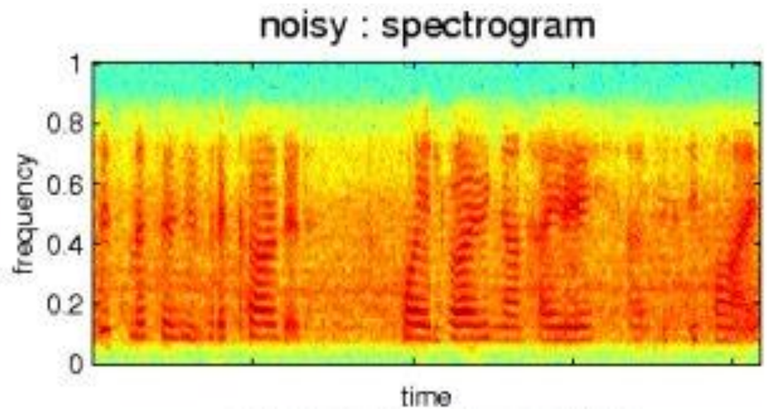


用 (带噪)相关观测量 估计 需要的目标量

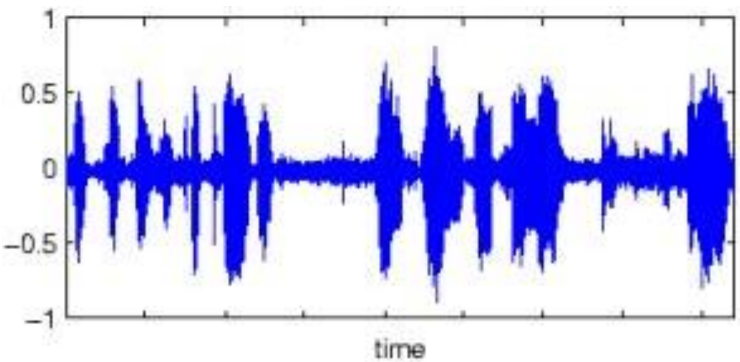
语音增强—演示

"this message is recorded while driving on highway... uh... sixty-five..."

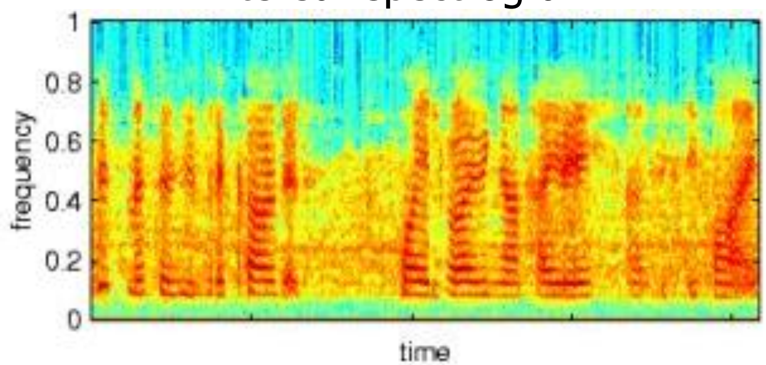




filtered: waveform



filtered: spectrogram



信号估计的概率方法

- 数学描述
  - 需要的目标量　　Z　　(未知)
  - (带噪)相关观测量的具体取值　　$X=x$　　(已知)
  - 联合概率分布　　$p_{Z,X}(z,x)$　　(给定)

求：对未知量的最佳预测 $z^{opt} = g(x)$?

- 衡量准则：最小均方误差 (Mean-Square-Error)

$$\min_g E\left[\left\|Z - g(X)\right\|^2\right]$$

- 答案：$g(x) = E[Z \mid X=x]$ 条件均值

推理(Inference)

Compute $p(H \mid O=o)$

	模式分类问题	信号估计问题
设	(W, X)	(Z, X)
求	最小分类错误率 $\min_{d} P\left(d(X) \neq W\right)$	最小均方误差 $\min_{g} E\left[\left\|Y - g(X)\right\|^2\right]$
解	$d(x) = \arg\max_{k=1,\cdots,K} P\left(W = k \mid X = x\right)$	$g(x) = E[Z \mid X=x]$

Infer unknown from observation

Infer unknown from observation

学习

- **Inference**

  Compute $p(H \mid O=o)$ using model $p(H, O)$

  - 所研究的对象的联合分布 $p(H, O)$ 是已知的！

- **Learning**

  Estimate $p(W, X)$ from data

  - 数据：随机变量 (W, X) 的若干实现/样本

    Samples: $(w^{(1)}, x^{(1)}), \ldots, (w^{(N)}, x^{(N)})$

60,000 images from
about 250 writers

Discriminative model example

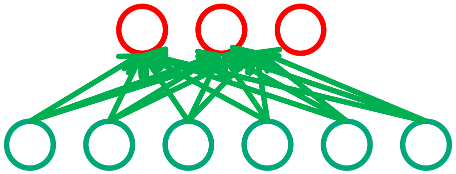
- Multi-class logistic regression / maxent classifier

$$p(W=k|X=x)=\frac{exp\left(w_k^T x+b_k\right)}{\sum_{j=1}^{K} exp\left(w_j^T x+b_j\right)}, w_k \in \mathbb{R}^{32}, b_k \in \mathbb{R}$$

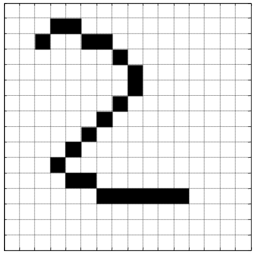
$$p(W=k|X=x)=\frac{exp(y_k)}{\sum_{j=1}^{K} exp\left(y_j\right)} \triangleq softmax(y_k)$$

$$where\ logit\ y_k = w_k^T x + b_k, k=1,\cdots,K.$$

$W=k$



$X=x$

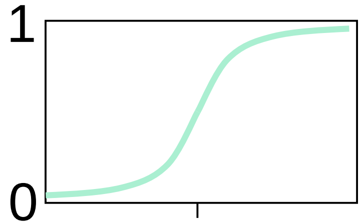


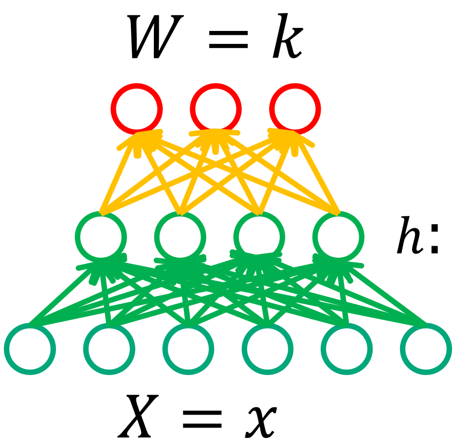
$$\begin{pmatrix} x_1 \\ \vdots \\ x_{32} \end{pmatrix}$$ 笔划密度特征

Discriminative model example

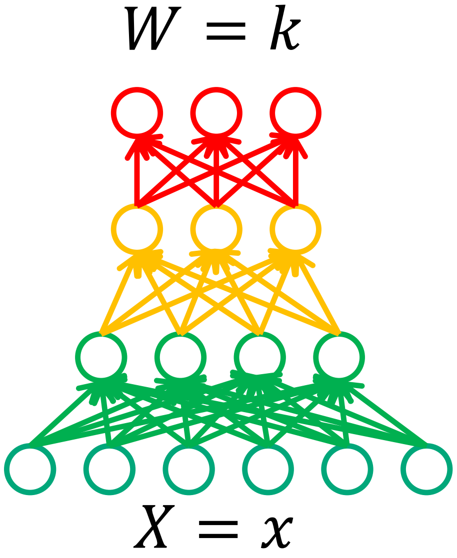
- Neural Networks

$$h = sigmoid(Ax + b)$$





h: Computational Node



推理和学习

- **不同领域的许多应用问题可归结为一种统计推理和学习**
  - 模式识别：未知变量为离散时的一种推理
  - 信号估计：未知变量为连续时的一种推理
  - 得到（联合）分布的过程是一种统计学习

  - 语音识别: $p(words \mid acoustics)$
  - 目标跟踪: $p(ObjectTrajectory \mid VideoInput)$
  - …

第一章 引言

1.1 统计推理和学习的概念

1.2 简单图论知识

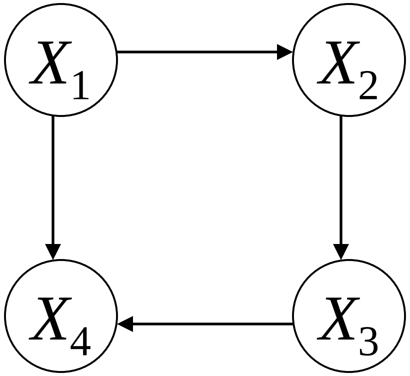
1.3 图模型入门

图论

- A graph is a pair G=(X, E)
  - X = { $X_1,\ldots, X_N$ } is a finite set of vertices, also called nodes, of G
    — 结点集合
  - E is a subset of the set $X \times X$ = $\{(X_i, X_j): i \neq j\}$, called edges of G
    — 边集合, 结点有序对的集合
  - Undirected edge: both (X_i, X_j) and (X_j, X_i) belong to E
    $X_i \sim X_j$
  - Directed edge (arc): $(X_i, X_j) \in E$ and $(X_j, X_i) \notin E$
    $X_i \rightarrow X_j$
    we say that X_i is a parent of X_j, X_j is a child of X_i

图论

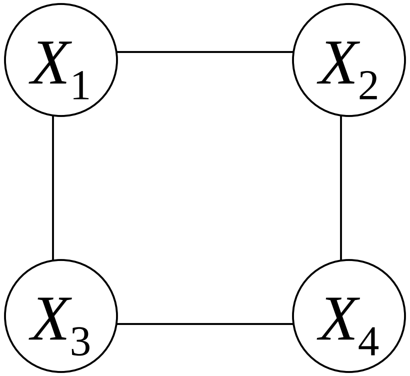
- Directed graph: All edges in the graph are directed



$$X = \{ X_1, X_2 , X_3, X_4 \}$$

$$E = \{\{X_1, X_2\}, \{X_1, X_4\}, \{X_2, X_3\}, \{X_3, X_4\}\}$$

- Undirected graph: All edges in the graph are undirected



$$X = \{ X_1, X_2 , X_3, X_4 \}$$

$$E = \{\{X_1, X_2\}, \{X_2, X_1\}, \{X_1, X_3\}, \{X_3, X_1\}, \{X_2, X_4\}, \{X_4, X_2\}, \{X_3, X_4\}, \{X_4, X_3\}\}$$

第一章 引言

1.1 统计推理和学习的概念

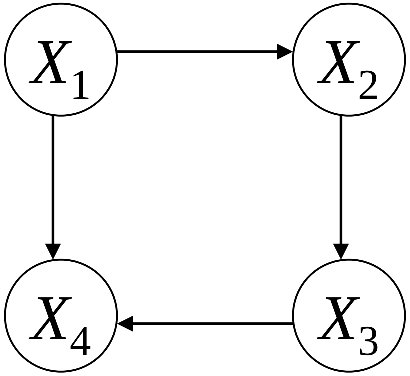
1.2 简单图论知识

1.3 图模型入门

图模型

- 图模型：在图上赋予概率分布得到的概率模型

- Directed graph
  - The nodes represent random variables
  - The edges (parent-child relationship) represent dependence
  - Let $X\pi_i$ represents the set of parents of node X_i



在图上赋予概率分布的过程，就是**图的语义**。

有向图的语义：

$$p(x_1, \cdots, x_N) \triangleq \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

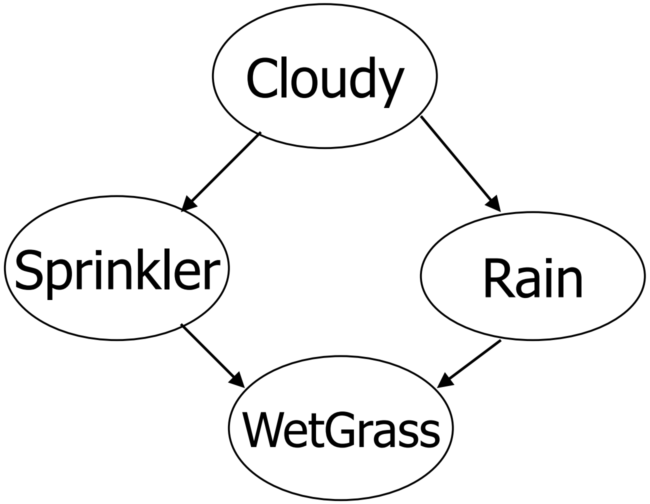
有向有环图上如上定义，一般来讲不是一个合法的概率分布。

Toy Example of a Bayes net

— 分析变量间关系，建立概率模型

$p(C=0)$	$p(C=1)$
0.5	0.5

C	$p(S=0\|C)$	$p(S=1\|C)$
0	0.5	0.5
1	0.9	0.1



C	$p(R=0\|C)$	$p(R=1\|C)$
0	0.8	0.2
1	0.2	0.8

S	R	$p(W=0\|S, R)$	$p(W=1\|S, R)$
0	0	1.0	0.0
1	0	0.1	0.9
0	1	0.1	0.9
1	1	0.01	0.99

Conditional Probability Table (CPT)

Toy Example of a Bayes net

- The joint probability of all the nodes:

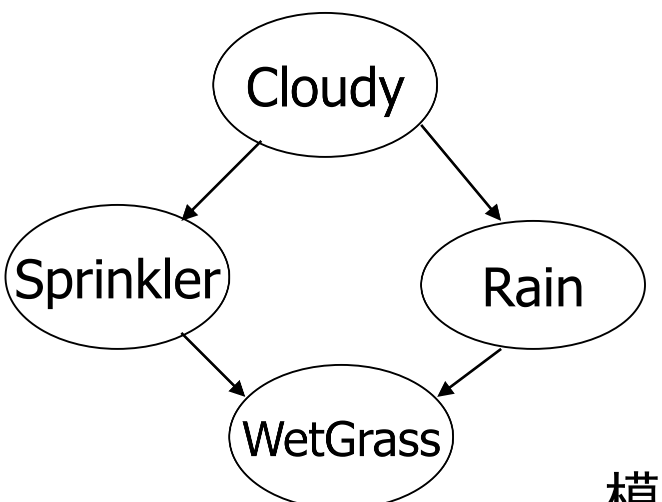
$$p(C, R, S, W) \;=\; p(C) \times p(R \mid C) \times p(S \mid C,R) \times p(W \mid C,R,S)$$

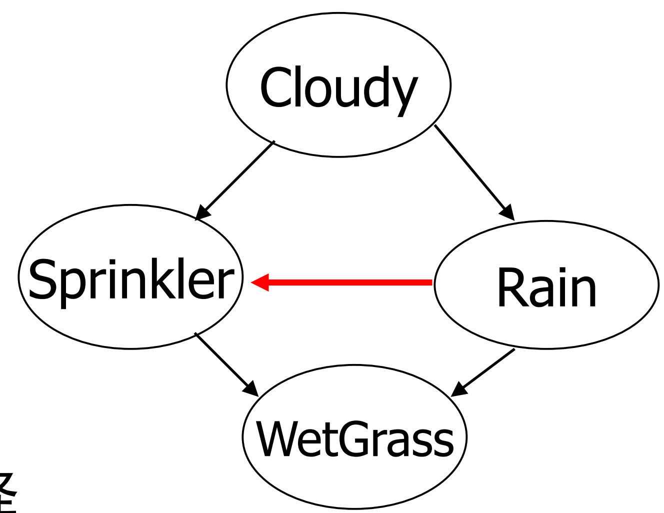
$$p(C, R, S, W) \;\triangleq\; p(C) \times p(R \mid C) \times p(S \mid C) \times p(W \mid R,S)$$

$R \perp S \mid C$

$W \perp C \mid S, R$

- 规定变量的联合分布具有“特定分解形式”：做假设/约束
- 建模就是做假设，求证（假设是否合理，是否恰当）的过程





模型选择

Why Graphical Model？

- 丰富的模型表达能力（Representation）
  - 统一了目前广泛应用的许多统计模型和方法
  - 通过 画图 以建模；通过 读图 以分析变量间关系
- 强大的推理计算能力（Inference）
  - 原理性算法，通用性，一般性
  - 面对具体的新模型，运用原理而不必自己重新设计推理算法

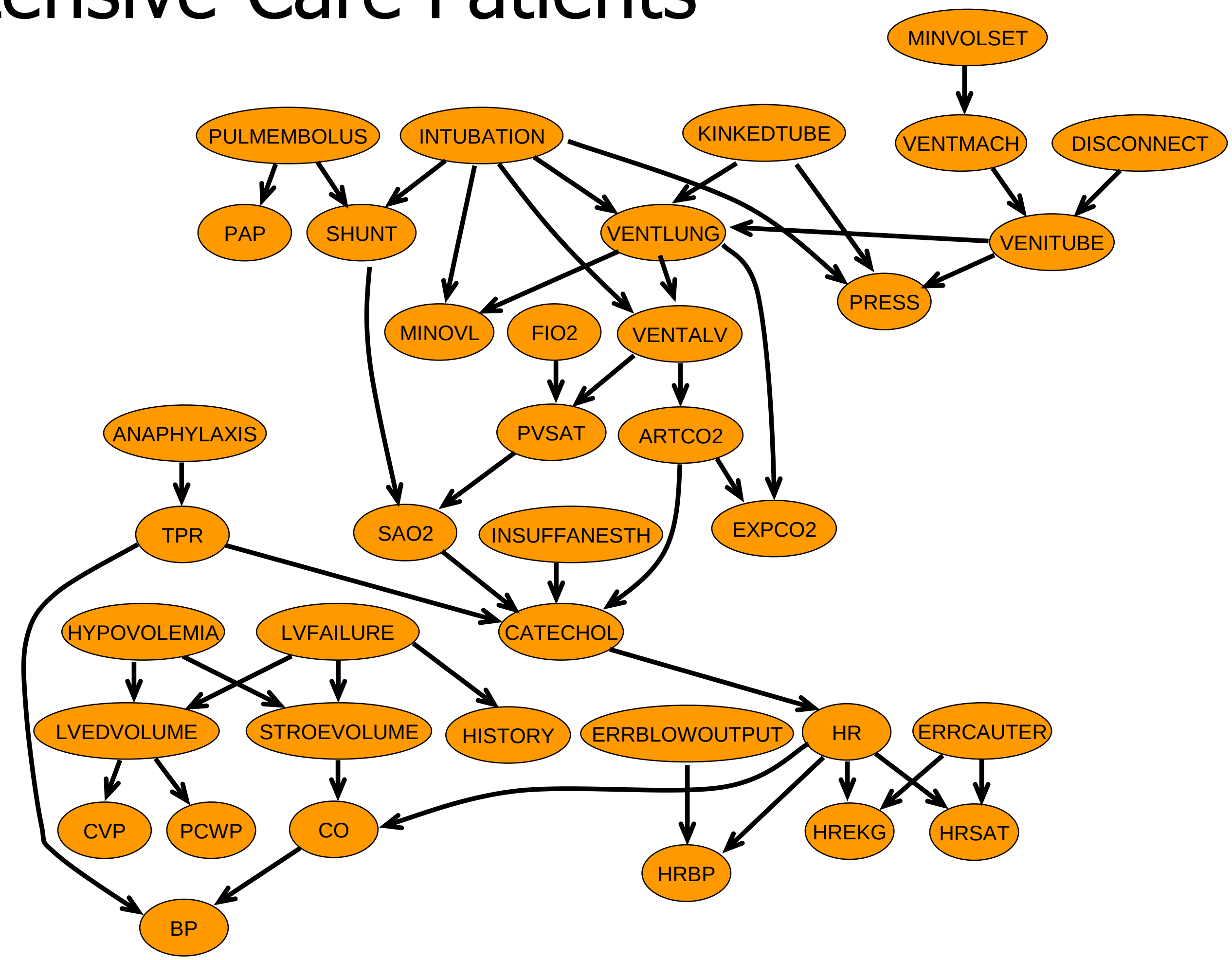
求： $p(S=1|W=1)=?$ $p(R=1|W=1)=?$

- 全面的学习理论（Learning）
  - 结构学习
  - 拍脑袋＋让数据来说话：领域知识与样本信息（数据）有机结合

A real Bayes net

Monitoring Intensive-Care Patients

- 37 variables



第一章 引言

1.1 统计推理和学习的概念

1.2 简单图论知识

1.3 图模型入门

View of a family of distributions 分布族的观点

- 一个图g的分布，一般而言，不是指一个具体的分布，而是指一族分布
  - 穷尽 x_v 的形式、每个结点处的局部条件分布$\{ p(x_v \mid x_{pa(v)}) \}$的可能形式，得到一个分布的集合，记为 M($g$)

$$\mathrm{M}(g) = \left\{ p(x_V) : p(x_V) = \prod_{v \in V} p\left(x_v \mid x_{pa(v)}\right) \right\}$$

$$p(V) = \prod_{v \in V} p\left(v \mid pa(v)\right)$$

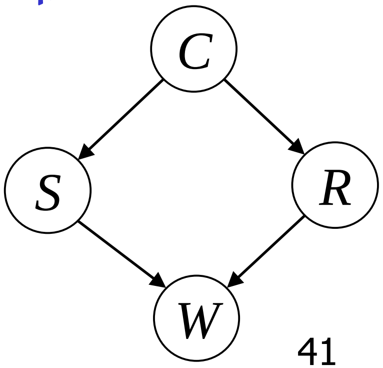
  - **一般讨论：**对这一族分布中的每个分布都成立的结论
  - **从一般到具体，可以有不同抽象层次看待一个图的分布。**

$$p(C,R,S,W) = p(C)\,p(R \mid C)\,p(S \mid C)\,p(W \mid R,S)$$

$$R \perp S \mid C$$

$$W \perp C \mid S, R$$

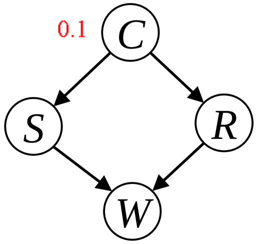
C	$p(S=0 \mid C)$	$p(S=1 \mid C)$
0	0.5	0.5
1	0.9	0.1



图模型建模的四要素

- 当使用图模型去表示一个具体的概率分布时（ to describe a particular distribution），需逐步明确以下四要素
- 语义（Semantics）
  - 定义了图和概率分布如何发生联系（有向图、无向图、因子图、链图、...）
- 结构（Structure）
  - 定义随机变量及之间联系（即明确图包含哪些结点以及边）

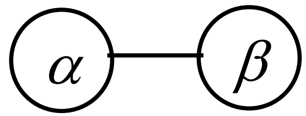
C	$p(S=0 \mid C)$	$p(S=1 \mid C)$
0	0.5	0.5
1	0.9	0.1



- 实现（Implementation）
  - 指定结点类型（ discrete/continuous ）以及局部函数的具体形式
- 参数（Parameter）
  - 凭经验指定或利用数据进行估计——局部函数的待定参数的取值

Graph Theory

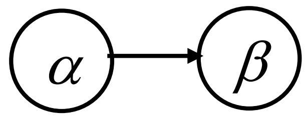
Basic concepts – undirected edge, neighbor



we have an undirected edge between α and β	$\alpha \sim \beta$	If both ordered-pairs (α, β) and (β, α) belong to E
we also say that α and β are neighbors, α is a neighbour of β , or β is a neighbor of α		

the set of neighbors of β 邻居集	$ne(\beta)$	$ne(\beta) = \{\alpha : \alpha \sim \beta\}$

Basic concepts – directed edge, parent, child



we have an directed edge from α to β	$\alpha \rightarrow \beta$	If $(\alpha, \beta) \in E$ but $(\beta, \alpha) \notin E$
we also say that α is a parent of β , and that β is a child of α		

the set of parents of β 父结点集	$pa(\beta)$	$pa(\beta) = \{\alpha : \alpha \rightarrow \beta\}$
the set of children of α 子结点集	$ch(\alpha)$	$ch(\alpha) = \{\beta : \alpha \rightarrow \beta\}$

Basic concepts – un/directed graph

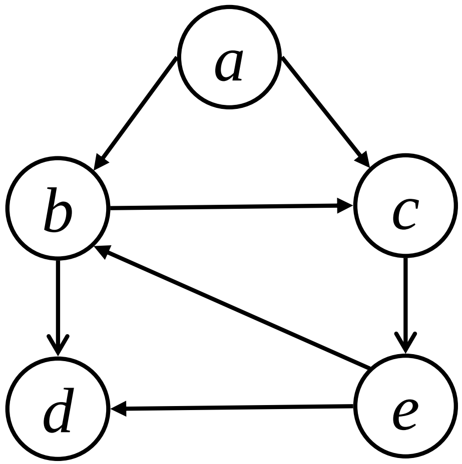
we say D is a directed graph		If *all* the edges of a graph D are directed
we say g is a undirected graph		If *all* the edges of a graph g are undirected

Basic concepts – path, trail

path 路径		从α_0到α_n的一条长为n的路径，是指存在n+1个不同结点构成的序列$\alpha_0,\ldots, \alpha_n$，且 $(\alpha_{i-1}, \alpha_i)\in E$，$i = 1, \ldots, n$.

- 一条路径不会和自己相交；不能逆着有向边的方向而行

we say that α leads to β α 可到达 β	$\alpha \mapsto \beta$	如果存在一条从α 到 β 的路径



Basic concepts – cycle, DAG, trail

n-cycle 环		从α_1到α_n的一条长为n的环，是指存在n个不同结点构成的序列$\alpha_1,\ldots,\alpha_n$，且$(\alpha_i,\alpha_{i+1})\in E$，$i=1,\ldots,n\text{-}1$，$(\alpha_n,\alpha_1)\in E$

acyclic graph		If it does not possess any cycles
directed acyclic graph (DAG)		A directed graph which is acyclic

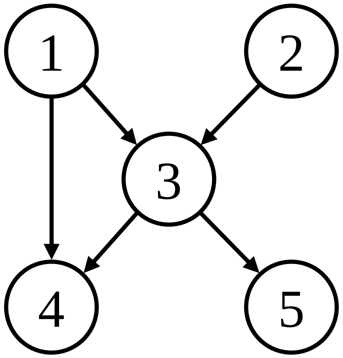
trail from α to β 迹		从α_0到α_n的一条长为n的迹，是指存在n+1个不同结点构成的序列$\alpha_0,\ldots,\alpha_n$，且$\alpha_{i\text{-}1}\rightarrow\alpha_i$, or $\alpha_i\rightarrow\alpha_{i\text{-}1}$, or $\alpha_{i\text{-}1}\sim\alpha_i$，$i=1,\ldots,n$.

- 一条迹不会和自己相交；可以逆着有向边的方向而行

Basic concepts – well-order of DAG

well-order of a DAG (Directed Acyclic Graph) 有向无环图的良序		我们总可以对一个DAG的结点进行排序/编号，使得：图中的有向边总是从低编号的结点指向高编号的结点。这样一种排序称为良序

predecessors of node α	$pr(\alpha)$	在DAG的一个良序中，结点编号比 α 小的结点集合称为 α 的前辈结点集

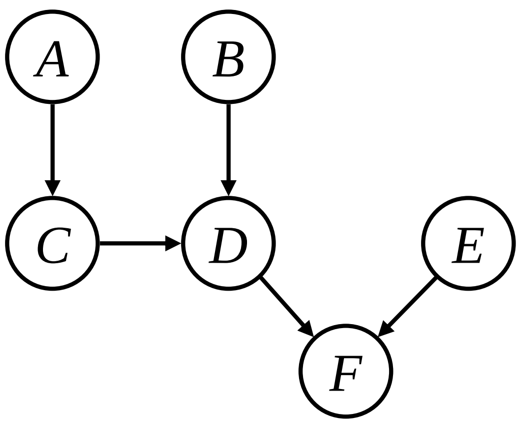


Here is a well-ordering of the nodes in the DAG

Basic concepts – ancestors, non/descendants

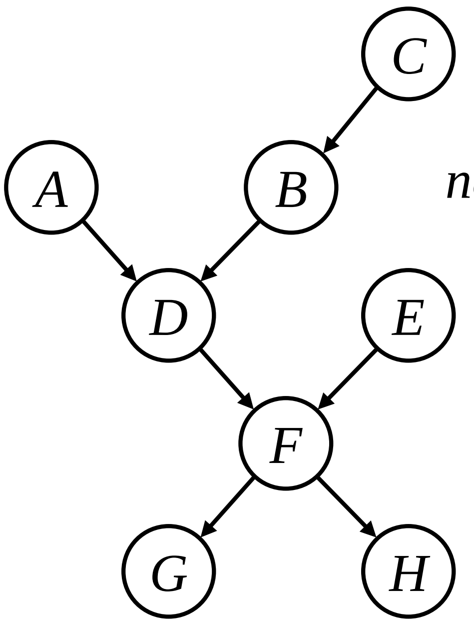
有向图

ancestors (祖先) of node β	$an(\beta)$	$an(\beta) = \{\alpha : \alpha \mapsto \beta \text{ but not } \beta \mapsto \alpha\}$
descendants (后代) of α	$de(\alpha)$	$de(\alpha) = \{\beta : \alpha \mapsto \beta \text{ but not } \beta \mapsto \alpha\}$
non-descendants (非后代) of α	$nd(\alpha)$	$nd(\alpha) = V \setminus (\alpha \cup de(\alpha))$



$an(C) =$

$de(C) =$



$nd(D) =$

Conditional independence (CI)

Definition

- $X \perp Y \mid Z$
  if and only if the following equivalent condition holds
  - $p(x \mid y, z) \equiv p(x \mid z)$ if $p(y, z) > 0$
  - $p(x \mid y, z)$ has the form $a(x, z)$ if $p(y, z) > 0$
  - $p(x, y \mid z) \equiv p(x \mid z)\, p(y \mid z)$ if $p(z) > 0$
  - $p(x, y \mid z)$ has the form $a(x, z)\, b(y, z)$ if $p(z) > 0$
  - $p(x, y, z) \equiv p(x \mid z)\, p(y \mid z)\, p(z)$
  - $p(x, y, z) \equiv p(x, z)\, p(y, z) / p(z)$ if $p(z) > 0$
  - $p(x, y, z)$ has the form $a(x, z)\, b(y, z)$

Properties of CI—条件独立性的运算法则

- The ternary relation $X \perp Y \mid Z$ has the following properties:
  - P1: Symmetry
    If $X \perp Y \mid Z$ then $Y \perp X \mid Z$
  - P2: Decomposition
    If $X \perp Y \mid Z$ and U is a function of X then $U \perp Y \mid Z$
  - P3: Weak Union
    If $X \perp Y \mid Z$ and U is a function of X then $X \perp Y \mid (Z, U)$
  - P4: Contraction
    If $X \perp Y \mid W$ and $X \perp Z \mid (W, Y)$ then $X \perp (Y, Z) \mid W$
  - P5: Intersection (hold under certain conditions)
    If $X \perp Y \mid (Z, W)$ and $X \perp Z \mid (Y, W)$ then $X \perp (Y, Z) \mid W$

Properties of CI - Example of P5 fails

- $X \perp Y \mid Z$ and $X \perp Z \mid Y \Rightarrow ?\ X \perp (Y, Z)$
  - Define: $p(X{=}Y{=}Z{=}1) = p(X{=}Y{=}Z{=}0) = 0.5$

    Other configurations of X, Y and Z have 0 probability.
  - $p(x \mid z) = \dfrac{p(x,z)}{p(z)} = \dfrac{0.5 \times \delta(x=z)}{0.5} = \delta(x=z)$ $\quad p(y \mid z) = \delta(y=z)$

    $$p(x,y \mid z) = \frac{p(x,y,z)}{p(z)} = \frac{0.5 \times \delta(x=y=z)}{0.5} = \delta(x=y=z)$$

    $$= \delta(x=z)\delta(y=z) = p(x \mid z)\,p(y \mid z)$$
  - This gives $X \perp Y \mid Z$. Similarly, $X \perp Z \mid Y$.
  - $p(x \mid y=1, z=1) = \dfrac{p(x, y=1, z=1)}{p(y=1, z=1)} = \dfrac{0.5 \times \delta(x=1)}{0.5} = \delta(x=1)$

    $$p(x=0) = p(x=1) = 0.5$$
  - $X \perp (Y, Z)$ is not true. Given (Y, Z), we know $X{=}Y{=}Z$.

Properties of CI - P5

- P5 is true when the distribution p is positive.

  If $\begin{matrix} X \perp Y \mid (Z,W) \\ X \perp Z \mid (Y,W) \end{matrix}$ and $p(x,y,z,w) > 0, \forall x,y,z,w$

  then $X \perp (Y,Z) \mid W$

- Proof: $p(x,y,z,w) = g(x,z,w)\,h(y,z,w)$

$$= k(x,y,w)\,l(y,z,w)$$

for suitable strictly positive functions g, h, k, l .

Thus, for all z we must have $k(x,y,w) = \dfrac{g(x,z,w)\,h(y,z,w)}{l(y,z,w)}$

Hence, choosing a fixed $z = z_0$, we have

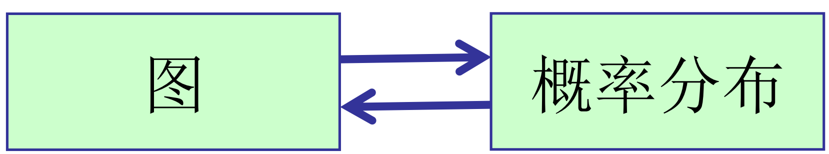
$$k(x,y,w) = g(x,z_0,w)\frac{h(y,z_0,w)}{l(y,z_0,w)} = \pi(x,w)\,\rho(y,w)$$

Thus, $p(x,y,z,w) = \pi(x,w)\,\rho(y,w) \times l(y,z,w) \Rightarrow X \perp (Y,Z) \mid W$

第二章 图模型的表示理论

语 义

一个图表示了怎样的概率分布



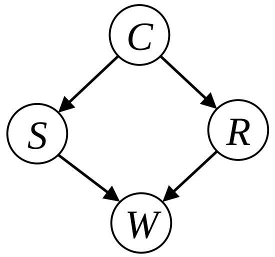
一个概率分布如何表示成一个图

Definition - (DF) property

赋予概率分布一新表述：先定义性质，性质诱导分布

- 称一个分布 $p(x_V)$ 服从 依图D的有向分解性，如果

$$p\left(x_V\right)=\prod_{v\in V} p\left(x_v \mid x_{pa(v)}\right)$$



由依图D的有向分解性诱导了一族分布：$\mathrm{M}_{DF}(D)$

$$\mathrm{M}_{DF}\left(D\right)=\left\{p(x_V):p(x_V)=\prod_{v\in V} p\left(x_v \mid x_{pa(v)}\right)\right\}$$

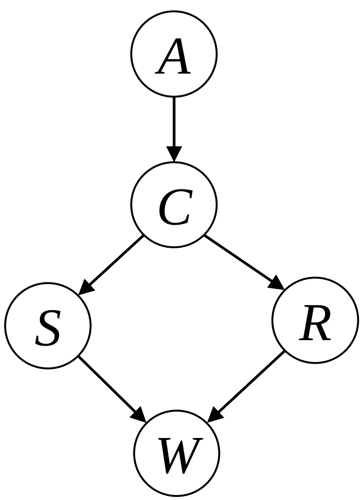
Definition - (DO) property

称一个分布 $p(x_V)$ 服从依图 D 的有向有序Markov性，如果在图D的**某个**良序下，成立

in some given well-ordering of the nodes (CDLS p73)

$$v \perp pr(v) \setminus pa(v) \mid pa(v)$$

由依图D的有向有序Markov性诱导了一族分布：$M_{DO}(D)$



$p(A, C, S, R, W)$

在**某个**良序下，成立 $v \perp pr(v) \setminus pa(v) \mid pa(v)$

⇓ 对图的结点数运用数学归纳法

$$p(V) = \prod_{v \in V} p(v \mid pa(v))$$

⇓

在**任意**良序下，成立 $v \perp pr(v) \setminus pa(v) \mid pa(v)$

不同良序下写出的CI是等价的，统称为DO

View of a family of distributions

依图D的 有向分解性	⇄	依图D的 有向有序Markov性
$p(V)=\prod_{v\in V} p(v \mid pa(v))$		$v \perp pr(v) \setminus pa(v) \mid pa(v)$

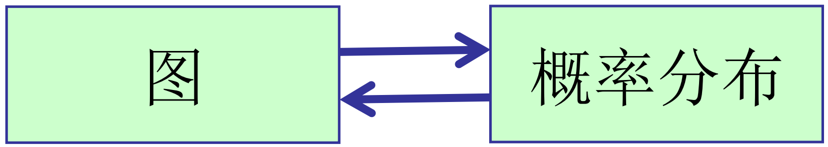
- Theorem: $M_{DF}(D)$ = $M_{DO}(D)$，合记为$M(D)$

At the core of the graphical model representation

- 在建模和算法分析中，灵活地解读一个概率图

DAG的语义

一个图表示了怎样的概率分布

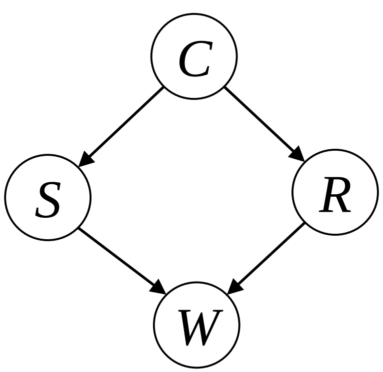


一个概率分布如何表示成一个图

一个DAG表示一个怎样的概率分布（by definition）

- 满足有向分解性
- 满足有向有序Markov性

一个概率分布如何表示成一个**DAG**



$$p(C,R,S,W)=p(C)\,p(R\,|\,C)\,p(S\,|\,C)\,p(W\,|\,R,S)$$

I-map (independency map)

- 将一个分布 $p(.)$ 所蕴含/满足/表示的 CI 记为 $\triangleq I(p)$

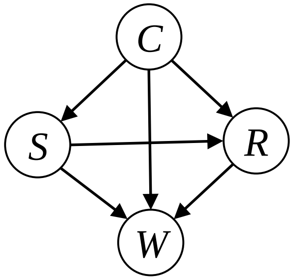
$$p(C,R,S,W)=p(C)\,p(R|C)\,p(S|C)\,p(W|R,S) \qquad \left\{\begin{array}{l} R\perp S\,|\,C \\ W\perp C\,|\,S,R \\ \cdots \end{array}\right\}$$

称一个BN D 是一个分布 p 的 I-map，如果

$$I(M(D))\subseteq I(p)$$

这时，我们也称分布 p 可以表示成 图D 。

- 相对分布 p 所表示的条件独立性，图 D 表示了较少的条件独立性。
- 分布 p 可以有一些条件独立性没有反映在图 D 中。
- 一个完备的有向无环图 是 任何分布的 I-map。



称一个BN D 是一个分布 p 的 完美贝叶斯网络表示 (P-map)，如果

$$I(M(D))=I(p)$$

P-map的存在性

- 不是每一个分布都有 完美贝叶斯网络表示。

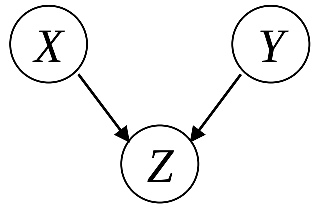
例子（KF p.81 example 3.6）

- 考虑三个布尔随机变量X, Y, Z的如下联合分布

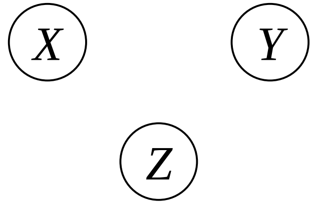
$$p(x,y,z)=\begin{cases}1/12 & x\oplus y\oplus z=0\\ 1/6 & x\oplus y\oplus z=1\end{cases}$$

x	y	$x\oplus y$
0	0	0
0	1	1
1	0	1
1	1	0

$$X \perp Y \in I(p)$$
$$X \perp Z \in I(p)$$
$$Y \perp Z \in I(p)$$
$$X \perp Y | Z \notin I(p)$$
$$XY \perp Z \notin I(p)$$

$$\mathrm{I}\left(\right.$$


$$\left.\right) \overset{?}{=} I(p)$$

$$\mathrm{I}\left(\right.$$


$$\left.\right) \overset{?}{=} I(p)$$

课程章节

- 第一章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第二章 图模型的推理理论（4）
  - 精确推理：**variable-elimination，cluster-tree，triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第三章 图模型的学习理论（2）
  - 参数学习：**maxlikelihoodEstimate，RFLearning, BayesEstimate**
  - 结构学习：**StructureLearning**

	pgm-1 semantics √		pgm-2 hmm-crf pgm-3 exact	pgm-4 kalman pgm-5 sampling
pgm-6 variational pgm-7 ML	pgm-8 Bayesian			