概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(32学时, 2学分, 2004 – 2012, 2017)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

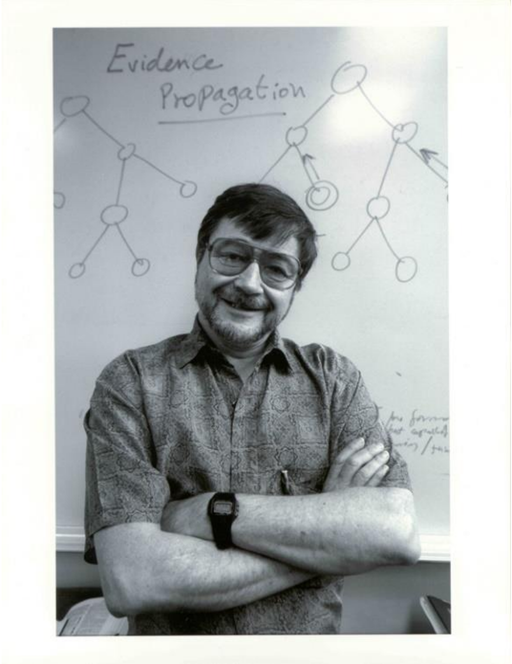
Tel: 62796193

Email: ozj@tsinghua.edu.cn

引言

- 概率建模，推理和学习
  - 机器智能
  - 处理不确定性是智能的一种重要表现
    - 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
  - 不确定性是客观世界中的一种真实、广泛的存在



- (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）

  表示理论

  - 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）
- 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 引言（1）
- 第二章 图模型的表示理论（3）
  - DGM-UGM
  - Semantics
  - HMM, CRF
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

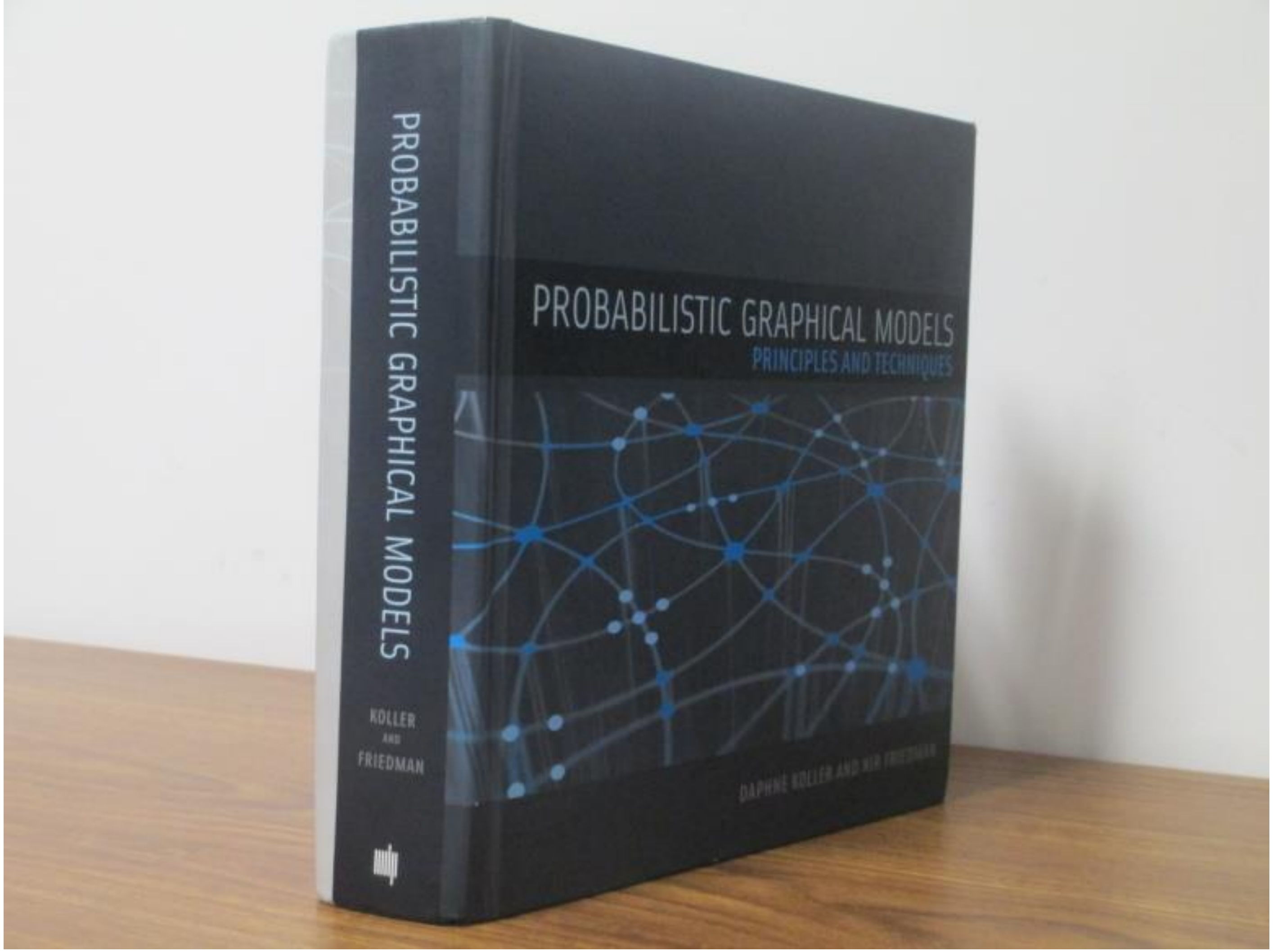
考核方式

- 课后作业 (4次): 20%
  - 掌握基本概念和原理
  - 按时交至网络学堂，迟交分数打九折。
- 阅读及笔记（scribe）一次: 20%
  - 13节课，课前微信群提交阅读摘要（1页）：总结，提出问题；
  - 课堂笔记；
  - 课后扩展阅读和调研，综合你的阅读和课堂笔记，形成一节资料，当次课一周后（周日23:59）email交初稿，两周后（周日23:59） email交终稿，Latex模板。
  - 若成组(<=3人)，按组给分，说明分工
- 大作业 (自行选题的project): 60%
  - Applying 应用解决实际问题
  - Theoretic 偏理论探索
  - 第17周报告
  - 若成组(<=2人)，按组给分，说明分工

Tools available

- Bayes Net Toolbox (BNT) for Matlab
  - http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html
- PNL (Probabilistic Network Library): A C++ version of BNT
  - http://sourceforge.net/projects/openpnl
- PMTK toolkit : primarily designed to accompany Kevin Murphy's textbook “Machine learning: a probabilistic perspective”
  - https://github.com/probml/pmtk3

参考书


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

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 toolkit

Murphy书
Comparison to other books on the market

- My book (MLaPP) is similar to Bishop's Pattern recognition and machine learning, Hastie et al's The Elements of Statistical Learning, and to Wasserman's All of statistics, with the following key differences:
  - MLaPP is more accessible to undergrads. It pre-supposes a background in probability, linear algebra, calculus, and programming; however, the mathematical level ramps up slowly, with more difficult sections clearly denoted as such. This makes the book suitable for both undergrads and grads. Summaries of the relevant mathematical background, on topics such as linear algebra, optimization and classical statistics make the book self-contained.
  - MLaPP is more practically-oriented. In particular, it comes with Matlab software to reproduce almost every figure, and to implement almost every algorithm, discussed in the book. It includes many worked examples of the methods applied to real data, with readable source code online.
  - MLaPP covers various important topics that are not discussed in these other books, such as conditional random fields, deep learning, etc.
  - MLaPP is "more Bayesian" than the Hastie or Wasserman books, but "more frequentist" than the Bishop book. In particular, in MLaPP, we make extensive use of MAP estimation, which we regard as "poor man's Bayes". We prefer this to the regularization interpretation of MAP, because then all the methods in the book (except cross validation...) can be viewed as probabilistic inference, or some approximation thereof. The MAP interpretation also allows for an easy "upgrade path" to more accurate methods of approximate Bayesian inference, such as empirical Bayes, variational Bayes, MCMC, SMC, etc.
  - The emphasis is on simple parametric models (linear and logistic regression, discriminant analysis/ naive Bayes, mixture models, factor analysis, graphical models, etc.), which are the ones most often used in practice. However, we also briefly discuss non-parametric models, such as Gaussian processes, Dirichlet processes, SVMs, RVMs, etc.
  - OZJ: More on UGMs than Bishop book.

有关图模型应用的期刊专辑

- Special Issue on New Computational Paradigms for Acoustic Modeling in Speech Recognition.
  - Computer Speech & Language, Volume: 17, Issue: 2-3, April - July 2003.
- Special Issue on Graphical Models in Computer Vision.
  - IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 25, Issue: 7, July 2003.
- Special Issue On Codes On Graphs And Iterative Algorithms.
  - IEEE Transactions on Information Theory, Volume: 47, Issue: 2, Feb 2001.
- Special issue: Probabilistic models of cognition
  - Trends in Cognitive Sciences, Volume 10, Issue 7, July 2006.
- …

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

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.)
  - 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)

$$p(x) \triangleq p\left(X = x\right)$$

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

$$p(x_1,\ldots,x_n) \triangleq p\left(X_1 = x_1,\ldots,X_n = x_n\right)$$

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)

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

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)
- 线性代数
  - 向量
  - 矩阵
  - 行列式
  - 矩阵的特征向量，特征值

第一章 引言

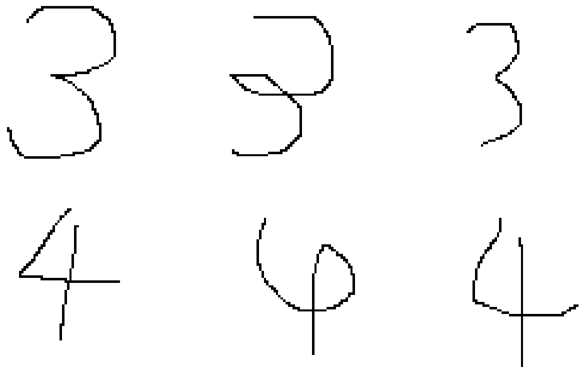
1.1 统计推理和学习的概念 { 模式识别 信号估计

1.2 简单图论知识

1.3 图模型入门

模式识别

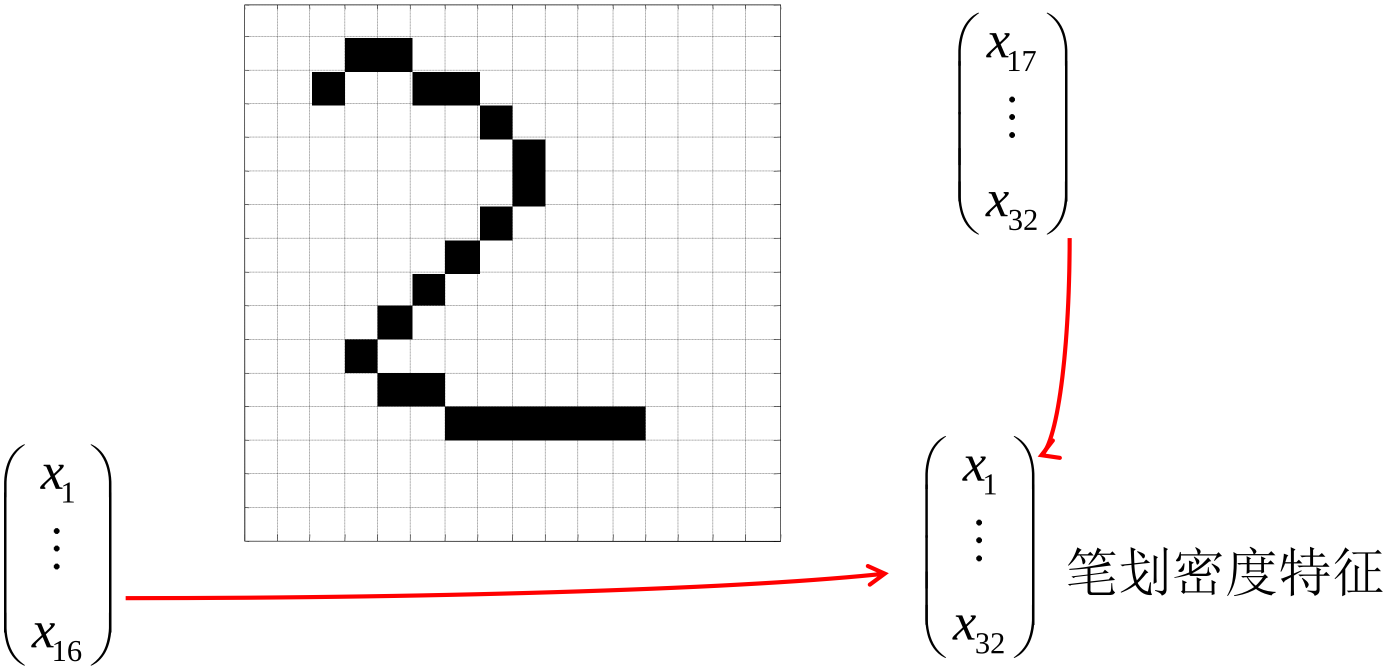
- 人们为了认识客观事物，把**事物**按相似的程度组成不同的**类别**



- 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)



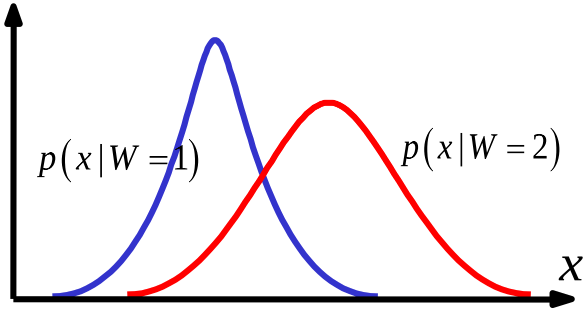
笔划密度特征

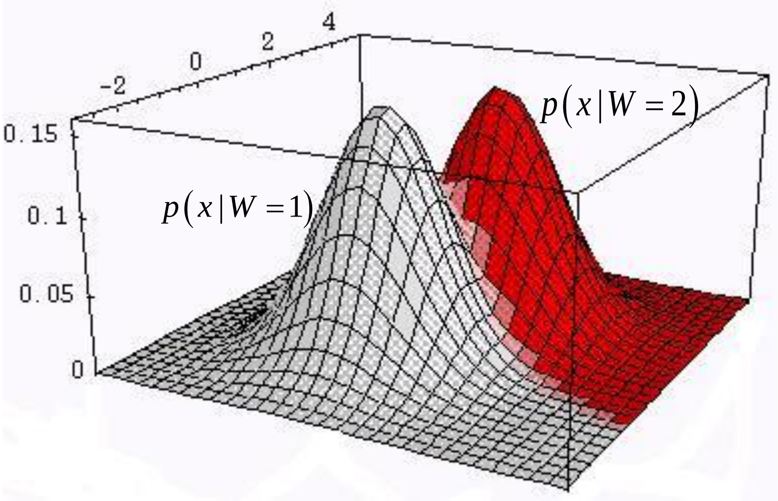
模式识别的概率模型

- 设有概率模型 $p(W, X) = p(W)p(X|W)$

$p(W)$ 类先验概率（e.g. 字频）

$p(X|W)$ 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：$\min_{d} P_e(d)=P(d(X)\neq W)$
- 解: $$d(x)=\arg\max_{k=1,\cdots,K} p(W=k\,|\,X=x)$$

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 $d(x)$

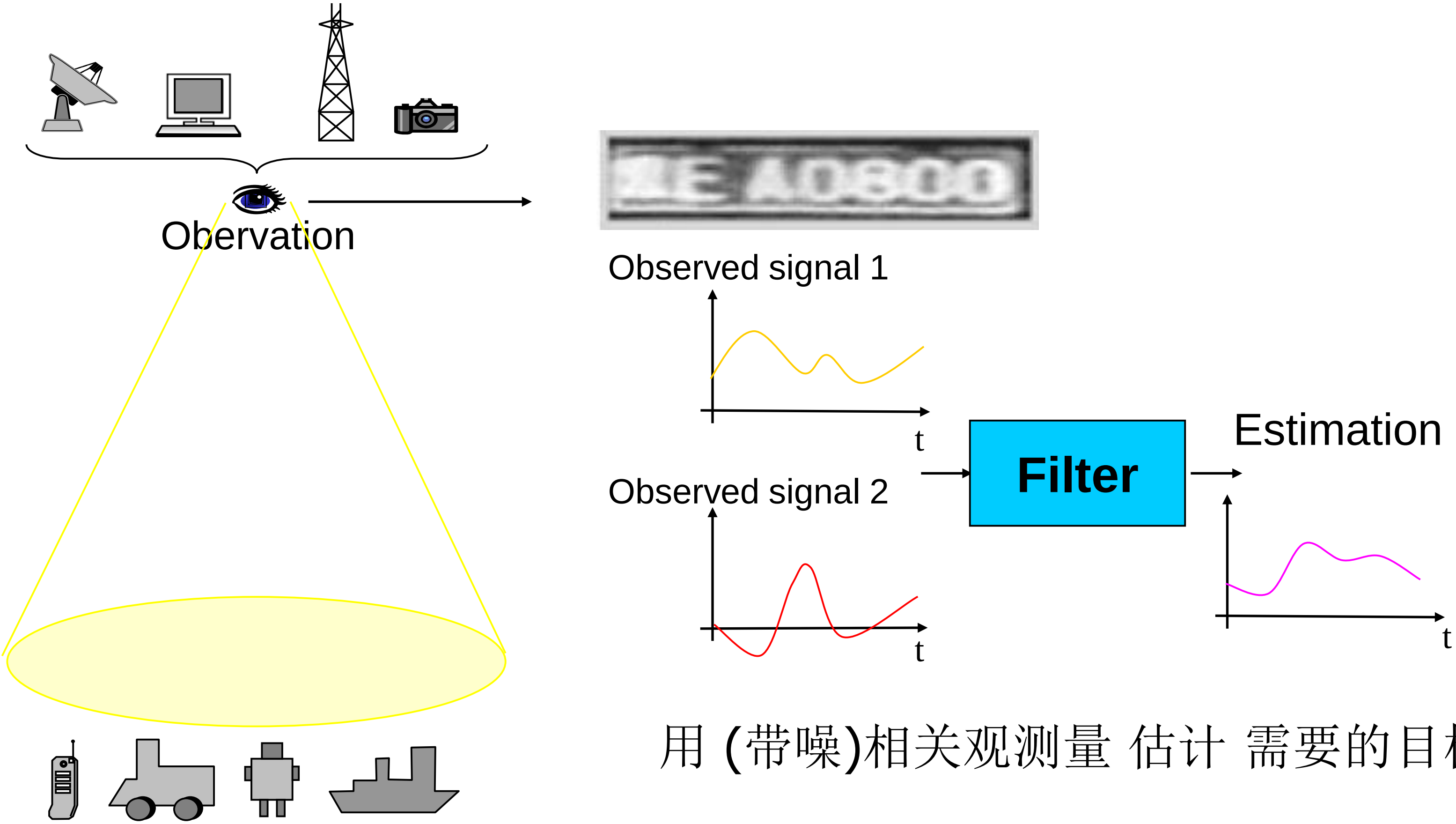
最大后验（MAP，Maximum A Posteriori）判决

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



Obervation

Observed signal 1

Observed signal 2

Filter

Estimation

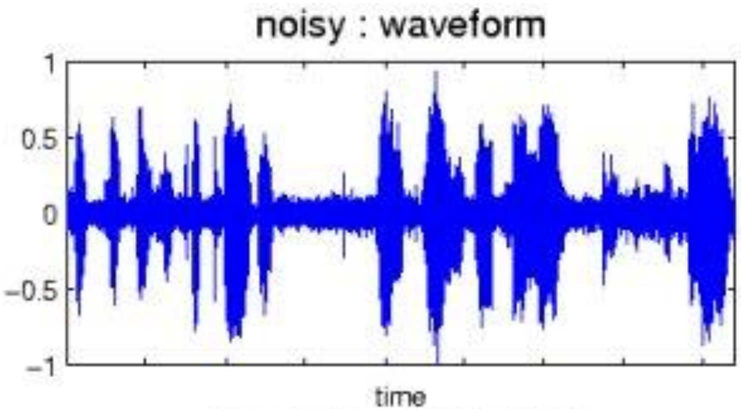
用 (带噪)相关观测量 估计 需要的目标量

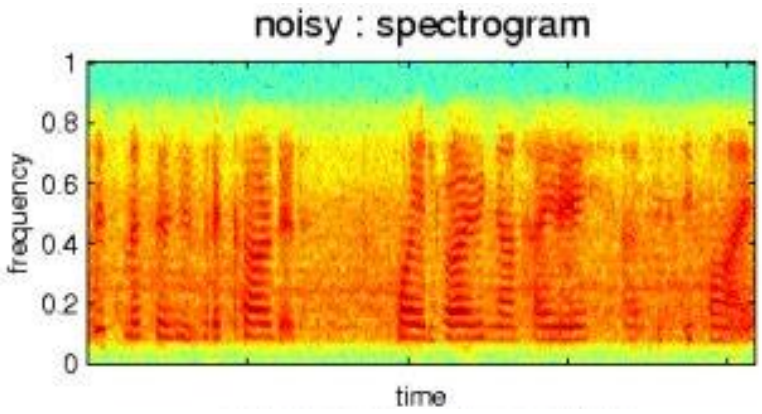
Filter(滤波器)－名词解释

- Any of various electronic devices used to reject signals of certain frequencies while passing others.
  - 一种电子学设备，用于限制特定频率的信号，同时通过其它信号。
  - 《美国传统辞典》
  - Classic filter: lowpass, highpass, or bandpass filter
- A device or program that separates data, signals or material in accordance with specified criteria.
  - 一种将数据、信号或材料 按规定的标准 进行分离的装置或程序。
  - 《现代英汉词典》
  - 概率意义上最好的分离/提取/估计

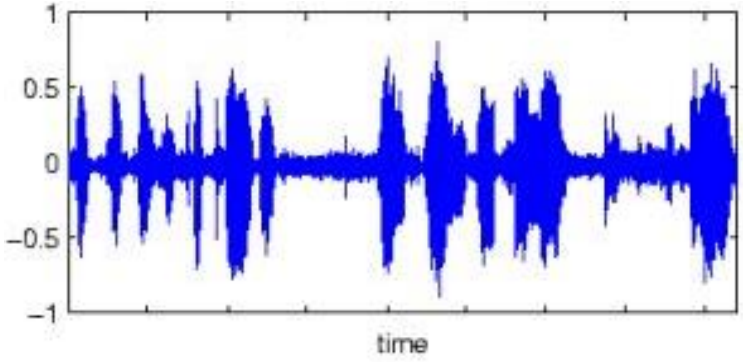
语音增强—演示

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

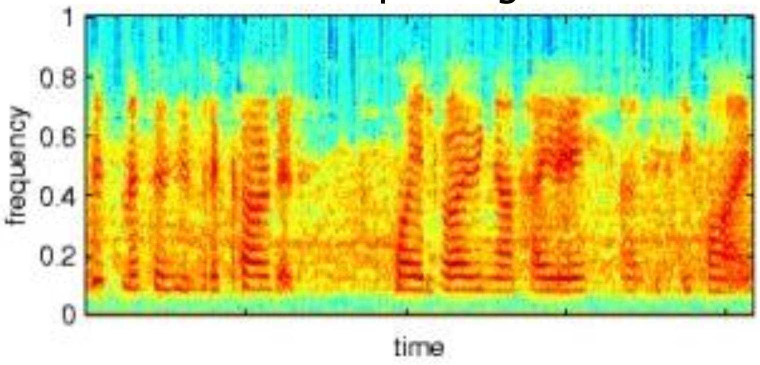




filtered: waveform



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)

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

- 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 $\min_{d} P\left(d\left(X\right)\neq W\right)$	最小均方误差 $\min_{g} E\left[\left\Vert Y-g\left(X\right)\right\Vert^{2}\right]$
解	$d\left(x\right)=\arg\max_{k=1,\cdots,K} P\left(W=k \mid X=x\right)$	$g(x) = E[Z \mid X=x]$

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

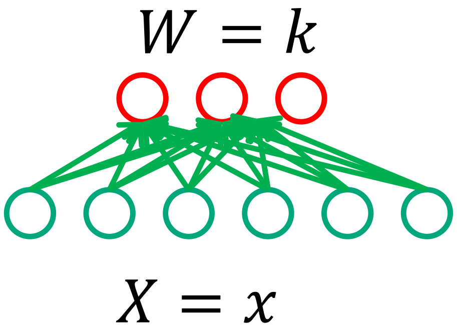
Discriminative model example

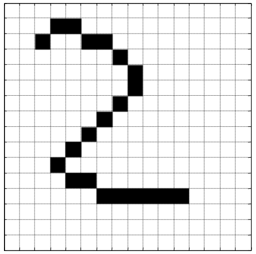
- Multi-class logistic regression / maxent classifier

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

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

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



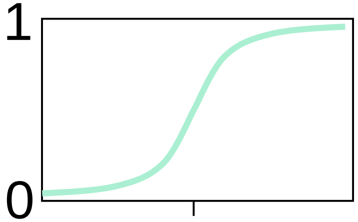


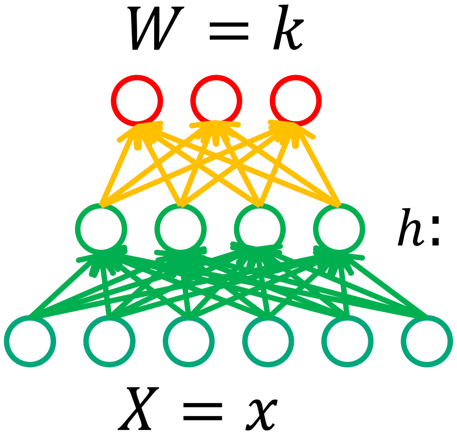
$$\begin{pmatrix} x_1 \\ \vdots \\ x_{32} \end{pmatrix}$$ 笔划密度特征

Discriminative model example

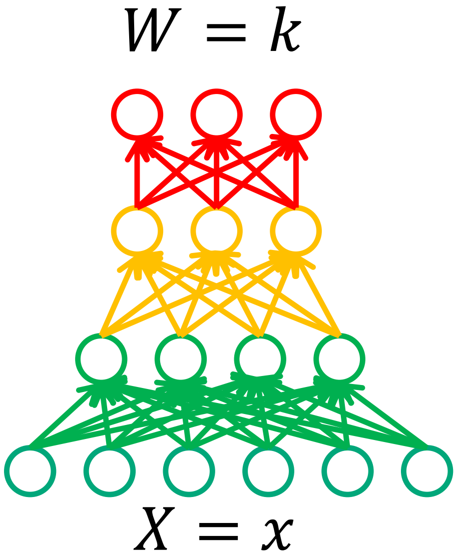
- Neural Networks

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





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

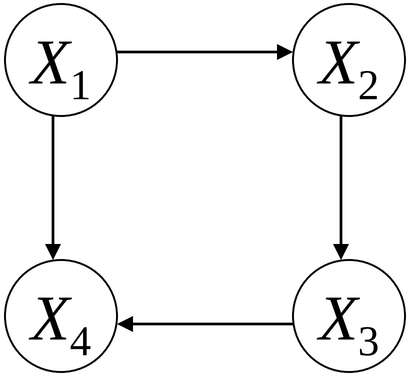
    $X_i \sim X_j$
  - Directed edge (arc): $(X_i, X_j) \in E$ and $(X_j, X_i) \notin E$

    $X_i \rightarrow X_j$

    we say that X_i is a parent of X_j, X_j is a child of X_i

图论

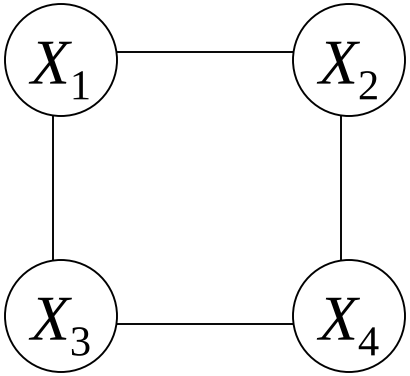
- Directed graph: All edges in the graph are directed



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

$$\begin{aligned} E = \{&\{X_1, X_2\}, \\ &\{X_1, X_4\}, \\ &\{X_2, X_3\}, \\ &\{X_3, X_4\}\} \end{aligned}$$

- Undirected graph: All edges in the graph are undirected



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

$$\begin{aligned} E = \{&\{X_1, X_2\}, \{X_2, X_1\}, \\ &\{X_1, X_3\}, \{X_3, X_1\}, \\ &\{X_2, X_4\}, \{X_4, X_2\}, \\ &\{X_3, X_4\}, \{X_4, X_3\}\} \end{aligned}$$

第一章 引言

1.1 统计推理和学习的概念

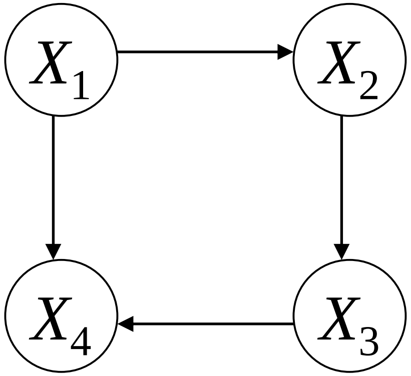
1.2 简单图论知识

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：

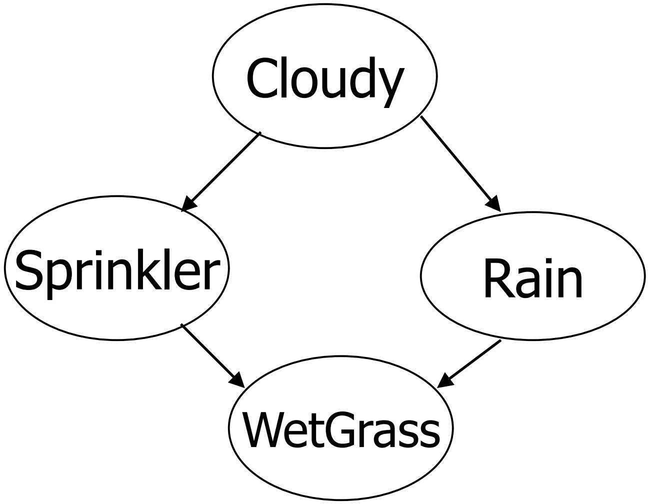
$$p(x_1,\cdots,x_N) \triangleq \prod_{i=1}^{N} p\left(x_i | x_{\pi_i}\right)$$

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:

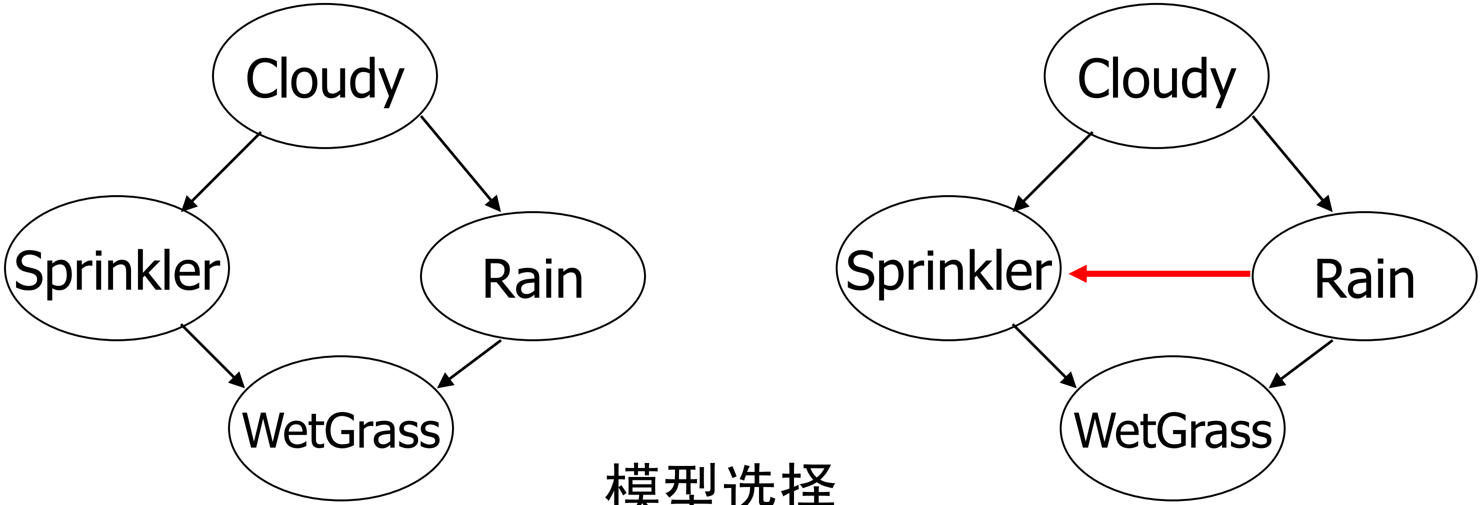
$R \perp S \mid C$

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

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

$W \perp C \mid S, R$

- 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

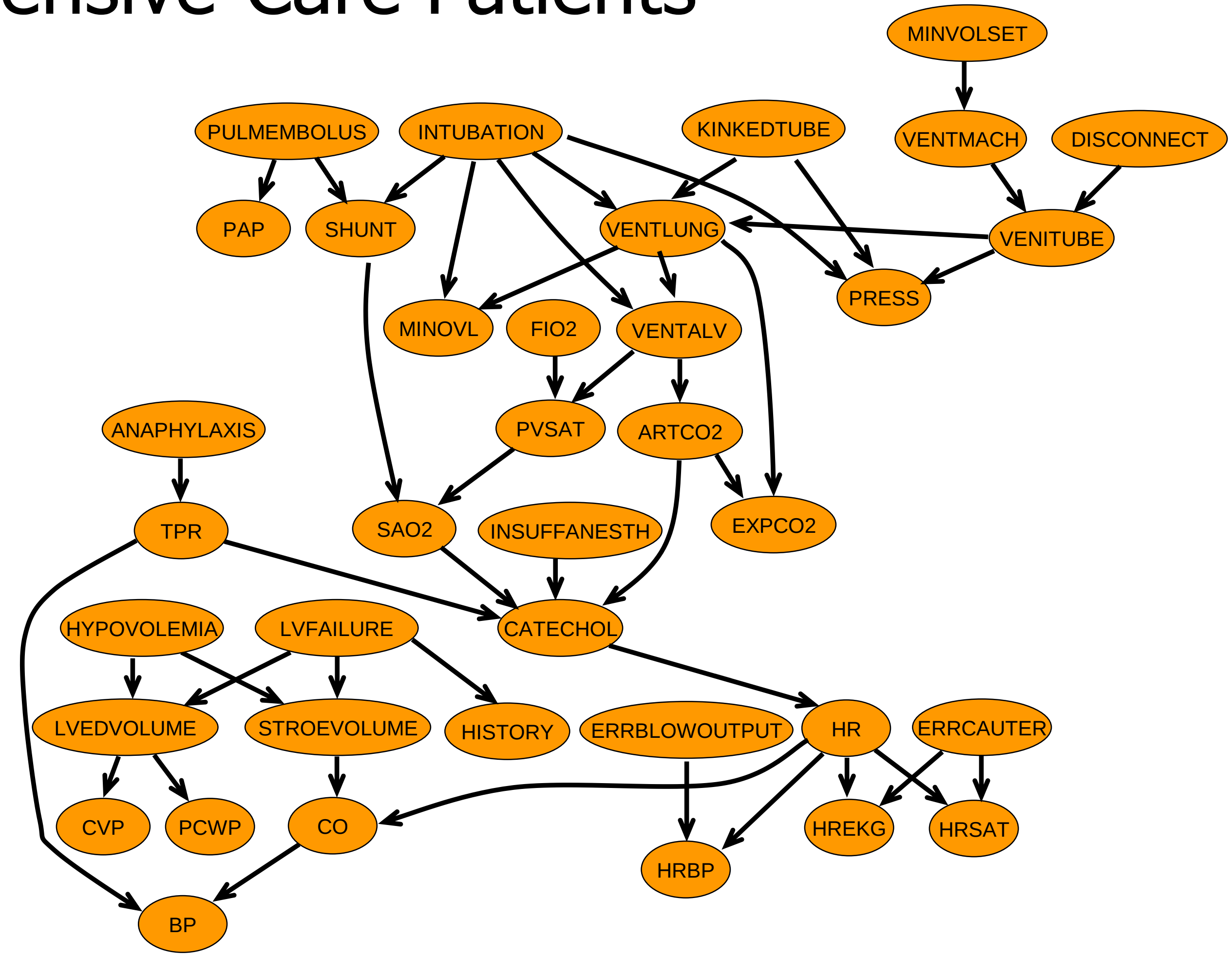
求： $p(S=1|W=1)=?$ $p(R=1|W=1)=?$

- 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