概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 8 – Bayesian Estimate)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

Tel: 62796193

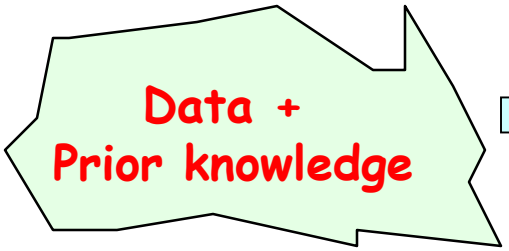
Email: ozj@tsinghua.edu.cn

课程章节

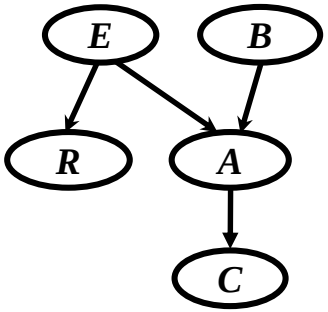
- 第一章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第二章 图模型的推理理论（4）
  - 精确推理：**variable-elimination，cluster-tree，triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第三章 图模型的学习理论（2）
  - 参数学习：**maxlikelihoodEstimate，RFLearning, BayesEstimate**
  - 结构学习：**StructureLearning**

	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			

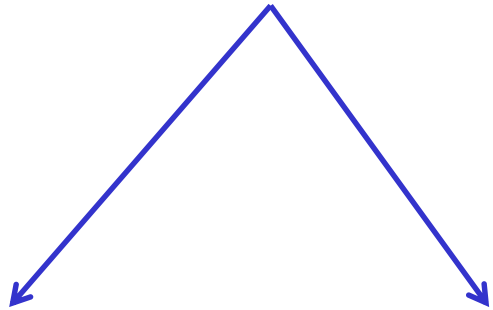
Learning



$p(X \mid g, \theta)$



E	B	$p(A \mid E, B)$	
e	b	**0.9**	**0.1**
e	$\bar{b}$	**0.7**	**0.3**
$\bar{e}$	b	**0.8**	**0.2**
$\bar{e}$	$\bar{b}$	**0.99**	**0.01**



Structure learning /model selection:

Learning g from data

Parameter learning /parameter estimation:

learning θ from data

The learning problem

	Known structure		Unknown structure
Complete data	ML	Bayesian	
Incomplete data	ML	Bayesian	

Why Bayesian ?

- 假设抛一块硬币5次，数字朝上(x=0) 3次
  - 最大似然参数估计 $p(x=0) = 3/5$
  - 这不是一个太好的估计
  - 考虑先验知识——参数 θ 的先验分布 $p(\theta)$
    数字朝上的概率应该近似等于0.5
  - 数据稀疏时避免过拟合
  - 贝叶斯方法在结构学习中具有独特的优势
  - Provide uncertainty measure

Parameter learning

— Bayesian (Known structure, complete data)

对单个分布的参数进行估计？

对一个贝叶斯网络的全体参数进行估计？

参数估计的贝叶斯方法

- 给定一个概率分布函数的参数表达式（parametric form）

$$p(x \mid \theta)$$

从独立同分布样本集 $D = (x[1], \ldots, x[M])$ 中估计出参数 θ？

- 将 θ 视为一个随机变量

$$p(\theta \mid D) \propto \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

- 采取某种准则，从后验分布 $p(\theta \mid D)$ 出发得到对 θ 的点估计 $\hat{\theta}$

$$\hat{\theta}^{MMSE} = \arg\min_{\hat{\theta}} E\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \int \theta\, p(\theta \mid D)\, d\theta = E(\theta \mid D)$$

最小均方误差下 贝叶斯估计： 后验均值

$$\hat{\theta}^{MAP} = \arg\max_{\hat{\theta}} E\left[\delta(\theta - \hat{\theta})\right] = \arg\max_{\hat{\theta}} p(\hat{\theta} \mid D) = \arg\max_{\hat{\theta}} p(D \mid \hat{\theta})\, p(\hat{\theta})$$

二值相似度下 贝叶斯估计： 最大后验估计

Fully Bayesian

Data: $\mathcal{D}= (x[1],\ldots, x[M])$

Prior $P(\theta)$

Posterior $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$

Prediction $P\left(x^{(t)}\middle|\mathcal{D}\right) = \mathrm{E}_{P(\theta|\mathcal{D})}\left[P\left(x^{(t)}\middle|\theta\right)\right] = \int_\theta P\left(x^{(t)}\middle|,\theta\right)P(\theta|\mathcal{D})\mathrm{d}\theta$

$P(\theta|\mathcal{D})$ can be approximated via Markov Chain Monte Carlo methods.
Prediction can be approximated by Monte Carlo averaging:

$$P\left(x^{(t)}\middle|,\mathcal{D}\right) \approx \frac{1}{n}\sum_{i=1}^{n} P\left(x^{(t)}\middle|,\theta_i\right) \quad ,\theta_i \sim P(\theta|\mathcal{D})$$

Multinomial distribution 多元分布

- $x \in \{1, 2, \ldots, K\}$ is discrete r.v.
- $\theta_k = p(x=k)$, $1 \le k \le K$, is the parameters, $\theta = \{\theta_k \mid 1 \le k \le K\}$
- 观测到独立同分布样本集 $D = (x[1], \ldots, x[M])$
- 希望估计 θ ？

似然函数 $p(x[1:M] \mid \theta) = \prod_{m=1}^{M} p(x[m] \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$

N_k : 在样本集中 $x[m]=k$ 出现的次数

考虑参数θ服从Dirichlet分布

$$p(\theta) = \frac{1}{Z(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$Z(\alpha)$ 是归一化常数， $\alpha=(\alpha_1, \ldots, \alpha_K)$ 称为hyperparameters

Dirichlet分布 $p(\theta) = \frac{1}{Z(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$

- $Z(\alpha)$ 是归一化常数

$$Z(\alpha) = \int_{\theta_1} \cdots \int_{\theta_K} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} d\theta_1 \cdots d\theta_K = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K)}$$

  - $\Gamma(\alpha)$ is gamma function: $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha - 1} e^{-t} dt$
  - For integers, $\Gamma(n+1) = n!$

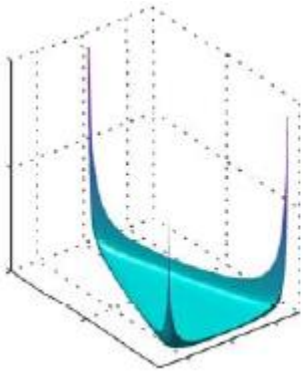
- 如果 $\theta = (\theta_1, \ldots, \theta_K) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_K)$

$$E[\theta_k] = \int \theta_k \cdot p(\theta) d\theta = \frac{\alpha_k}{\sum_{\ell} \alpha_{\ell}}$$

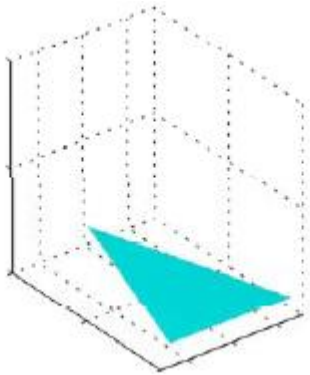
Dirichlet分布（K=3）$p(\theta)=\frac{1}{Z(\alpha)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1}$

- $\theta=(\theta_1,\ldots,\theta_K)\sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_K)$ $\sum_k \theta_k=1,\ \theta_k\geq 0$

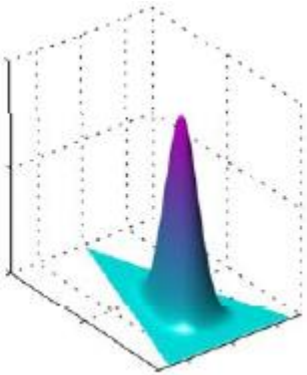
  θ 定义在维数为 K-1 的单纯形上



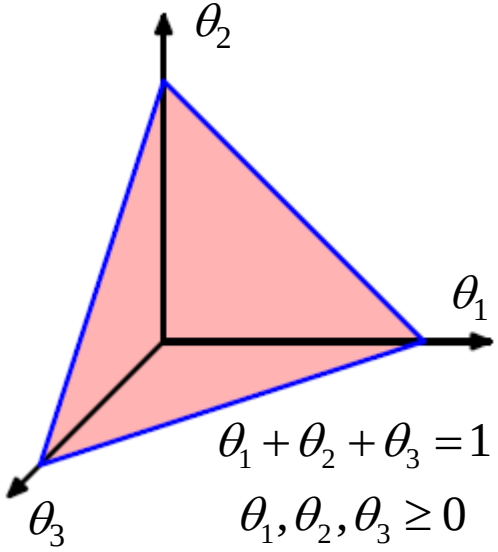
$\alpha_1=\alpha_2=\alpha_3=0.1$



$\alpha_1=\alpha_2=\alpha_3=1$



$\alpha_1=\alpha_2=\alpha_3=10$



Dirichlet后验分布 ∝ 多元分布似然函数 · Dirichlet先验分布

$$p(\theta|D) \propto p(D|\theta)\,p(\theta)$$

- The likelihood function $p(D|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$
- The Dirichlet prior $p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
- The posterior probability θ of given D ~ Dirichlet $(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

$$p(\theta|D) \propto \prod_{k=1}^{K} \theta_k^{N_k} \times \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1}$$

Dirichlet is the conjugate prior for multinomial

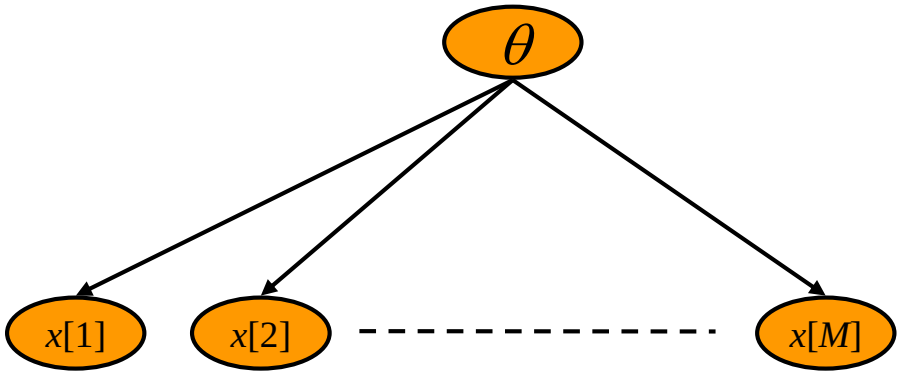
- (最小均方误差下)贝叶斯估计

$$\hat{\theta}_k^{MMSE} = \frac{\alpha_k + N_k}{\sum_l (\alpha_l + N_l)} \qquad \hat{\theta}_k^{ML} = \frac{N_k}{\sum_{l=1}^{K} N_l}$$

超参数 $\alpha_1, \ldots, \alpha_K$ 可视为一种根据先验知识而设定的先验次数（prior count）

Plate notation

- 服从总体分布 $p(x \mid \theta)$，独立同分布采样 $D = (x[1], \ldots, x[M])$ 的贝叶斯网络表示



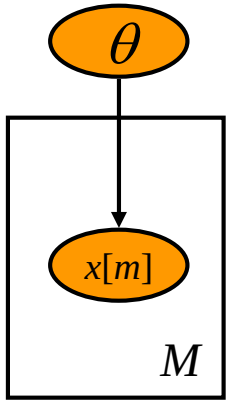


Plate notation

- 用于表示重复结构（repetitive structure）

1）将盒子内的图结构重复多次。
重复的次数由右下角的数字（e.g. M）来指定，
盒子内变量的编号相应变动（e.g. m）

2）进入盒子、离开盒子的有向边 进行相应的重复。

Learning parameters for BNs (complete data)

- 考虑贝叶斯网络 $x = \{x_1, \ldots, x_N\}$
  假设：各个条件分布 $p(x_1 \mid pa_1), \ldots, p(x_N \mid pa_N)$ 有各自表征参数 $\{\theta_1, \ldots, \theta_N\}$

头姿类别 $x_1 \in 1{:}K$

观测图像 $x_2 \in R^{44*28}$

总体分布：$p(x_1, x_2 \mid \theta)$

$N(x_2 \mid \mu_k, \Sigma_k)$

$p(x_2 \mid x_1, \theta_2)$的参数：$\{\mu_k, \Sigma_k\}_{k=1:K}$　　$p(x_1 \mid \theta_1)$的参数：$\{\pi_k\}_{k=1:K}$

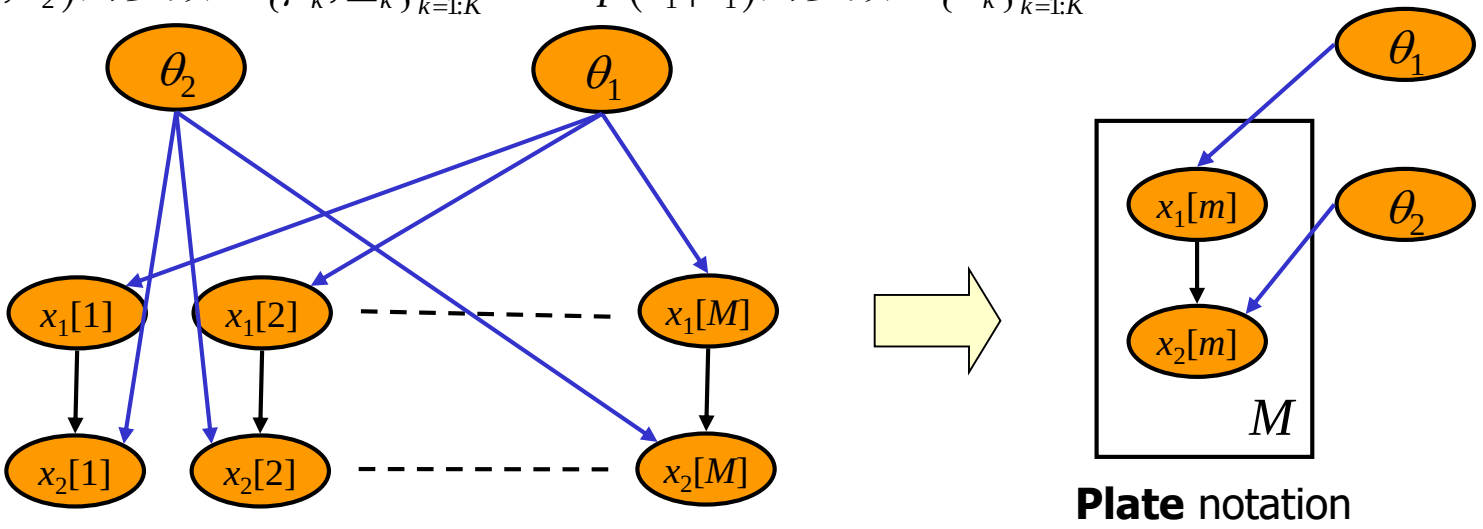
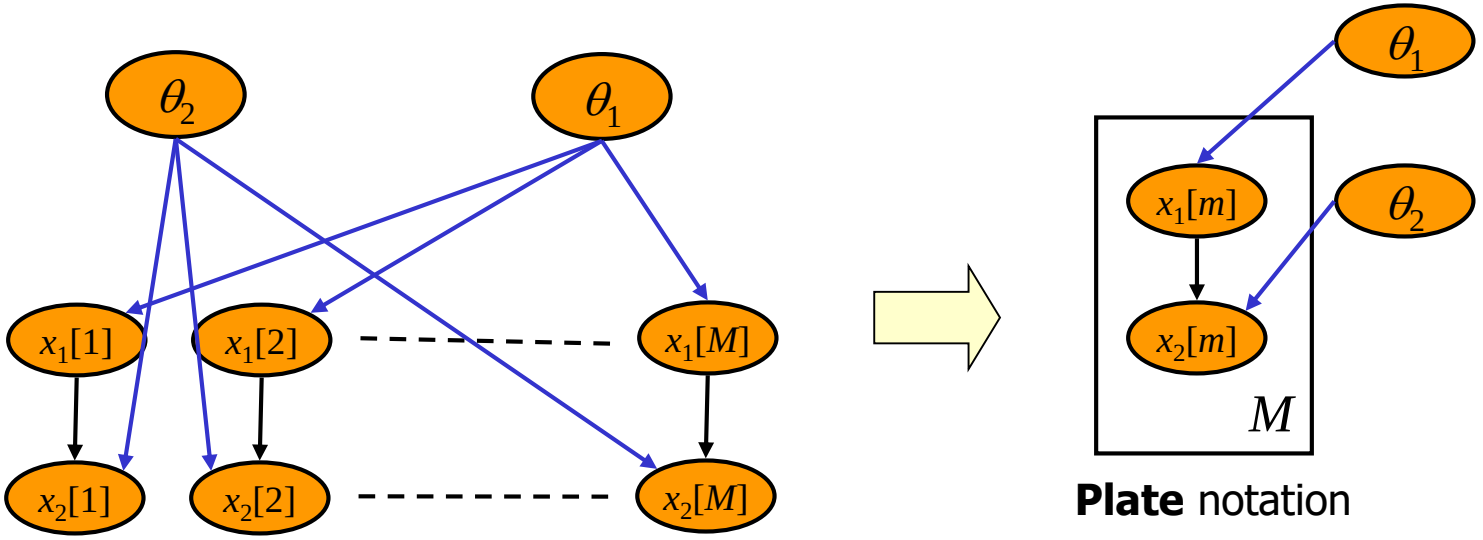


Plate notation

Learning parameters for BNs (complete data)



- Definition: Global parameter independence $p(\theta) = \prod_{n=1}^{N} p(\theta_n)$

$$p(\theta \mid D) \propto p(\theta) \times p(D \mid \theta) = p(\theta) \times \prod_{m=1}^{M} p(x[m] \mid \theta)$$

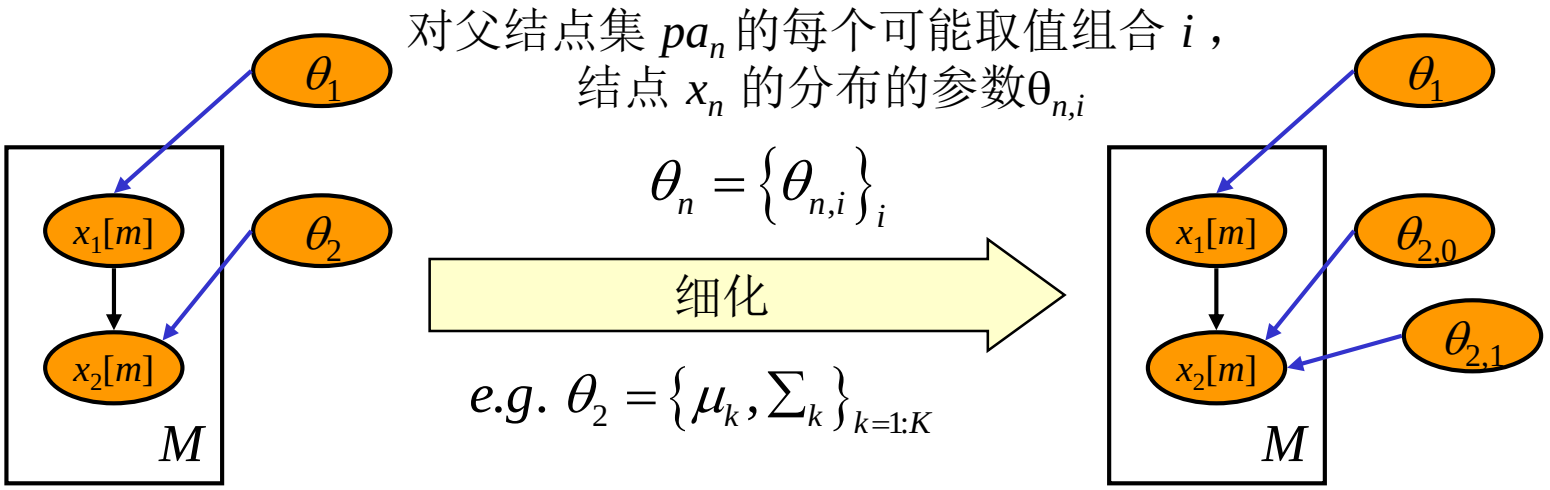
$$= p(\theta) \times \prod_{m=1}^{M} \prod_{n=1}^{N} p(x_n[m] \mid pa_n[m], \theta_n)$$

$$p(\theta \mid D) \propto \prod_{n=1}^{N} \left\{ p(\theta_n) \times \prod_{m=1}^{M} p(x_n[m] \mid pa_n[m], \theta_n) \right\}$$

$$p(\theta \mid D) = \prod_{n=1}^{N} p(\theta_n \mid D)$$

总体参数的后验分布
= 每个结点处表征参数的后验分布的连乘积

Learning parameters for BNs (complete data)



$$e.g.\ p\left(\{\mu_k, \Sigma_k\}_{k=1:K}\right) = \prod_k p\left(\mu_k, \Sigma_k\right)$$

- Definition: Local parameter independence $p(\theta_n) = \prod_i p(\theta_{n,i})$

$$p(\theta_n \mid D) \propto p(\theta_n) \times \prod_{m=1}^{M} p(x_n[m] \mid pa_n[m], \theta_n)$$

$$= \prod_i \left\{ p(\theta_{n,i}) \cdot \prod_{\substack{1 \le m \le M \\ s.t.\ pa_n[m]=i}} p(x_n[m] \mid pa_n[m] = i, \theta_{n,i}) \right\}$$

$$p(\theta_n \mid D) = \prod_i p(\theta_{n,i} \mid D)$$

Bayes estimate for multinomial Bayes net

对每个结点 n 及父结点集 pa_n 的每个可能取值组合 i
一个单独的多元分布的参数 $\theta_{n,i,k}$

- 充分统计量：次数

$$N_{n,i,k} = \sum_{m=1}^{M} \delta\left(pa_n[m] = i, x_n[m] = k\right)$$

$$\left[\hat{\theta}_{n,i,k}\right]^{ML} = \frac{N_{n,i,k}}{\sum_{l=1}^{K_n} N_{n,i,l}}$$

- 假设局部参数均服从Dirichlet分布 $p\left(\theta_{n,i}\right) \sim Dirichlet\left(\alpha_{n,i,1}, \alpha_{n,i,2}, \cdots, \alpha_{n,i,K_n}\right)$

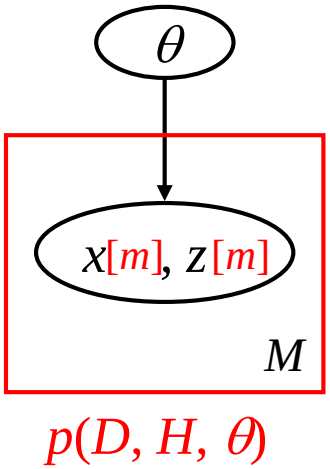
$$\left[\hat{\theta}_{n,i,k}\right]^{MMSE} = \frac{\alpha_{n,i,k} + N_{n,i,k}}{\sum_{l=1}^{K_n}\left(\alpha_{n,i,l} + N_{n,i,l}\right)}$$

Parameter learning

— Bayesian (Known structure, incomplete data)

一般原理

- 总体分布 $p(x, z \mid \theta)$，参数先验分布 $p(\theta)$
  - 总体分布的IID 样本集 $D = (x[1], \ldots, x[M])$：观测数据
    $H = (z[1], \ldots, z[M])$



- 观测到数据 D，求参数的后验分布 $p(\theta \mid D)$？
  - 参数 θ 视为一种特殊的隐变量，化归为 推理计算
- 变分贝叶斯方法（Variational Bayesian）
  - 基于变分推理 求解 参数后验分布 $p(\theta \mid D)$
  - 用变分分布 $q(H, \theta)$ 去近似真实后验分布 $p(H, \theta \mid D)$

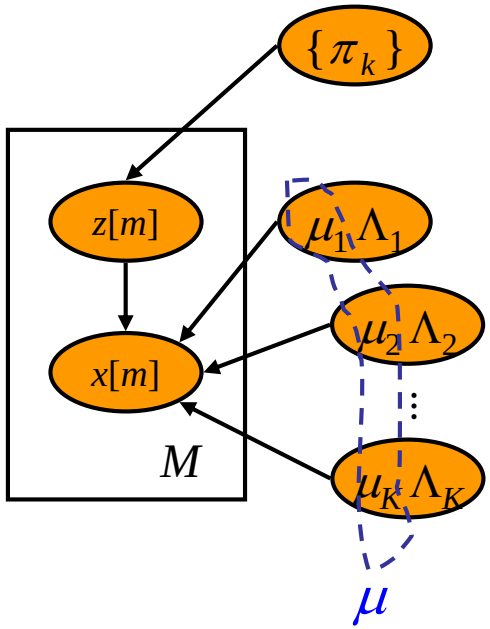
$$q(H,\theta) = q(H)q(\theta)$$

$p(H \mid D)$　$p(\theta \mid D)$

轮换求解：

- $\log q(\theta) = E_{q(H)}\left[\log p(H, D, \theta) \mid \theta\right] + const = \sum_{H} q(H) \log p(H, D, \theta) + const$
- $\log q(H) = E_{q}\left[\log p(H, D, \theta) \mid H\right] + const = \sum_{\theta} q(\theta) \log p(H, D, \theta) + const$

高斯混合模型参数估计的贝叶斯方法



总体分布 $p(z=k,x)=p(z=k)\,p(x\,|\,z=k)$

$$=\pi_k\ N(x\,|\,\mu_k,\Lambda_k)$$

参数先验分布 $p(\theta)=p(\pi)\prod_{k=1}^{K}p(\mu_k,\Lambda_k)$

Dirichlet prior for mixing coefficients

$$p(\boldsymbol{\pi})=C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

Normal-Wishart prior for means and precisions

$$p(\mu_k,\Lambda_k)=\mathcal{N}\left(\mu_k|\mathbf{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0)$$

高斯混合模型参数估计的贝叶斯方法

- 假设如下的变分分布

$$q\left(z[1:M],\pi,\mu,\Lambda\right)=q\left(z[1:M]\right)q\left(\pi,\mu,\Lambda\right)$$

$$q(\quad H \quad , \quad \theta \quad) = q(\quad H \quad)q(\quad \theta \quad)$$

No other assumptions!

- 变分推理结果

$$q\left(\pi,\mu,\Lambda\right)=q\left(\pi\right)\prod_{k=1}^{K}q\left(\mu_k,\Lambda_k\right)$$

~ Dirichlet

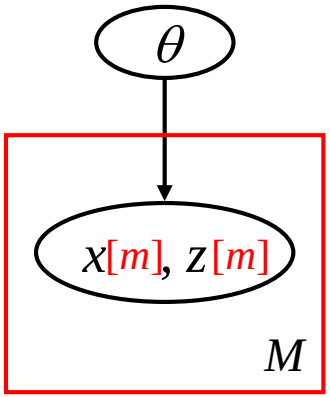
~ Normal-Wishart

$$q\left(z[1:M]\right)=\prod_{m=1}^{M}q\left(z[m]\right)$$

~ Multinomial

详见Bishop书 10.2 Illustration: Variational Mixture of Gaussians

VB discussion：点估计



- 总体分布 $p(x, z \mid \theta)$，参数先验分布 $p(\theta)$
  - 总体分布的IID 样本集

$D = (x[1], \ldots, x[M])$：观测数据

$H = (z[1], \ldots, z[M])$

- 观测到数据 D，求参数的后验分布 $p(\theta \mid D)$？
- 变分贝叶斯方法（Variational Bayesian）

$$p(H, \theta \mid D) \approx q(H, \theta) = q(H)q(\theta)$$

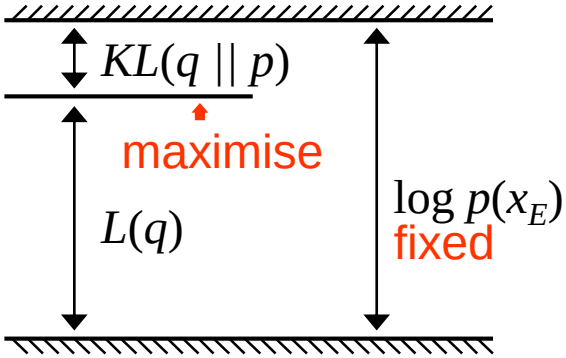
轮换求解：

$$\log q(\theta) = E_q\left[\log p(H, D, \theta) \mid \theta\right] + const = \sum_H q(H) \log p(H, D, \theta) + const$$

$$\log q(H) = E_q\left[\log p(H, D, \theta) \mid H\right] + const = \sum_\theta q(\theta) \log p(H, D, \theta) + const$$

只关心 θ 的点估计 $\xrightarrow[q(\theta) = \delta(\theta - \theta^*)]{\text{约束}}$ $\min_{\text{约束}q(\theta)\text{为}\delta\text{函数}} KL\left[q(H)q(\theta) \| p(H, \theta \mid D)\right]$

VB discussion：点估计



- 变分贝叶斯方法（Variational Bayesian）

$$\min_{\text{约束}q(x_k)\text{为}\delta\text{函数}} KL\left[\prod_{i\in H} q(x_i) \,\middle\|\, p(x_H \mid x_E)\right]$$

针对$q(x_k)$的最优化：将 $L(q)$ 视为 $q(x_k)$ 的函数，与 $q(x_k)$ 无关项并入常数

$$L(q) = H\left(q(x_k)\right) + \sum_{x_k} q(x_k)\log \tilde{p}(x_k \mid x_E) + \text{常数}$$

$$= -KL\left(q(x_k) \,\|\, \tilde{p}(x_k \mid x_E)\right) + \text{常数}$$

无约束最优化：$\log q(x_k) = \log \tilde{p}(x_k \mid x_E) = E_q\left[\log p(x_H, x_E) \mid x_k\right] + \text{常数}$

有约束最优化：$q(x_k) = \delta(x_k - x_k^*)$，其中 $x_k^* = \arg\max_{x_k} \log \tilde{p}(x_k \mid x_E)$

$$= \arg\max_{x_k} E_q\left[\log p(x_H, x_E) \mid x_k\right]$$

约束$q(x_k)$为δ函数

From VB to 不完备数据下MAP估计

$$q(H,\theta)=q(H)q(\theta)$$

轮换求解：

- $\log q(\theta)=E_q\left[\log p(H,D,\theta)\,|\,\theta\right]+const$

$$x_k^*=\arg\max E_q\left[\log p(x_H,x_E)\,|\,x_k\right]$$

只关心 θ 的点估计 $\xrightarrow[q(\theta)=\delta(\theta-\theta^*)]{\text{约束}}$ $\theta^*=\arg\max_{x_k,\theta} E_q\left[\log p(H,D,\theta)\,|\,\theta\right]$

$$\theta^*=\arg\max_{\theta}\sum_H p\left(H\,|\,D,\theta^{(old)}\right)\log p(H,D,\theta)$$

$$\theta^*=\arg\max_{\theta}\left\{\left[\sum_H p\left(H\,|\,D,\theta^{(old)}\right)\log p(H,D\,|\,\theta)\right]+\log p(\theta)\right\}$$

- $\log q(H)=E_q\left[\log p(H,D,\theta)\,|\,H\right]+const$

$$\log q(H)=\sum_\theta q(\theta)\log p(H,D,\theta)+const$$

$$\delta\left(\theta-\theta^{(old)}\right)$$

$$=\log p\left(H,D,\theta^{(old)}\right)+const$$

$$q(H)\propto p\left(H,D,\theta^{(old)}\right)$$

$$ML:\ \max_\theta p(D\,|\,\theta)$$

$$MAP:\ \max_\theta p(D\,|\,\theta)\,p(\theta)$$

From VB to ICM (Iterative contional modes)

$$q(H,\theta)=q(H)q(\theta)$$

众数，最频值，最常出现的变量值

轮换求解：

- $\log q(\theta)=E_q\left[\log p(H,D,\theta)\,|\,\theta\right]+const$

只关心 θ 的点估计 $\xrightarrow[q(\theta)=\delta(\theta-\theta^*)]{\text{约束}}$ $\theta^*=\arg\max_\theta E_q\left[\log p(H,D,\theta)\,|\,\theta\right]$

$$=\sum_H q(H)\log p(H,D,\theta)$$

$$\theta^*=\arg\max_\theta \log p(\theta,|D,H^*)$$

- $\log q(H)=E_q\left[\log p(H,D,\theta)\,|\,H\right]+const$

只关心 H 的点估计 $\xrightarrow[q(H)=\delta(H-H^*)]{\text{约束}}$ $H^*=\arg\max_H E_q\left[\log p(H,D,\theta)\,|\,H\right]$

$$=\sum_\theta q(\theta)\log p(H,D,\theta)$$

$$H^*=\arg\max_H \log p(H,|D,\theta^*)$$

推理计算：$p(x_H \mid x_E)$

Gibbs采样：

$$\hat{x}_k - sampling\ from\ p\left(x_k \mid x_{H\setminus\{k\}}, x_E\right)$$

$$\hat{x}_k - sampling\ from\ p\left(x_H, x_E\right)$$

轮换采样：

$x_1, x_2, x_3, \cdots, x_K$

$\hat{x}_1, x_2, x_3, \cdots, x_K$

$x_1, \hat{x}_2, x_3, \cdots, x_K$

均值场变分：

$$\log \hat{q}\left(x_k\right) = E_q\left[\log p\left(x_H, x_E\right) \mid x_k\right] + const$$

轮换求期望：

$q(x_1), q(x_2), q(x_3), \cdots, q(x_K)$

$\hat{q}(x_1), q(x_2), q(x_3), \cdots, q(x_K)$

$\hat{q}(x_1), \hat{q}(x_2), q(x_3), \cdots, q(x_K)$

ICM：$\max\limits_{x_H} p\left(x_H \mid x_E\right)$

$$\min_{\text{约束}q(x_H)\text{为}\delta\text{函数}} KL\left[q\left(x_H\right) \middle\| p\left(x_H \mid x_E\right)\right]$$

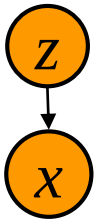
$$x_k^* = \arg\max_{x_k} p\left(x_H, x_E\right)$$

轮换求最大化：

$x_1, x_2, x_3, \cdots, x_K$

$x_1^*, x_2, x_3, \cdots, x_K$

$x_1, x_2^*, x_3, \cdots, x_K$



Iterative learning tradeoff efficiency and accuracy

H是否采取点估计 \ θ是否采取点估计	No	Yes
No	$q(H), q(\theta)$ Expectation-Expectation (EE) Variational Bayes (VB) VB Mixtures of Gaussians	$q(H), \theta^*$ Expectation-Maximization (EM) Mixture of Gaussians
Yes	$H^*, q(\theta)$ Maximization-Expectation (ME) Bayesian K-Means	H^*, θ^* Maximization-Maximization (MM) Iterative Conditional Modes (ICM) K-Means

Max Welling and Kenichi Kurihara. Bayesian K-Means as a "Maximization-Expectation" Algorithm. SIAM Conference on Data Mining (SDM2006)
http://www.ics.uci.edu/~welling/publications/publications.html

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 9, SEPTEMBER 2005 1

A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models

Brendan J. Frey, *Senior Member*, *IEEE*, and Nebojsa Jojic

The learning problem

	Known structure		Unknown structure
Complete data	ML	Bayesian	
Incomplete data	ML	Bayesian	

Structure learning

— from complete data

Focus on score-and-search approach

结构与数据的匹配程度

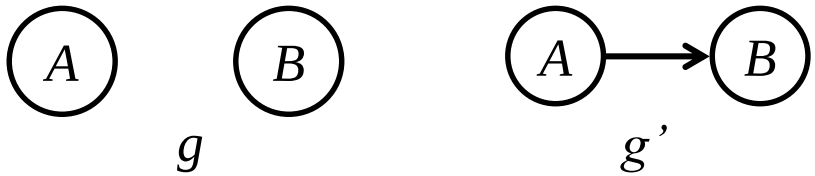
Likelihood score

- IID样本集 $D = (x[1], \ldots, x[M])$ 下似然函数

$$\max_{\theta^g} p\left(D \mid g, \theta^g\right) = \prod_{m=1}^{M} p\left(x[m] \mid g, \theta^g\right)$$

- 结构g的似然得分　　最大似然参数估计: $\theta_{ML}^g = \arg\max_{\theta^g} p\left(D \mid g, \theta^g\right)$

$$likelihood\left(g : D\right) = \max_{\theta^g} \log p\left(D \mid g, \theta^g\right) = \log p\left(D \mid g, \theta_{ML}^g\right)$$



g　　g'

  - 引理：g 添加边得到更复杂结构 g'，$likelihood(g' : D) \geq likelihood(g : D)$

$$\max_{\theta^{g'}} p\left(D \mid g', \theta^{g'}\right) \geq \max_{\theta^g} p\left(D \mid g, \theta^g\right)$$

  - 全连接结构具有最大的似然得分
  - 似然得分不适合于做模型选择

Bayesian score

- Bayesian approach: 将未知量视为随机变量
  - 将 g 视为一个随机变量
  - 给定数据 D 下结构 g 的后验分布

$p(g)$: 结构先验分布, e.g. $p(g) \propto c^{|g|}$, $c<1$

$$\max_g p(g|D) = \frac{p(D|g)\,p(g)}{p(D)}$$

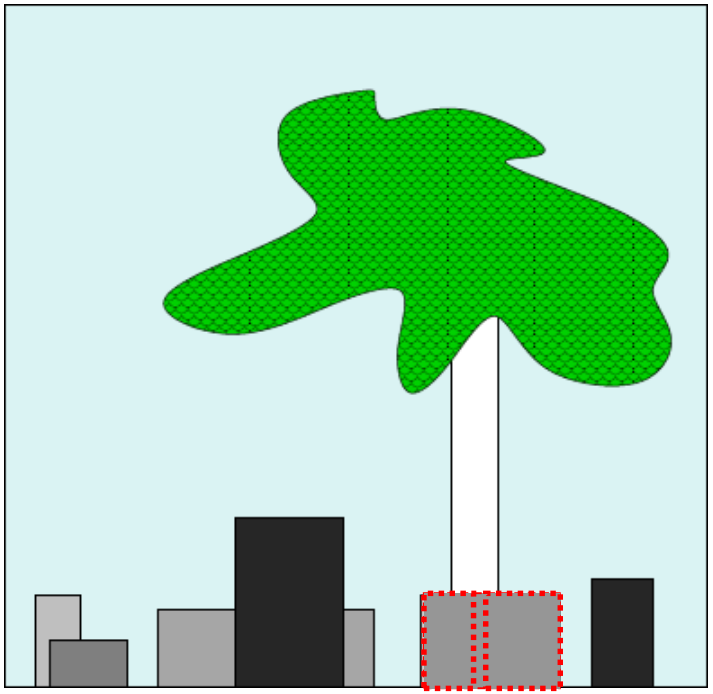
$p(D|g)$: marginal likelihood

$$p(D|g) = \int p(D|g,\theta^g)\,p(\theta^g|g)\,d\theta^g$$

- Bayesian score: $Bayes(g:D) = \log p(D|g) + \log p(g)$

贝叶斯模型选择

Occam's Razor : 接受能拟合数据的最简单模型



树后是一个盒子，

还是两个盒子?

- 贝叶斯模型选择 体现 奥克姆剃须刀原理
  - “如无必要，勿增实体”

贝叶斯模型选择 体现 奥克姆剃须刀原理

- 接受能拟合数据的最简单解释

Data: $D = \{ x[1], x[2], \ldots x[M] \}$

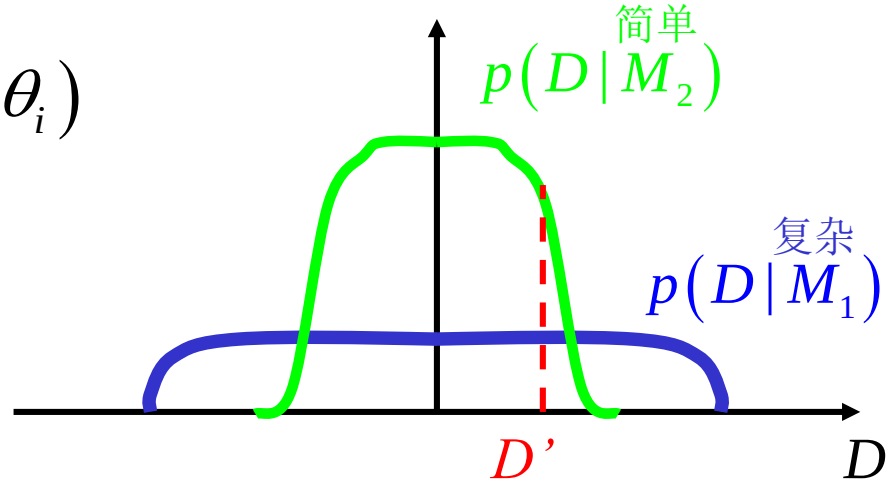
Models M_i : $p(\theta_i \mid M_i), p(x \mid M_i, \theta_i)$

贝叶斯模型选择

$$\max_{M_i} p(M_i \mid D) = \frac{p(M_i)\, p(D \mid M_i)}{p(D)}$$

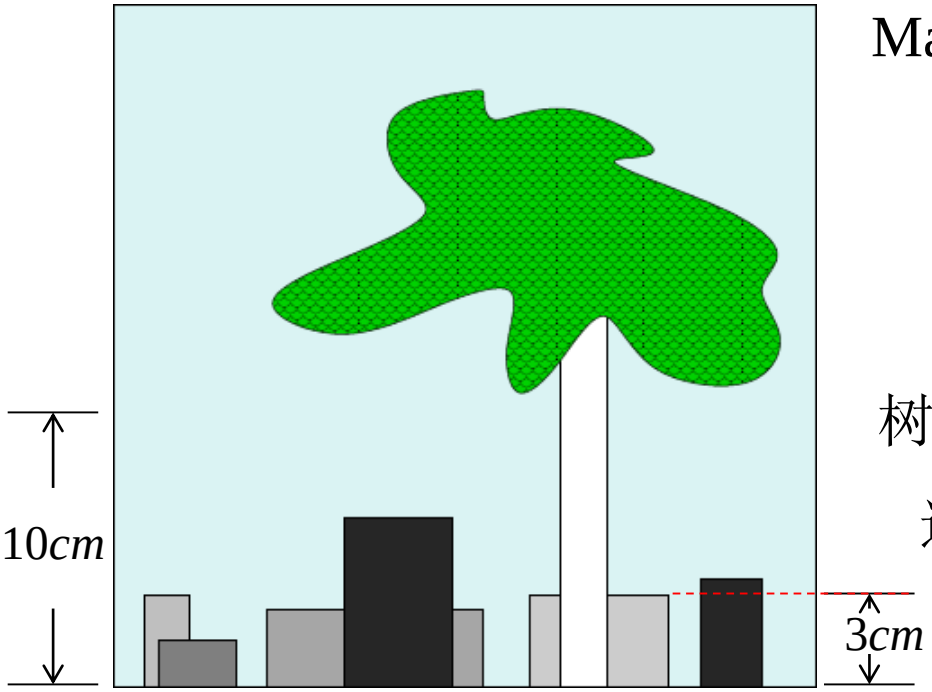
$$\max_{M_i} \; p(D \mid M_i)$$

Marginal likelihood



贝叶斯模型选择 体现 奥克姆剃须刀原理

Occam's Razor : 接受能拟合数据的最简单模型



Mackay书, Section 28

树后是一个盒子，M_1, α

还是两个盒子？ M_2, β_1, β_2

$$p(D|M_1) = \int p(D|M_1,\alpha)\, p(\alpha|M_1)\, d\alpha \quad = 0.1$$

($\alpha = 3cm$)

$$p(D|M_2) = \int p(D|M_2,\beta_1,\beta_2)\, p(\beta_1|M_2)\, p(\beta_2|M_2)\, d\beta_1 d\beta_2 \quad = 0.01$$

($\beta_1 = 3cm$, $\beta_2 = 3cm$)

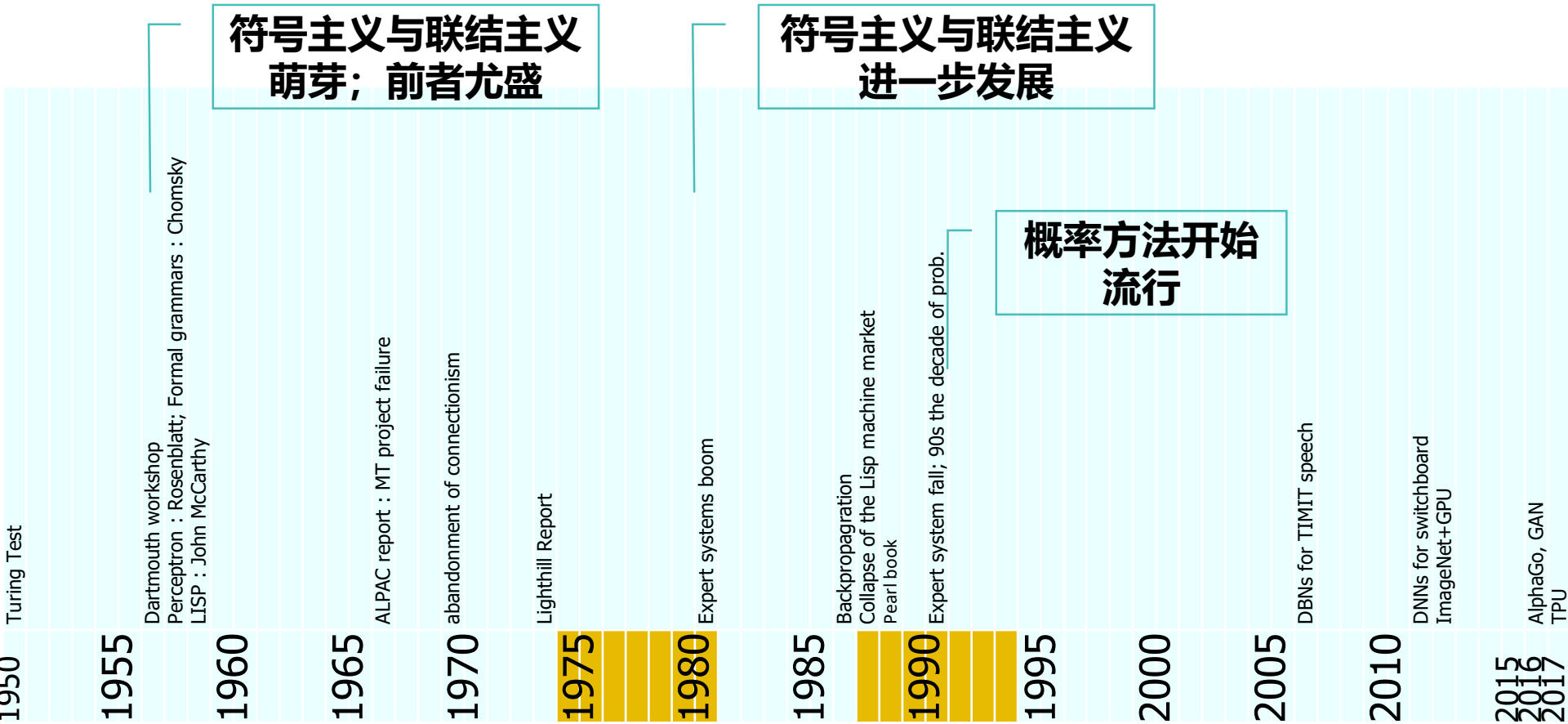
课程章节

- 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**

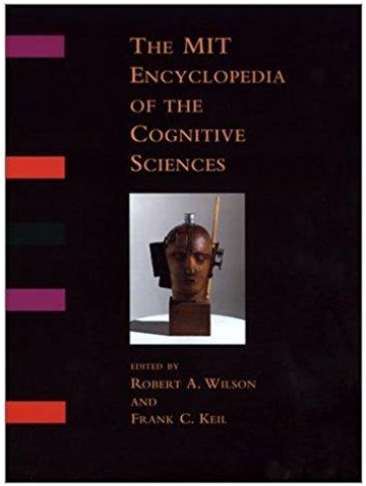
	pgm-1 semantics √		pgm-2 hmm-crf √ pgm-3 exact √	pgm-4 kalman √ pgm-5 sampling √
pgm-6 variational √ pgm-7 ML √	pgm-8 Bayesian √			

Paths to the future

人工智能的发展曲线


符号主义与联结主义
萌芽；前者尤盛
符号主义与联结主义
进一步发展
概率方法开始
流行
1950
Turing Test
1955
Dartmouth workshop
Perceptron : Rosenblatt; Formal grammars : Chomsky
LISP : John McCarthy
1960
1965
ALPAC report : MT project failure
1970
abandonment of connectionism
1975
Lighthill Report
1980
Expert systems boom
1985
Backpropagration
Collapse of the Lisp machine market
Pearl book
1990
Expert system fall; 90s the decade of prob.
1995
2000
2005
DBNs for TIMIT speech
2010
DNNs for switchboard
ImageNet+GPU
2015
2016
AlphaGo, GAN
2017
TPU

MITECS 1999



The three issues raised in the preceding paragraphs—sensorimotor connections to the external world, handling real-valued inputs and outputs, and robust handling of noisy and uncertain information—are primary motivations for the connectionist approach to cognition. (The existence of networks of neurons in the brain is obviously another.) Neural network models show promise for many low-level tasks such as visual pattern recognition and speech recognition. The most obvious drawback of the connectionist approach is the difficulty of envisaging a means to model higher levels of cognition (see BINDING PROBLEM and COGNITIVE MODELING, CONNECTIONIST), particularly when compared to the ability of symbol systems to generate an unbounded variety of structures from a finite set of symbols (see COMPOSITIONALITY). Some solutions have been proposed (see, for example, BINDING BY NEURAL SYNCHRONY); these solutions provide a plausible neural *implementation* of symbolic models of cognition, rather than an *alternative*.

Another problem for connectionist and other propositional approaches is the modeling of *temporally extended* behavior. Unless the external environment is completely observable by the agent's sensors, such behavior requires the agent to maintain some internal state information that reflects properties of the external world that are not directly observable. In the symbolic or logical approach, sentences such as "My car is parked at the corner of Columbus and Union" can be stored in "working memory" or in a "temporal knowledge base" and updated as appropriate. In connectionist models, internal states require the use of RECURRENT NETWORKS, which are as yet poorly understood.

In summary, the symbolic and connectionist approaches seem not antithetical but complementary—connectionist models may handle low-level cognition and may (or rather *must*, in some form) provide a substrate for higher-level symbolic processes. Probabilistic approaches to representation and reasoning may unify the symbolic and connectionist traditions. It seems that the more relevant distinction is between propositional and more expressive forms of representation.

mental representations. Accounts of such phenomena based on probability theory are now widely accepted within AI as an *augmentation* of the purely symbolic view; in particular, probabilistic models are a natural generalization of the logical approach. Recent work has also shown that some connectionist representations (e.g., Boltzmann machines) are essentially identical to probabilistic network models developed in AI (see NEURAL NETWORKS).

感谢对课程的支持！