自然语言处理

2022年秋季

黄河燕、鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

知识体系、问题及方法论

(三) 与贝叶斯理论、概率图模型与信息论

黄河燕，鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

大纲

- NLP的几个问题
- NLP的几个研究范式
- NLP与语言学
- NLP与知识工程
- NLP与机器学习
- NLP与贝叶斯理论和概率图模型
- NLP与信息论

NLP与贝叶斯理论和概率图模型

- NLP的概率统计方法可以用概率图模型(probabilistic graphical model)大一统
- PGM的基础是Bayes理论

NLP与贝叶斯理论和概率图模型

- Bayesian theorem
- PGM

Bayesian theorem

- 概率(probability)　　**{0,1} → [0,1]**

  “probability theory is, au fond, nothing but common sense reduced to calculus”

  ——Laplace (1812)

- 统计(statistical)
  - 数理统计是归纳，从观察值推出背后的数学模型(变量之间的关系)

Bayesian theorem

- 举例：描述身高

1. A来自广东: 180cm,

   B来自广西: 179cm

——所以，A比B高

2. 一组广东人的平均身高: 180cm，

   一组广西人的平均身高: 179cm

——所以，广东人和广西人的平均身高没有差异

统计意义上的描述

Bayesian theorem

What is in a trillion-word data

- "the" appears 23 billion times (2.2% of the trillion words), making it the most common word.
- In three-word sequences, "Find all posts" appears 13 million times (.001%), about as often as "each of the", but well below the 100 million of "All Rights Reserved" (.01%)

——《数据之美》中一个文本统计的例子

Bayesian theorem

- Zipf' law (齐夫定律)

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

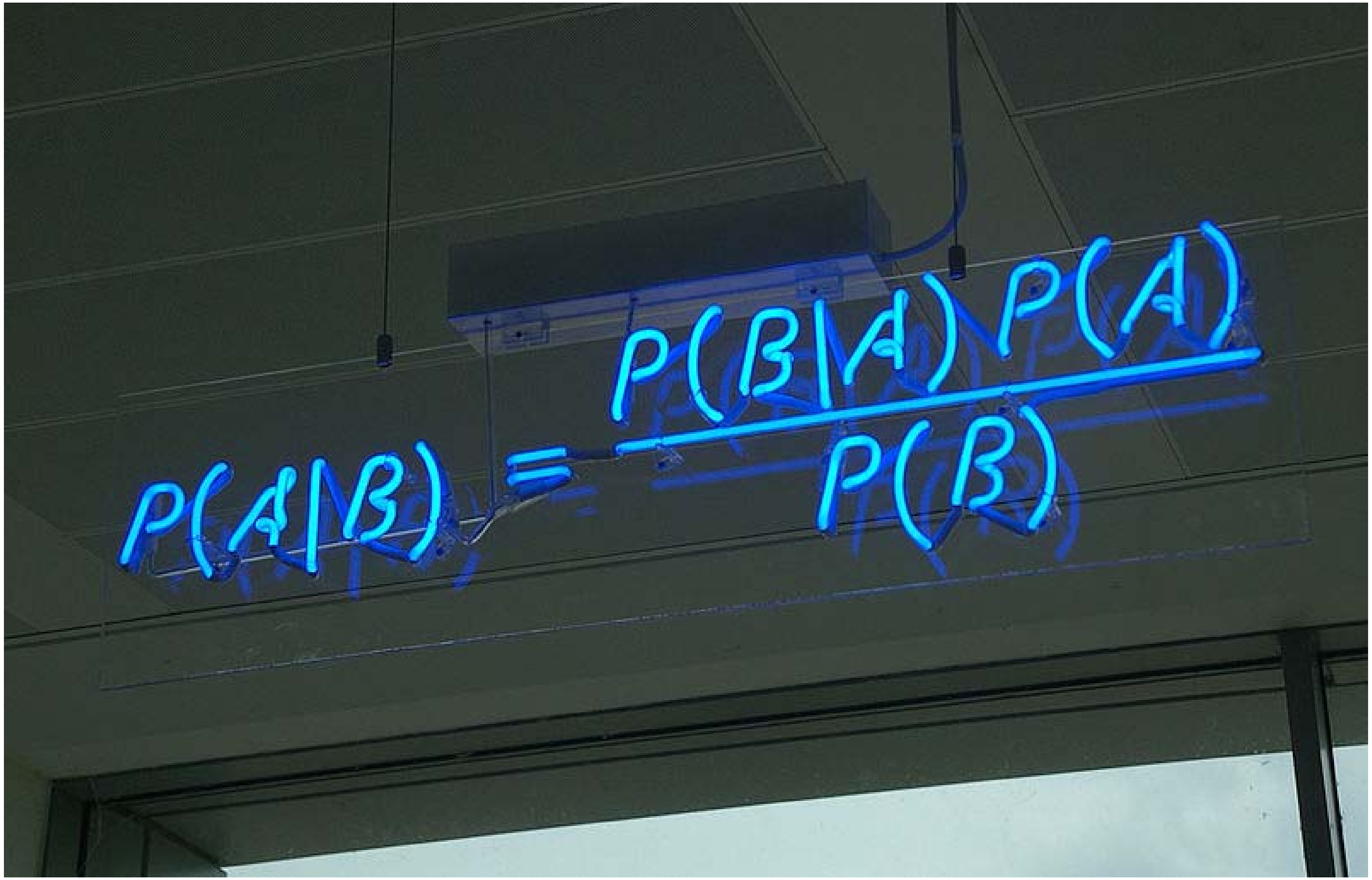
Bayesian theorem

□ Zipf' law

- 如果将某一种语言的每一个词按照它在大规模语料中出现的频率排序，那么词的频率f和它的排位r之间具有这样的关系：

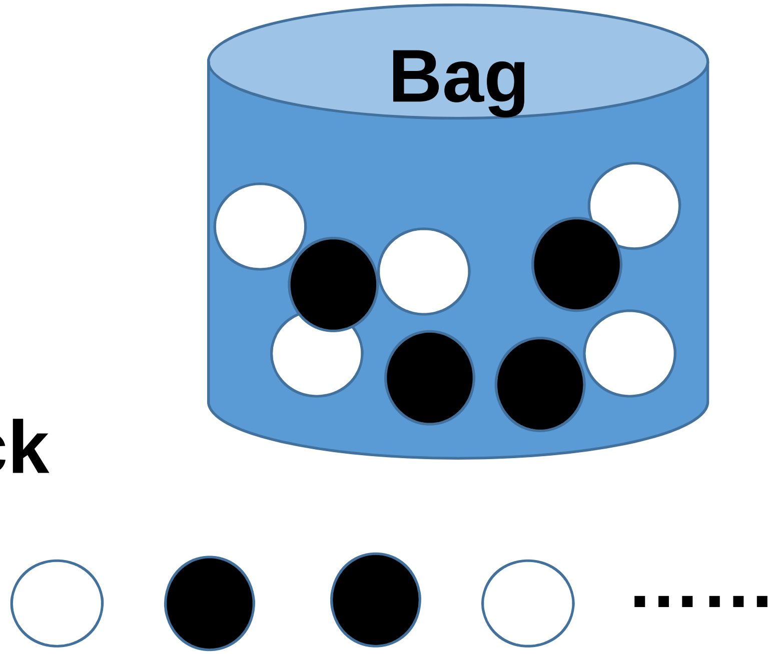
$$f \propto \frac{1}{r}$$

Bayesian theorem



Bayesian theorem

◻ 经典统计学(频率主义)：



Pick

取出一个球，是黑球的概率？

✓ 利用最大似然估计，不断做实验，用观察值来估计——得到某个参数值能够使样本出现的概率为最大

Bayesian theorem

- You cannot know the world exactly
- 我们必须根据observation来“猜”
- 进一步的，还会有一个先验“prior”

似然 likelihood

$$P(h|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|h)P(h)}{P(\boldsymbol{D})}$$

后验

posterior probability

Bayesian theorem

➢ 举例：**orthographic correction**

用户输入: **"thew"**

推断：**"the" or "thaw"?**

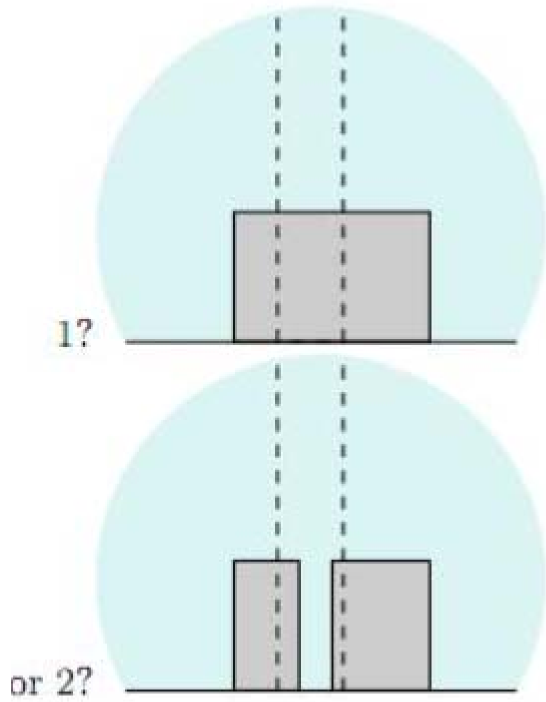
$$P(h|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|h)P(h)}{P(\boldsymbol{D})}$$

$$\max\{P(\text{the}|\text{thew}),\ P(\text{thaw}|\text{thew})\}$$

$$P(\text{thew}|\text{the})\text{P}(\text{the}) \qquad P(\text{thew}|\text{thaw})\text{P}(\text{thaw})$$

✓ 很多情况下，引入逆概，是为了引入先验

- Ockham’s Razor：如非必需，勿增实体
- Bayes定理中的Ockham’s Razor
  - Ockham’s Razor：$P(h)$ 越简单假设越少概率越高
  - Bayes Ockham’s Razor：$P(\boldsymbol{D}|h)$



Bayesian theorem

- Bayes定理在理论上是完美的
- 几乎所有的统计模型都可以归在Bayes框架下
  - Probabilistic graphical model
  - 后验概率估计，Bayes估计
  - Bayesian deep learning
  - 在参数中加入噪声增强模型泛化能力
  - …

回顾

机器学习的一个主要挑战——线性回归为例

- 概率论—最大后验估计(贝叶斯估计的点估计)
  - 假设参数$\boldsymbol{w}$为一个随机向量，并服从一个先验分布

$$p(\boldsymbol{w}|v) = \mathcal{N}(\boldsymbol{w}|0, v^2\boldsymbol{I})$$

  - 最大后验估计(**maximum a posteriori estimation, MAP)**是指最优参数为下述后验分布中概率密度最高的参数$\boldsymbol{w}$：

$$\boldsymbol{w} = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}, v, \sigma)$$

$$= \underbrace{\boxed{\underbrace{\boxed{\operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma)}}_{\text{最大似然估计}} p(\boldsymbol{w}|v)}}_{}$$

最大后验估计

Bayesian theorem

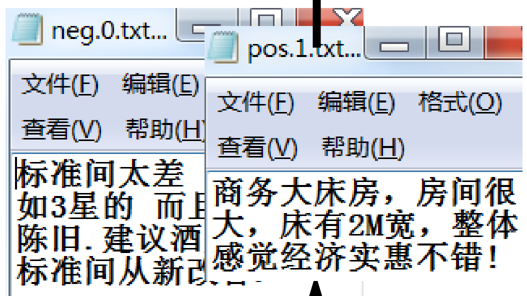
$$e^* = \arg\max_e P(e|f) = \arg\max_e \frac{P(f|e)P(e)}{P(f)}$$

$$= \arg\max_e P(f|e)P(e)$$

翻译模型　　语言模型

噪声信道SMT

$$C^* = \underset{c}{\operatorname{argmax}} P(C_i|W)$$



标注语料中查$P(w_k|C)$

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)}$$

$$= \frac{\prod_k P(w_k|C)P(C)}{P(W)}$$

Naïve Bayes

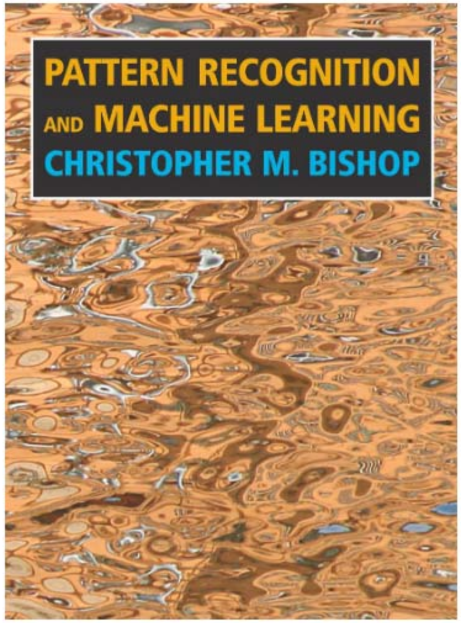
房间 设施 需要 改造

基于朴素贝叶斯的情感分析

NLP与贝叶斯理论和概率图模型

- Bayesian theorem
- PGM

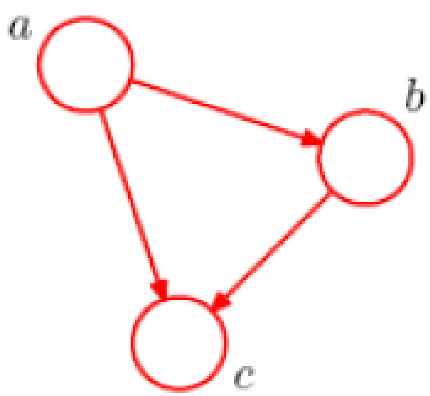
概率图模型(PGM)：统计学习的重要分支，丰富的框架，用于通过概率分布或随机过程来建模有限或无限个可观察或潜在变量之间的复杂交互作用——随机变量为节点，概率相关性为边的图。



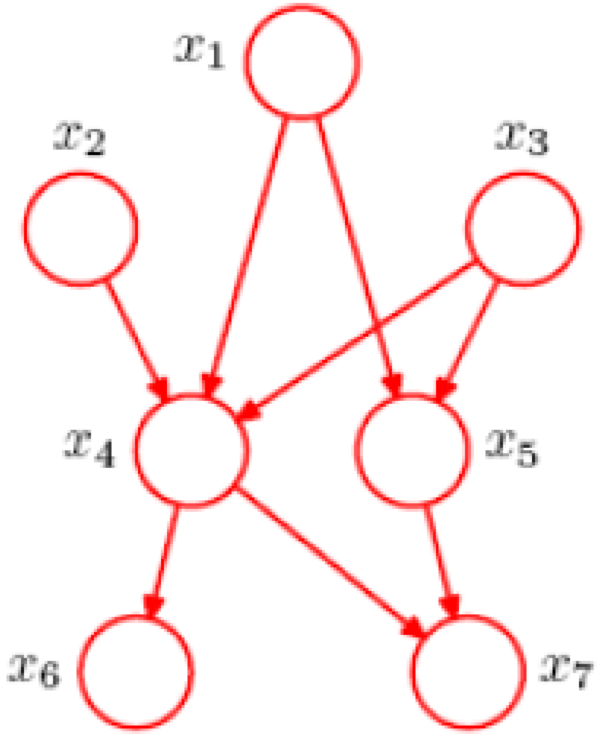
PGM

- Probabilistic graphical models——

  Structured probabilistic models



$P(a, b, c) = ?$

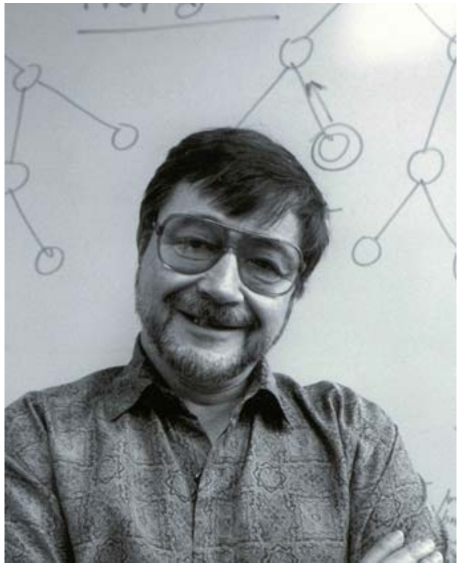


$P(\boldsymbol{x}) = P(x_1, x_2, x_3 \ldots) = ?$

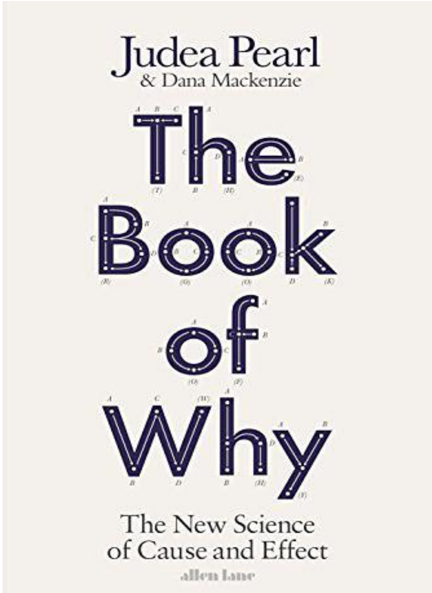
PGM

- Bayesian networks

$$P(h, \boldsymbol{D})$$



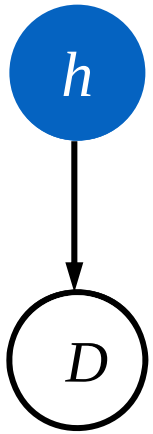
Judea Pearl



PGM

- Bayesian model

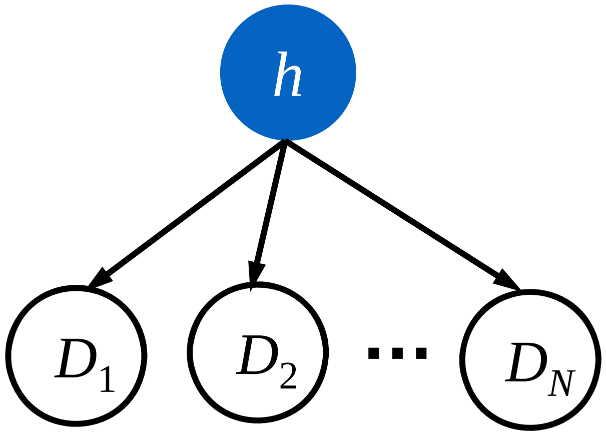
$$P(h, D) = P(h)P(D|h)$$



- Naïve Bayes

$$P(h, \boldsymbol{D}) = P(h)P(\boldsymbol{D}|h) = ?$$

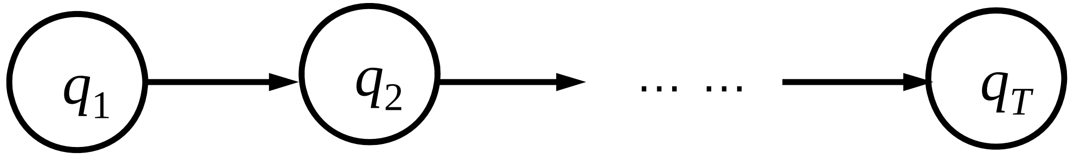
$$P(h)\prod_{i=1}^{N} P(D_i|h)$$



Static Bayesian network

PGM

□ Markov model



$$P(\boldsymbol{Q}) = \prod_{t=1}^{T} P(q_t | q_{t-1})$$

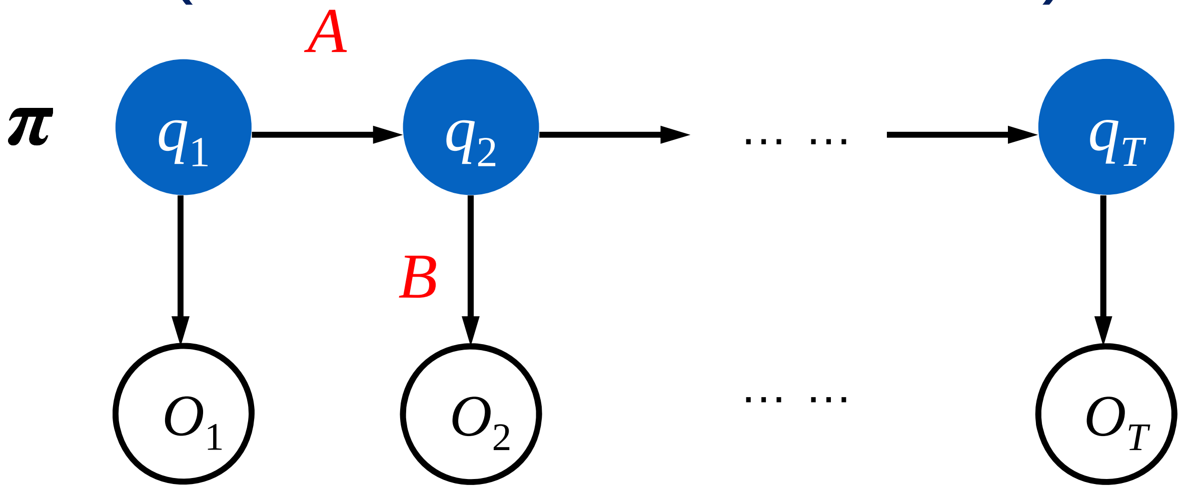
Dynamic Bayesian network

典型应用： ***n*-gram model**

$$P(\boldsymbol{W}) = \prod_{i=1}^{m} P(w_i | w_{i-n+1} w_{i-n+2} \ldots w_{i-1})$$

PGM

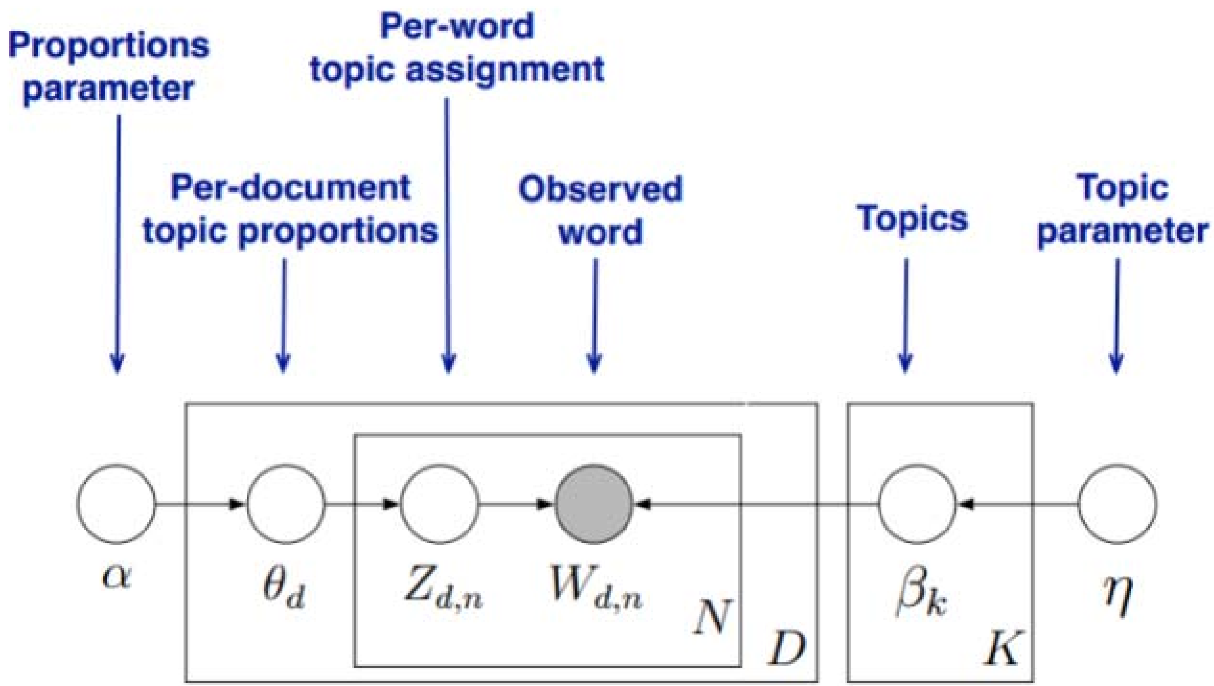
- HMM (hidden Markov model)



$$P(O,Q) = P(O|Q)P(Q) = \prod_{t=1}^{T} P(o_t|q_t)\,P(q_t|q_{t-1})$$

PGM

- LDA (latent Dirichlet allocation)

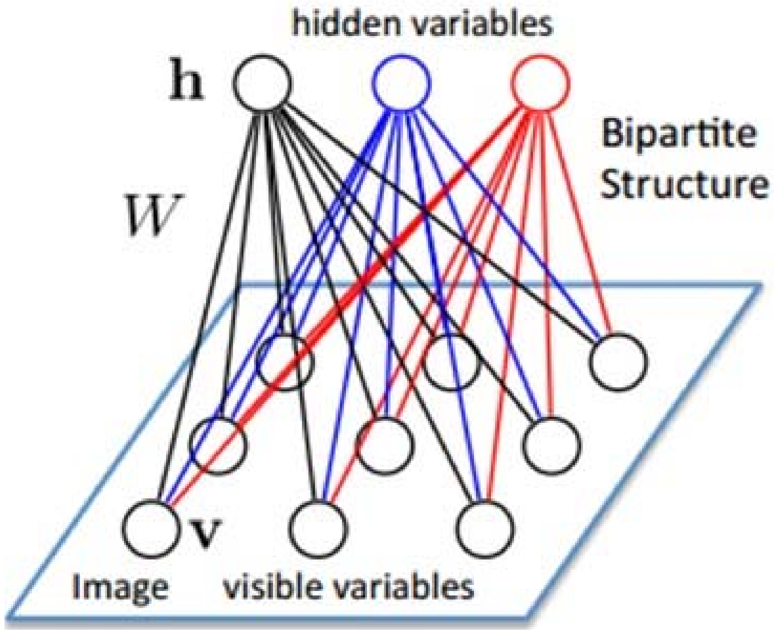


典型的主题模型**(topic model)**，一个三层**Bayes**模型：

$$P(w|d) = \underset{\beta}{P(w|t)}\,\underset{\theta}{P(t|d)}$$

基于隐变量(latent variable)的PGM例子

□ 受限玻尔兹曼机



$$L(\boldsymbol{W}, a, b) = -\sum_{i=1}^{m} \ln\left(P\left(\boldsymbol{v}^{(i)}\right)\right)$$

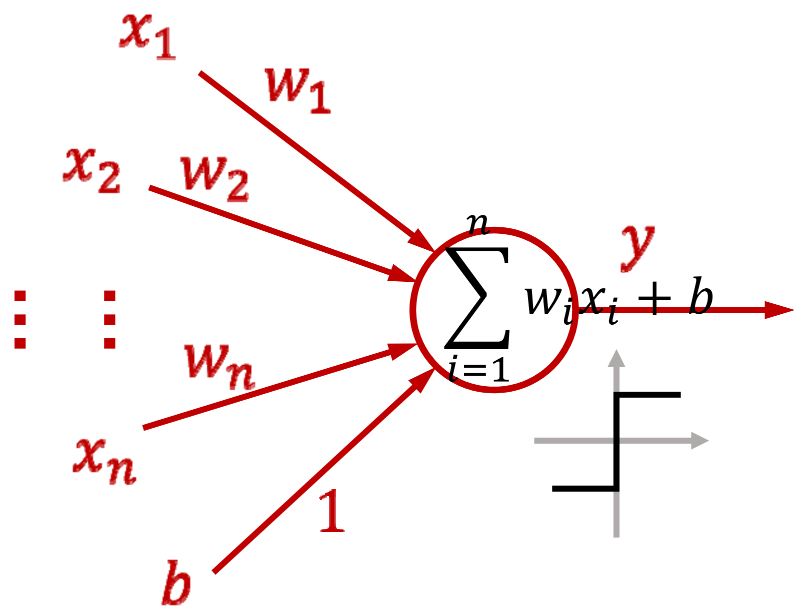
$$P(\boldsymbol{v}, \boldsymbol{W}) = \sum_{h} P(\boldsymbol{h}, \boldsymbol{W}) P(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{W})$$

PGM

- 以上为生成式模型(generative model)，思考其特点

PGM

◻ 感知机(perceptron)

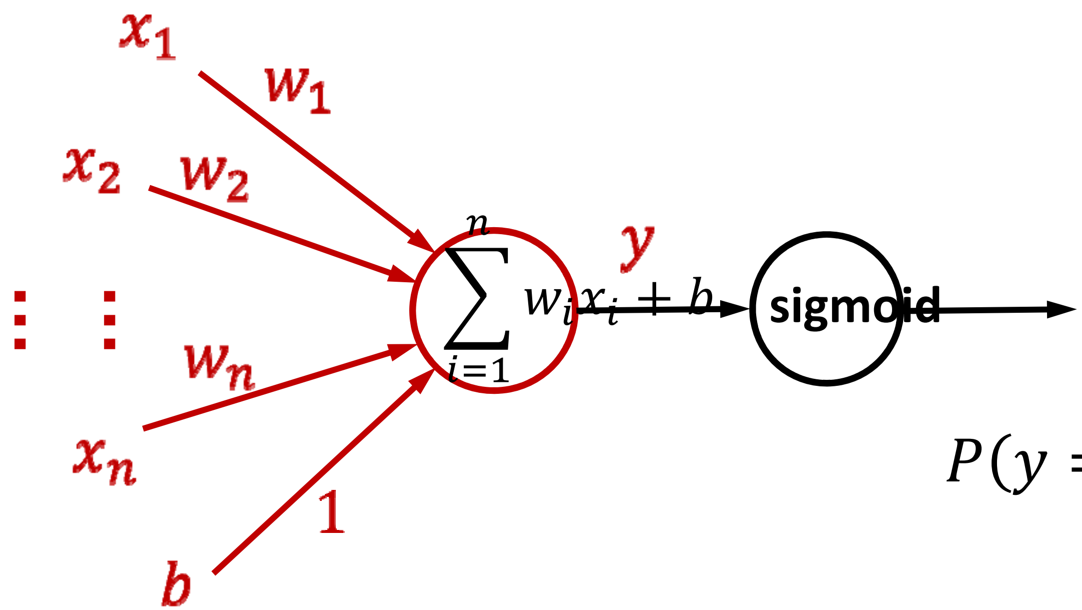


$$y = \mathrm{sgn}(\boldsymbol{w}^T \boldsymbol{x} + b)$$

- 线性模型，有监督判别式学习
- 为了逻辑分类，使用了符号函数
- 损失函数使用均方误差，不可导
- 可以使用错误分类点到分类面的距离

PGM

- LR(logistic regression, 逻辑回归)

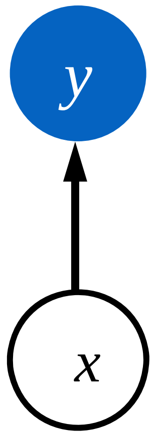


$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-z)}$$

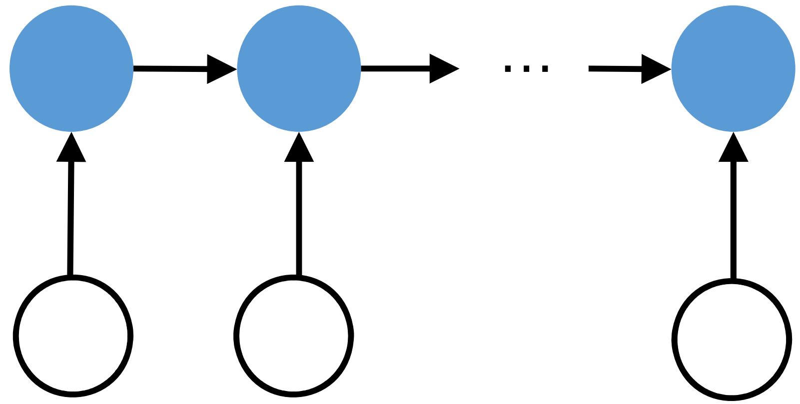
PGM

ME



$$p^*(y|x) = \frac{1}{Z(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

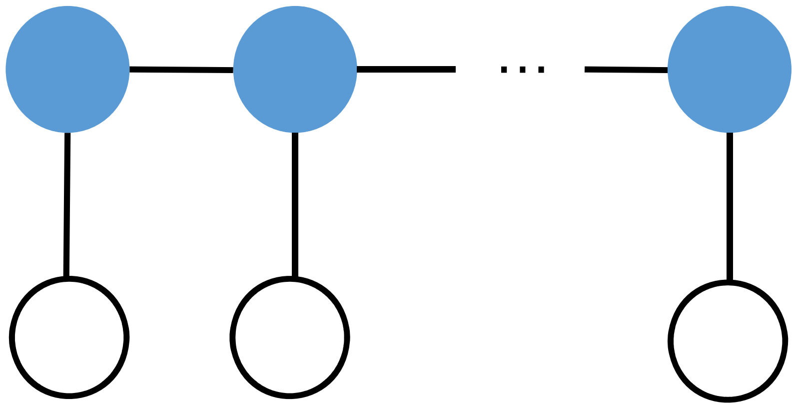
MEMM



最大熵马尔可夫，用于序列标注
有向判别模型。

PGM

◻ CRF(conditional random field)



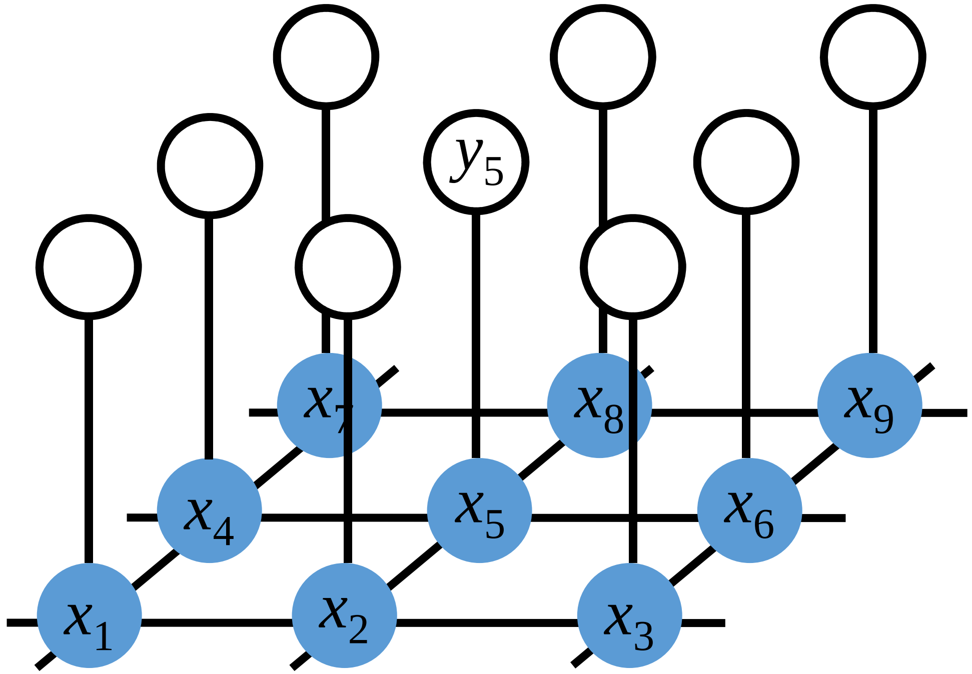
链式条件随机场，用于序列标注(几乎是最好的序列标注模型)。
无向判别模型。

PGM

- 以上为判别式模型(discriminative model)，思考其特点
- DNN是什么类型的模型？

PGM

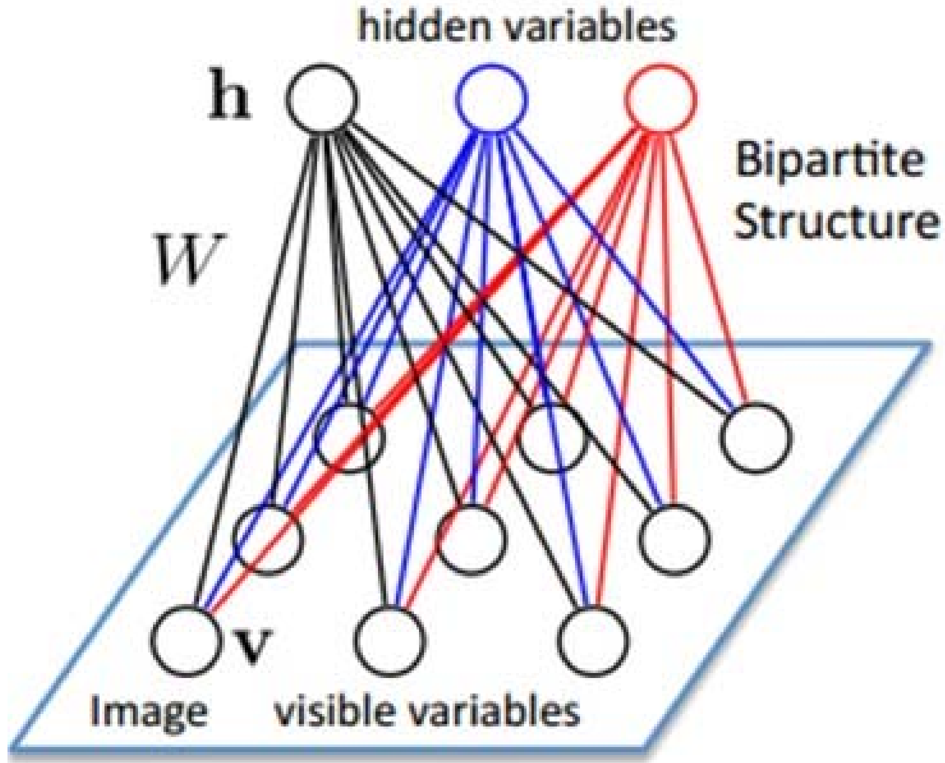
MRF(Markov random field)



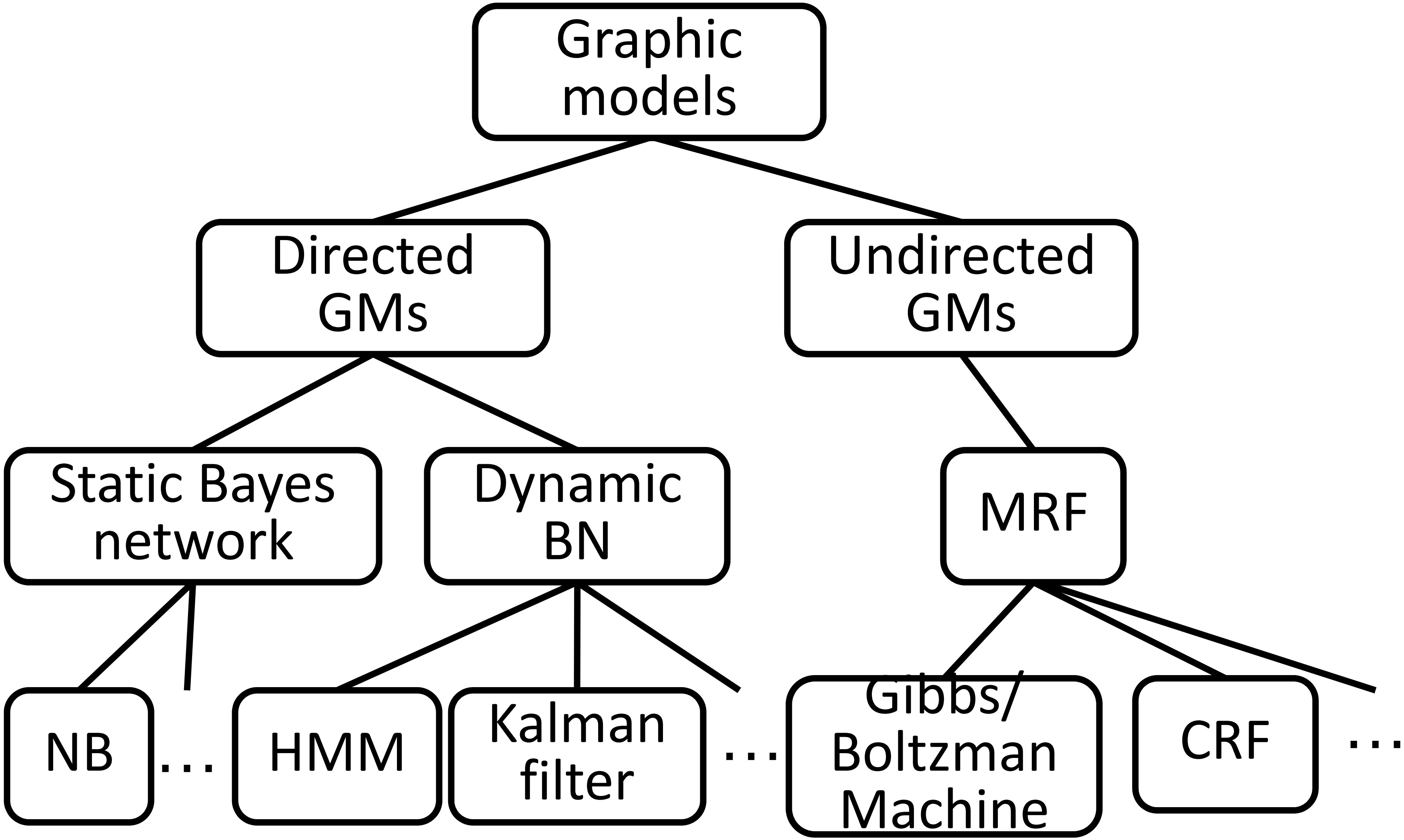
马尔科夫随机场，用于图像处理。无向生成模型。

PGM

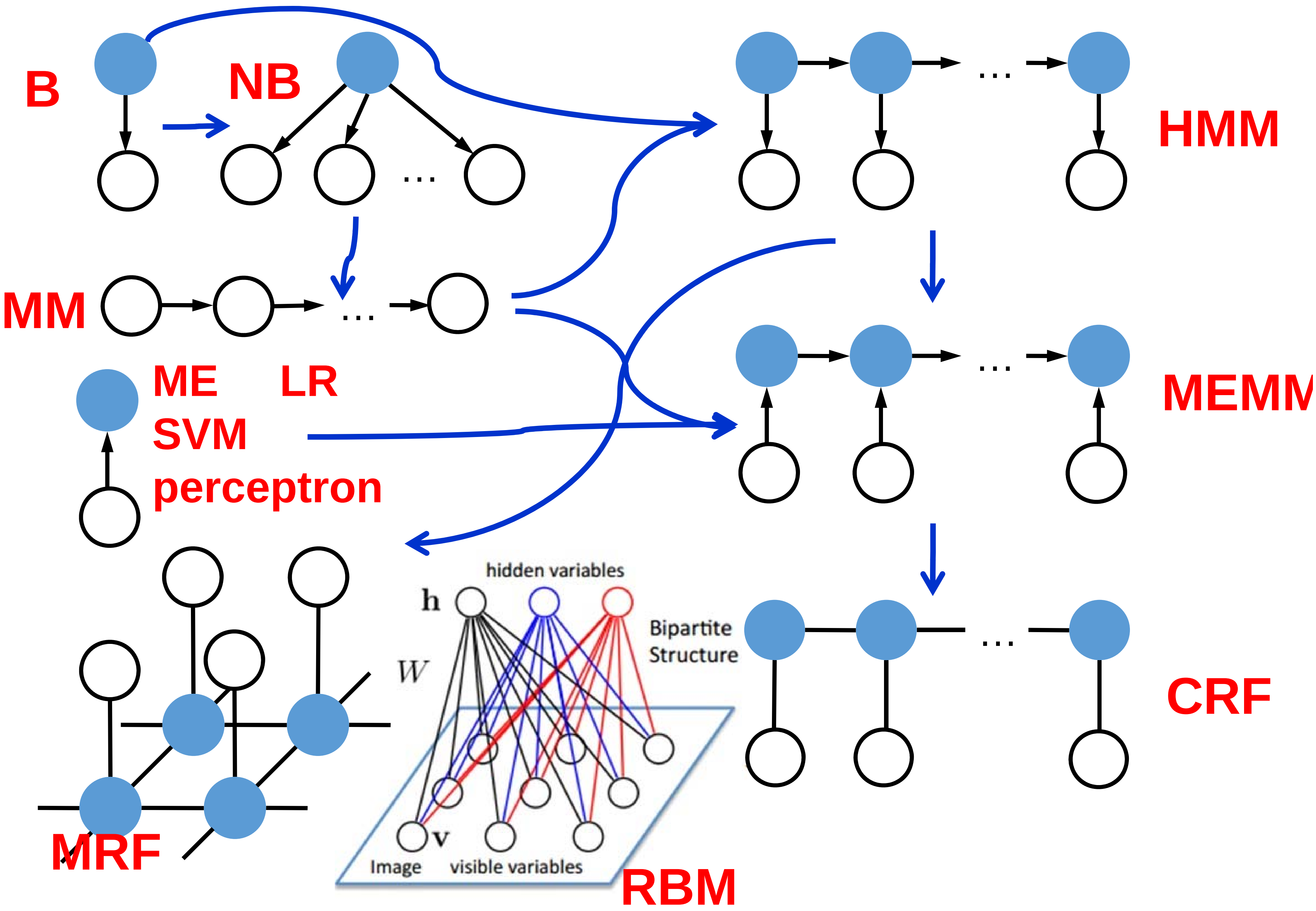
- RBM(restricted Boltzmann machine)



受限玻尔兹曼机，无向生成模型



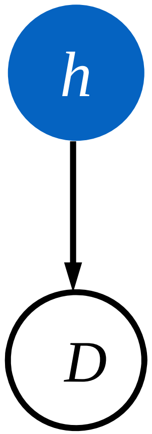
Graphic models


B
NB
HMM
MM
ME
LR
SVM
perceptron
MEMM
hidden variables
h
Bipartite
Structure
W
CRF
v
Image
visible variables
MRF
RBM

PGM

ML的三大问题

1. Modeling
2. Inference
3. Learning

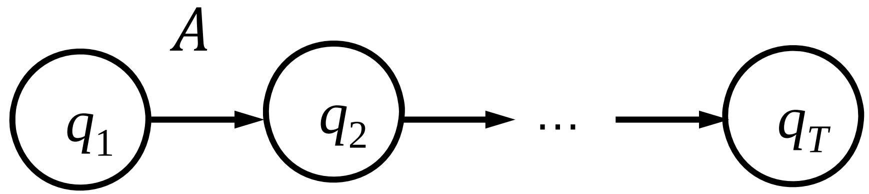


Bayesian model

1. $P(h,D) = P(h)P(D|h)$
2. $h^* = \underset{h}{\operatorname{argmax}}\, P(h|D) = \underset{h}{\operatorname{argmax}} \frac{P(h)P(D|h)}{P(D)}$
3. $P(h)?\, P(D|h)?$

PGM

- n-gram model



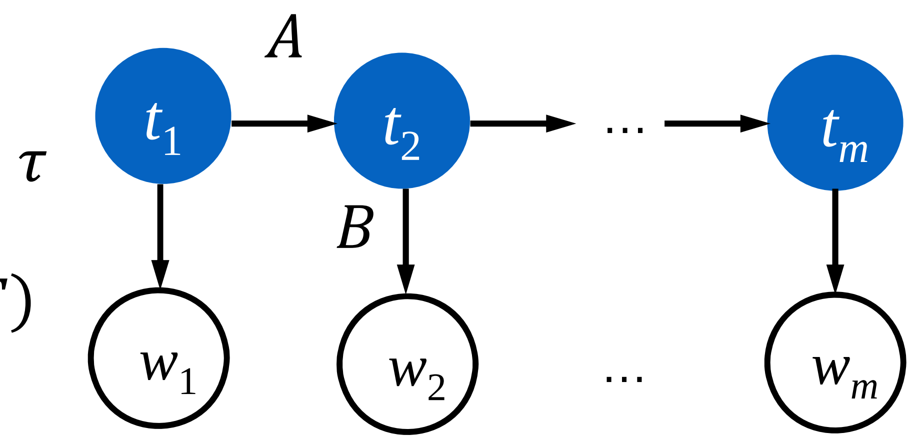
1. $P(W) = \prod_{i=1}^{m} P(w_i | w_{i-n+1} w_{i-n+2} \dots w_{i-1})$
2. $W^* = \underset{W}{\operatorname{argmax}} P(W)$

   $w_i^* = \underset{w}{\operatorname{argmax}} P(w_i | w_{i-n+1} w_{i-n+2} \dots w_{i-1})$
3. $P(w_i | w_{i-n+1} \dots w_{i-1})$?

$$P(w_i | w_{i-n+1} \dots w_{i-1}) = \frac{f(w_{i-n+1} \dots w_i)}{f(w_{i-n+1} \dots w_{i-1})}$$

最大似然估计
maximum likelihood estimation, MLE

PGM



HMM

1. $P(W,T) = P(W|T)P(T)$

$$= \prod_{i=1}^{m} P(w_i|t_i)\,P(t_i|t_{i-1})$$

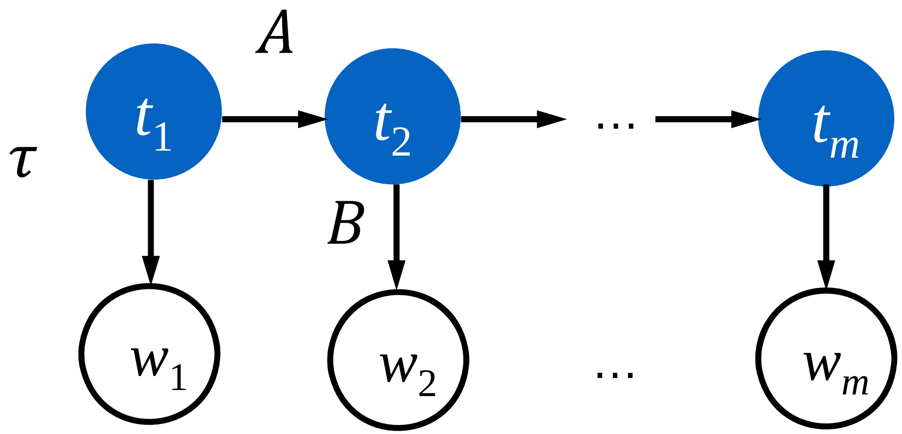
$$P(W) = \sum_{T} P(T)\,P(W|T)$$ 也可以这样**modeling**

2. $T^* = \operatorname*{argmax}_{T} P(T|W) = \operatorname*{argmax}_{T} \dfrac{P(W|T)P(T)}{P(W)}$

$$= \operatorname*{argmax}_{T} \prod_{i=1}^{m} P(w_i|t_i)\,P(t_i|t_{i-1})$$

PGM

□ HMM



3. $T^* =$

$$\underset{T}{\operatorname{argmax}} \prod_{i=1}^{m} P(w_i|t_i)\, P(t_i|t_{i-1})$$

Learning:

① 有监督：最大似然估计(maximum likelihood estimation, MLE)

② 无监督：期望最大化算法(expectation-maximum, EM)

大纲

- NLP的几个问题
- NLP的几个研究范式
- NLP与语言学
- NLP与知识工程
- NLP与机器学习
- NLP与贝叶斯理论和概率图模型
- NLP与信息论

如何度量信息？

- 谁赢得了世界杯？

——吴军《数学之美》

1 2 3 4 5 6 7 8 …… 32

1 2 3 4 …… 16 ? Yes

1 2 …… 8 ? No 9 10 …… 16✓

9 …… 12 ?

……We need 5.

实际上，我们最多需要5

如何度量信息？

- 谁赢得了世界杯？

$$= -(p_1 \log p_1 + p_2 \log p_2 + \cdots + p_{32} \log p_{32})$$

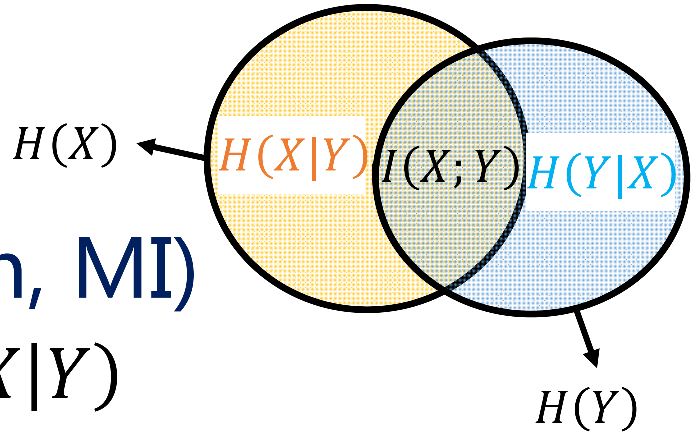
$$\leq \; 5 = -\left(\frac{1}{32}\log\frac{1}{32} + \frac{1}{32}\log\frac{1}{32} + \cdots + \frac{1}{32}\log\frac{1}{32}\right)$$

$$H(X) = -\sum_{x \epsilon X} p(x) \log_2 p(x)$$ **熵(entropy)**

它体现了信息的不确定性——

随机变量的随机性有多大

信息论相关概念



- 互信息(mutual information, MI)

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

在已知Y的值后X的不确定性的减少，即随机变量Y揭示了多少关于X的信息量

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

点互信息 pointwise mutual information

信息论相关概念

□ 交叉熵(cross entropy)

$$H(X,q) = -\sum_{x \in X} p(x) \log q(x)$$

- 用来衡量在给定的真实分布下，使用非真实分布所指定的策略消除系统的不确定性所需要付出的努力的大小

真实分布$p(x)$：$\left\{\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{56},\frac{1}{56}\cdots,\frac{1}{56}\right\}$

非真实分布$q(x)$：$\left\{\frac{1}{32},\frac{1}{32},\frac{1}{32},\cdots,\frac{1}{32}\right\}$

使用了一个非最优估计，将花费更大的代价

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -\left(\frac{1}{8}\log\frac{1}{8} + \frac{1}{8}\log\frac{1}{8} + \cdots + \frac{1}{56}\log\frac{1}{56}\right)$$

$$H(X,q) = -\sum_{x \in X} p(x) \log q(x) = -\left(\frac{1}{8}\log\frac{1}{32} + \frac{1}{8}\log\frac{1}{32} + \cdots + \frac{1}{56}\log\frac{1}{32}\right)$$

信息论相关概念

◘ 相对熵(Kullback-Leibler divergence, KL散度)

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = H(X, q) - H(X)$$

- 衡量两个取值为正的函数或概率分布之间的差异

✓ 在机器学习中的分类算法中，我们总是最小化交叉熵——因为交叉熵越低，由算法所产生的策略最接近最优策略，算法所算出的非真实分布越接近真实分布

信息论相关概念

困惑度(perplexity)

- 用来度量一个概率分布或概率模型预测样本的好坏程度
- 在语言模型中：

$$\left(\prod_{t=1}^{m} P(w_t | w_{t-1}^{t-n+1})\right)^{-\frac{1}{m}}$$

从熵的角度解释：当前状态下(已知前$n-1$个词)，后面能填入的平均的词的数量

0 language model: 997

2-gram collocation: 60

2-gram model: 20

信息论相关概念

最大熵(maximum entropy)

- 举例：词“获得”有这样的翻译候选

  { get, obtain, attain, gain, make } 如何估计分布？

Constrain：

$$p(\text{get}) + p(\text{obtain}) + p(\text{attain}) + p(\text{gain}) + p(\text{make}) = 1$$

直觉上：

$$p(\text{get}) = p(\text{obtain}) = p(\text{attain}) = p(\text{gain}) = p(\text{make}) = \frac{1}{5}$$

信息论相关概念

One more constrain:

$$p(\text{get}) + p(\text{obtain}) = \frac{2}{3}$$

直觉上：

$$p(\text{get}) = p(\text{obtain}) = \frac{1}{3}$$

$$p(\text{attain}) = p(\text{gain}) = p(\text{make}) = \frac{1}{9}$$

信息论相关概念

One more constrain:

$$p(\text{get}) + p(\text{make}) = \frac{1}{2}$$

直觉上：???

直觉是如何工作的

1. 满足全部已知的条件
2. 对未知的情况不做任何主观假设

Uniform distribution　Make entropy maximum

✓ 有证据就用，没有就认为它是随机的——因为不随机的你已经考虑完了，如果再考虑非随机分布你就没有理由了， 不是最好地模拟已知或世界了。

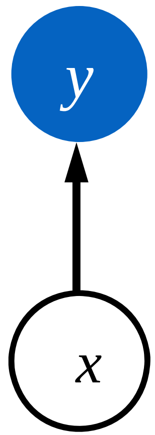
信息论相关概念

最大熵模型(ME)

- 给定先验数据，使熵最大的概率分布就是最好的分布

$$p^* = \underset{p \in P}{\operatorname{argmax}} H(Y|X) \quad \text{constrains: } P(f) = \bar{P}(f)$$

$$p^* = \underset{p \in P}{\operatorname{argmax}} \sum_{(x,y)} p(y|x)\bar{p}(x) \log \frac{1}{p(y|x)}$$



……

$$p^*(y|x) = \frac{1}{Z(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

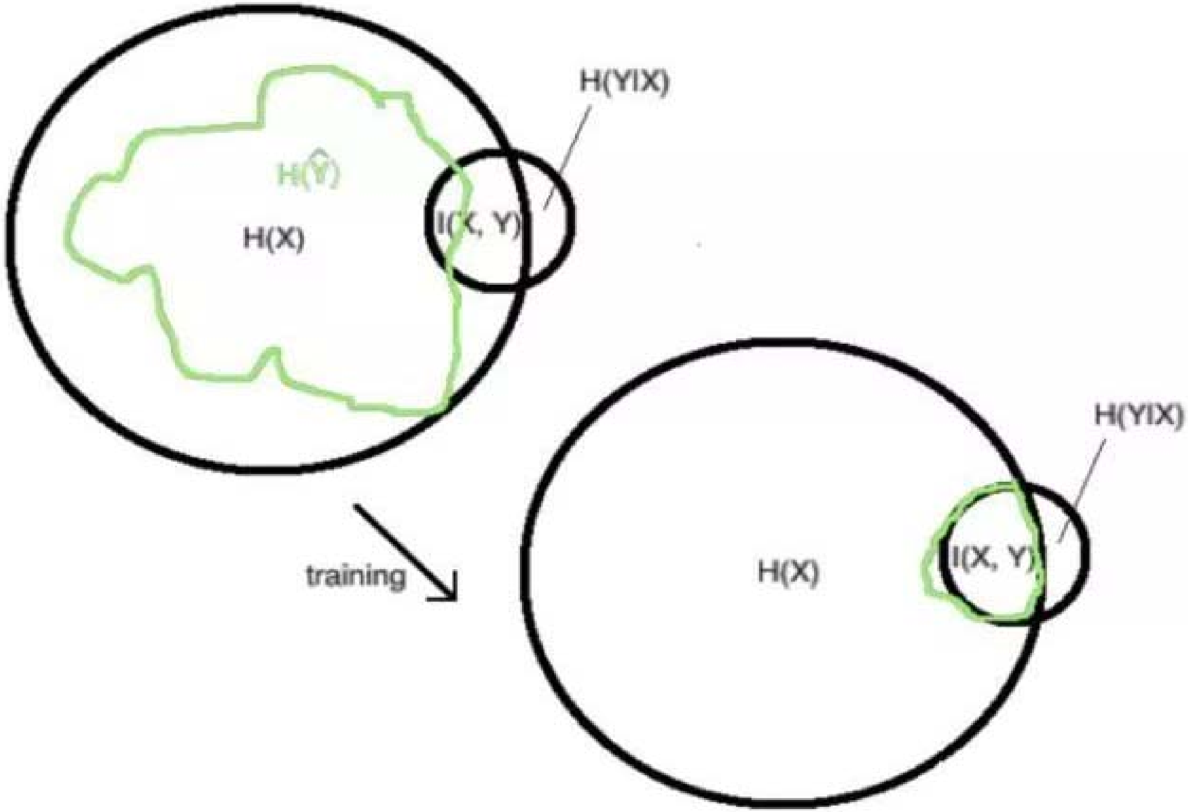
信息论相关概念

◻ 信息瓶颈(information bottleneck)

- 1999年，Naftali Tishby
- 2017年，用来解释深度学习的本质：拟合+挤压

$$\mathcal{L}_{IB} = -I(Y;\tilde{X}) + \beta I(X;\tilde{X})$$

假设***X***是一个复杂的数据集，就像一张狗的照片的像素，而***Y***是这些数据代表的一个更为简单的变量，比如单词“狗”。你可以任意压缩***X***而不丢失预测***Y***的能力，将***X***中所有与***Y***“相关”的信息捕获下来。



学习最重要的部分实际上是忘记

拟合+挤压

信息论可以精确定义“相关”

假设*X*是一个复杂的数据集，就像一张狗的照片的像素，而*Y*是这些数据代表的一个更为简单的变量，比如单词“狗”。你可以任意压缩*X*而不丢失预测*Y*的能力，将*X*中所有与*Y*“相关”的信息捕获下来。