Probabilistic Graphical Models

Nando de Freitas
September, 2012
University of British Columbia

Outline of the lecture

概率图建立

1. Joint distributions
2. The curse of dimensionality
3. Definition of a ***DAG (aka Bayesian network)***
4. Conditional independence in DAGs

概率图推断

5. marginalization and conditioning inference in DAGs using dynamic programming

经典概率图

6. HMM
7. Naïve Bayes

1. Joint distributions

针对某个任务的数据集包含了若干个随机变量，e.g., a, b, c, d, e。

这些随机变量的联合概率分布p(a, b, c, d, e) **全面地**反映了变量之间的关系。

知道了联合分布p(a, b, c, d, e)，我们就知道了现实世界中这几个变量关系的一切。

在统计推断(inference)中开启了“上帝视角”。

可以得出具有实际意义的，我们关心的任何条件分布p(a | b, c, e) 、单一变量分布p(d) 。

例如:	a	b	c	d	e
	沪深指数	标准普尔指数	纳斯达克指数	世界局势	市场走势

p(a, b, c, d, e)

p(市场走势=牛市|沪深指数=1500)

p(世界局势=恶化)

2. The curse of dimensionality

变量数提升(维度提升)对联合分布的计算(存储)带来的恶化

This curse tells us that to represent a joint distribution of ***d*** *binary variables, we need* $(2^d)-1$ *terms!*

联合概率分布的朴素表示：

给出每一种联合分布的情况，并给出对应的概率，组成一个如下的概率表。随着变量数目d的增加，组，表中条目将成指数级别增长，难以维持！

$d=3$

A	B	C	
0	0	0	$P(A=0, B=0, C=0)$
0	0	1	$P(A=0, B=0, C=1)$
0	1	0	⋮
0	1	1	
1	0	0	
⋮			

7

怎么办？

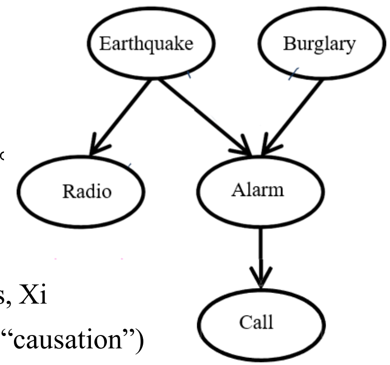
3. Directed Acyclic Graph (DAG)

已知变量：Earthquake Burglary Radio Alarm Call

如何求 $P(B, E, A, R, C)$?

首先，根据先验(领域)知识简化变量之间的依赖关系。

由此，我们得到了DAG(有向无环图)。



Nodes – random variables, Xi

Edges – direct influence (“causation”)

DAG假设：一个变量独立于除父结点之外的前辈结点、同辈结点。

X_i independent of $X_{ancestors}$ | $X_{parents}$

$$P(B, E, A, R, C)$$
$$= P(B)P(E|\cancel{B})P(A|B, E)P(R|\cancel{A}, \cancel{B}, E)P(C|\cancel{R}, A, \cancel{B}, \cancel{E})$$

初始想法是，一个变量只可能依赖于与它平级or前辈的结点。
但DAG告诉我们，一个变量只依赖于父结点，而与其他前辈独立。
P(C|A)，而不是P(C|A, E, B)。

3. Directed Acyclic Graph (DAG)

基于有向无环图的联合概率分布化简：

*The DAG tells us that if we have **n** variables x_i, the joint distribution of these variables **factorizes** as follows:*

$$P(x_{1:n}) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$$

一个变量只依赖父结点

一个复杂的联合概率化简为多个简单的 单一变量分布 和 条件概率分布。

DAG课堂笔记：

有向无环图考虑的是**局部**的信息关联，在很多inference中，变量只考虑和父亲节点的关系。

训练时只需要观测记录**局部**影响，我们再把这些局部关系拼接到一块时，做inference时就可以已知问到离它100步之外的**全局**问题。

拿数据，自动建立概率图，然后inference。这种全自动只在papers中。

图模型不是监督学习，不是从x学到y，不是学映射，而是学分布。图模型是说：如果我有很多数据，能不能把这些数据的联合分布以某种方式逼近，并不存在输入输出的关系。图模型不是监督、非监督、强化学习。角度不一样。

训练：在训练集中，用MLE逼近简化联合分布中的各个概率。频率学派，given a=某个值，c能取值的概率，就是counting。

图模型和神经网络结合VAE。变分推断下的auto encoder。通过图模型 让DL模型进行逻辑推断。

3. Directed Acyclic Graph (DAG)

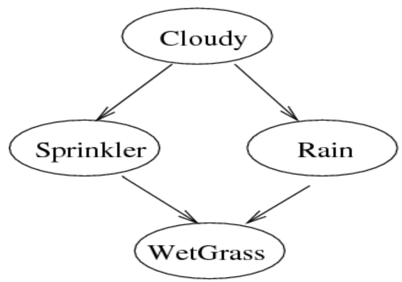
对于这个具有四个变量的联合分布来讲，我们可以选择存储一个完整的概率表(15个条目)
或者说，利用DAG给出变量之间的依赖关系简化计算、并存储各个简单概率表(9个条目)

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

c	s	r	w	prob
0	0	0	0	0.200
0	0	0	1	0.000
0	0	1	0	0.005
0	0	1	1	0.045
0	1	0	0	0.020
0	1	0	1	0.180
0	1	1	0	0.001
0	1	1	1	0.050
1	0	0	0	0.090
1	0	0	1	0.000
1	0	1	0	0.036
1	0	1	1	0.324
1	1	0	0	0.001
1	1	0	1	0.009
1	1	1	0	0.000
1	1	1	1	0.040

P(C=F)	P(C=T)
0.5	0.5

P(c)



条件

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

P(s|c)

P(R|c)

条件

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

条件

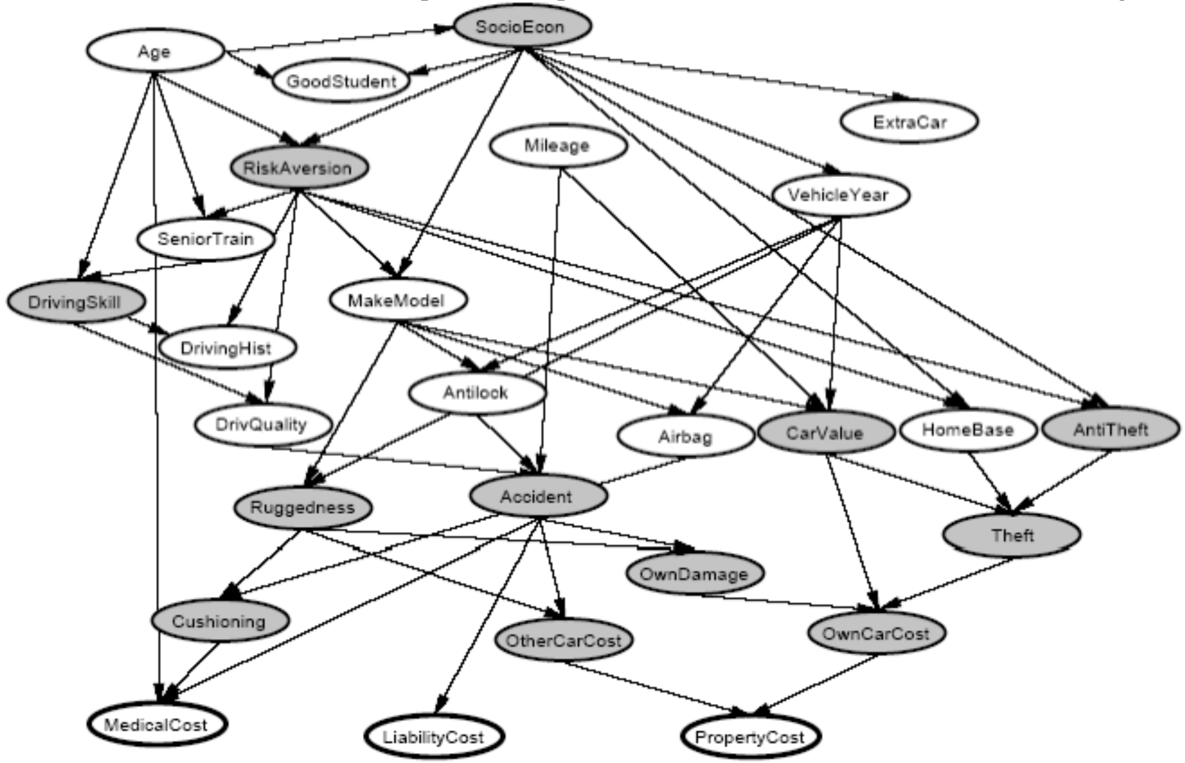
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

P(w|R,s)

将指数级别复杂度变成多项式级别复杂度。

Example: Vehicle insurance

http://www.cs.princeton.edu/courses/archive/fall10/cos402/assignments/bayes/



The 12 **shaded variables** are considered **hidden** or **unobservable**, while the other 15 are observable. The network has over 1400 parameters. An insurance company would be interested in predicting the bottom three "cost" variables, given all or a subset of the other observable variables.

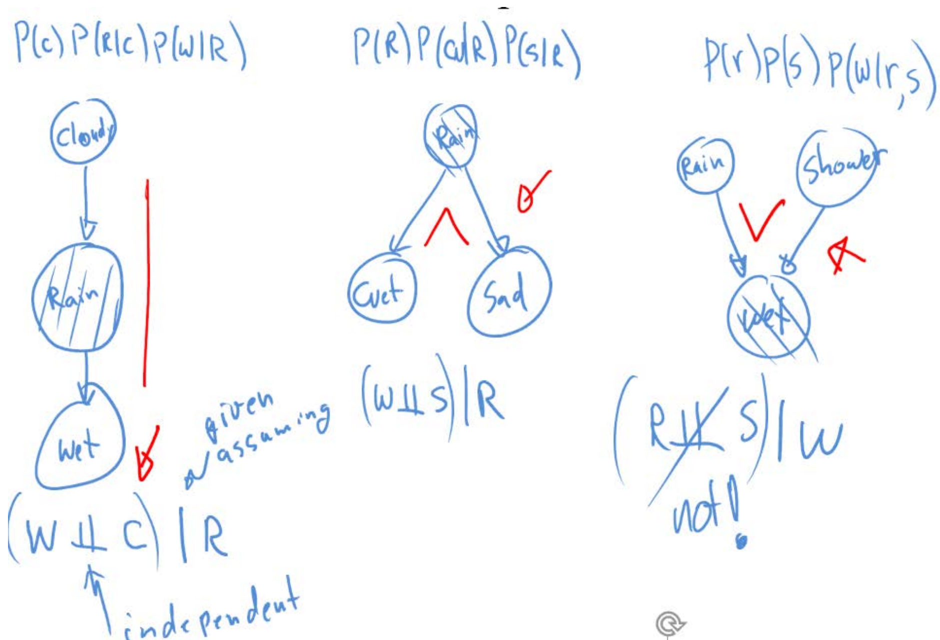
建立变量之间的依赖关系，计算联合分布，进行推断。

Unobservable的变量是否会影响概率的计算？

保单中的变量用概率图表示。在理赔中可以问各种问题。帮助计算风险。设计理财产品、理财方案。 given我这些结点被观测到，请问另一些结点的概率是多少？

4. Conditional independence in DAGs

3 cases of conditional independence to remember



DAG假设告诉我们：

Case1： Wet结点只依赖于父结点Rain，而与更久远的前辈结点独立；

Case2： Wet结点只依赖于父结点Rain，而与同辈份结点Sad独立；

Case3： Wet结点依赖于父结点Rain和Shower，但Rain和Shower不独立！

5. marginalization and conditioning inference in DAGs

5.1 利用DAG-based联合分布对*marginal probability* 的求解

Let us use 0 to denote false and 1 to denote true.

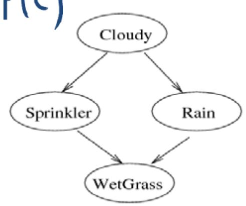
*What is the **marginal probability**, **P(S=1)**, that the sprinkler is on?*

$$P(S=1) = \sum_{C=0}^{1}\sum_{R=0}^{1}\sum_{W=0}^{1} P(C,R,W,S=1)$$

$$= \sum_{C}\sum_{R}\sum_{W} P(C)P(S=1|C)P(R|C)P(W|S=1,R)$$ 怎么算?

$P(C)$

P(C=F)	P(C=T)
0.5	0.5



$P(S|C)$

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

$P(R|C)$

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

$P(W|R_S)$

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

方法一：“朴素”的求解方法，挨个求和项->对应的C,R,W取值->查表并计算

Brute force (exponential) approach

What is the ***marginal probability****,* ***P(S=1)****, that the sprinkler is on?*

$$\sum_{C} \sum_{R} \sum_{W=0}^{1} P(W|S=1,R)\, P(S=1|C)\, P(R|C)\, P(C)$$

000
$$= \overset{0.1}{P(W=0|S=1,R=0)} \times \overset{0.5}{P(S=1|C=0)} \times \overset{0.8}{P(R=0|C=0)} \times \overset{0.5}{P(C=0)}$$

001
$$+ P(W=1|S=1,R=0)\, P(S=1|C=0)\, P(R=0|C=0)\, P(C=0)$$

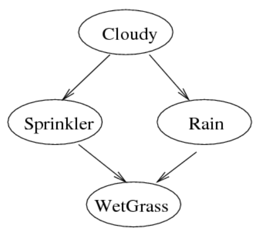
010
$$+ P(W=0|S=1,R=1)\, P(S=1|C=0)\, P(R=1|C=0)\, P(C=0)$$

011
$$+ \cdots$$

$$\vdots$$

111
$$+$$

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Brute force (exponential) approach

What is the ***marginal probability****,* ***P(S=1)****, that the sprinkler is on?*

```
PROD
FOR R=0:1:1
    FOR C=0:1:1
        FOR W=0:1:1
            PROD = PROD + P(C)P(R|C)P(S|C)P(W|S=1,R)
        END
    END
END
```

程序实现需要三重循环，效率感人。接下来介绍利用动态规划的解法。

方法二：消元法 or 动态规划 or 乘法分配律

variable elimination, *aka* dynamic programming, *aka* distributive law

What is the ***marginal probability***, ***P(S=1)***, *that the sprinkler is on?*

当R C固定时，

$P(R|C)\,P(C)\,P(S=1|C)$

固定。

$\sum_{W} P(W|S=1,R)\,P(R|C)\,P(C)\,P(S=1|C)$

相当于

$\sum_{W}$(P(W|S=1,R)*固定值)

所以，将固定值的部分提到求和外。

$$P(S=1) = \sum_{C}\sum_{R}\sum_{W} P(W|S=1,R)\,P(R|C)\,P(C)\,P(S=1|C)$$

$$= \sum_{C}\sum_{R} P(R|C)\,P(C)\,P(S=1|C)\boxed{\sum_{W} \cancel{P(W|S=1,R)}} \sim \phi = 1$$

$$= \sum_{C} P(C)\,P(S=1|C)\boxed{\sum_{R} \cancel{P(R|C)}} \sim \psi = 1$$

$$= P(S=1|C=0)\,P(C=0) + P(S=1|C=1)\,P(C=1)$$

$$= 0.3$$

What is the ***marginal probability****,* ***P(S=1)****, that the sprinkler is on?*

$\psi = 0$

$\phi = 0$

$\Theta = 0$

FOR $W = 0:1:1$

$\quad \phi_R = \phi_R + P(W \mid S=1, R)$

END

FOR $R = 0:1:1$

$\quad \psi_C = \psi + P(R \mid C)\,\phi_R$

END

FOR $C = 0:1:1$

$\quad \Theta = \Theta + P(S=1 \mid C)\,P(C)\,\psi_C \longrightarrow \Theta = 0.3$

END

程序实现没有for循环的叠加了，效率大幅提高。

动态规划课堂笔记:

凡是问题有**递归**的思想，考虑用动态规划。

递归类似于数学归纳法。不断向前归于简单问题的求解。

机器学习中的动态规划:
1. 概率图
2. HMM
3. VC维（刻画模型复杂度的一种方式，机器学习理论的基石）
4. 强化学习贝尔曼方程MDP
5. BP的DP实现

5.2 利用DAG-based联合分布对***posterior probability***的求解

What is the ***posterior probability***, ***P(S=1|W=1)***, *that the sprinkler is on given that the grass is wet?*

关键在于利用贝叶斯公式将条件概率进行转化：

$$P(S=1|W=1) = \frac{P(S=1, W=1)}{P(W=1)}$$

由此，就可利用5.1中的动态规划思想求解：

$$P(W=1) = \sum_S \sum_C \sum_R P(S, W=1, C, R)$$

$$\underline{P(S=1, W=1)} = \sum_C \sum_R \underline{P(S=1, W=1, C, R)}$$

5.2 利用DAG-based联合分布对***posterior probability***的求解

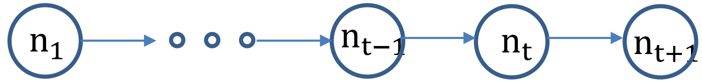
What is the ***posterior probability****,* ***P(S=1|W=1,R=1)****, that the sprinkler is on given that the grass is wet and it is raining?*

$$P(S=1|W=1,R=1) = P(S=1\,|\,(W=1,R=1))$$

$$= \frac{P(S=1,W=1,R=1)}{P(W=1,R=1)}$$

6. HMM

马尔可夫性：下一刻仅取决上一时刻状态。



Given n_t， $n_{t+1} \perp$ (独立于)$n_{t-1}, \dots, n_1$

Outline

概率图是得到"研究的变量之间联合概率分布"的有效手段。

那么,如何用概率图模型(or贝叶斯网)来进行 inference? or 如何利用概率图计算

This lecture is devoted to the problem of inference in probabilistic graphical models (aka Bayesian nets). The goal is for you to:

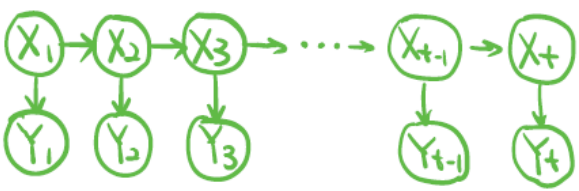
利用 动态归化　　利用 Bayes rule + 动态规化.

- ❑ Practice marginalization and conditioning.
- ❑ Practice Bayes rule.

之前内容

- ❑ Learn the HMM representation (model).
- ❑ Learn how to do **prediction** and **filtering** using HMMs

HMM: 用于时序分析的概率图.



如何计算HMM问题中最关心的概率

$$P(X_t \mid Y_{1:t})$$

即, 基于t时刻之前所有观测 $Y_{1:t}$

来预测t时刻隐变量 X_t 出现概率

方法为 optimal filtering, 分为两步:

① prediction

② filtering

6.1 实例:单一时间点的概率图

Assume you have a little robot that is trying to ***estimate*** *the* ***posterior*** *probability that you are* ***happy*** *or* ***sad****, given that the robot has observed whether you are* ***watching Game of Thrones (w)****,* ***sleeping (s)****,* ***crying (c)*** *or* ***face booking (f)****.*

Let the ***unknown state*** *be* ***X=h*** *if you're happy and* ***X=s*** *if you're sad.*

Let ***Y*** *denote the* ***observation****, which can be* ***w****,* ***s****,* ***c*** *or* ***f*** *.*

We want to answer queries, such as:

$P(X=h|Y=f)$?

$P(X=s|Y=c)$?

6.1 实例:单一时间点的概率图（续）

变量: X 心情 Y 行为　　画DAG	计算

心情是隐状态,不可知,只能预测
(当然,在训练集中,已经给出心情用于训练)

假设 x y 的关系为
心情决定要做的事.

Robot 只能观察到外在状态
Robot 想通过观察看本质.

$P(x\mid y)$?

根据概率图,写出
$P(x,y)=P(x)\cdot P(y\mid x)$

$P(x\mid y)$?

训练集中可观测到

$P(x)=$

X: S	H
0.2	0.8

$P(x=s)=0.2$

训练集中可观测到

$P(y\mid x)=$

X \ Y	w	s	c	f	
S	0.1	0.3	0.5	0.1	=1
H	0.4	0.4	0.2	0	=1

$$P(x=H\mid y=w)=\frac{\overset{0.4}{P(y=w\mid x=H)}\,\overset{0.8}{P(x=H)}}{\underbrace{\overset{0.4}{P(y=w\mid x=H)}\,\overset{0.8}{P(x=H)}+\overset{0.1}{P(y=w\mid x=s)}\,\overset{0.2}{P(x=s)}}_{=P(y=w)}}$$

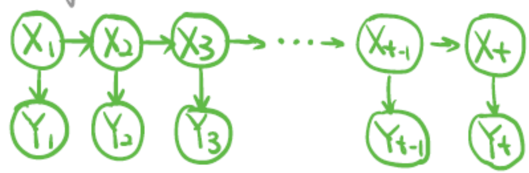
6.2 HMM：用于时序分析的概率图模型

*But what if instead of an absolute prior, what we have instead is a temporal (**transition prior**). That is, we assume a **dynamical system***

隐变量状态转移矩阵

$$\begin{array}{cc|cc} & & \multicolumn{2}{c}{X_2} \\ & & S & H \\ \hline X_1 & S & 0.99 & 0.01 \\ & H & 0.1 & 0.9 \end{array}$$

具有时序关联的Xi Yi的DAG



求：*Given a history of observations, say **Y_1=w, Y_2=f, Y_3=c**, we want to compute the posterior distribution that you are happy at step 3. That is, we want to estimate:*

$$P(X_3=h \mid Y_1=w, Y_2=f, Y_3=c)$$

*Clearly, to know if you're happy when crying it helps to know if the sequence of observations is **wcw** or **ccc**.*

$Y_1 Y_2 Y_3$

6.2.1 HMM的联合概率分布

In general, we assume we have an ***initial*** *distribution* $\boldsymbol{P(X_0)}$, *a* ***transition*** *model* $\boldsymbol{P(X_t | X_{t-1})}$, *and an* ***observation*** *model* $\boldsymbol{P(Y_t | X_t)}$.

①随时序数据已知的分布

②变量之间的依赖关系. or DAG or 概率图.

隐变量状态的先验分布

$$P(x_0) = \begin{array}{c} \begin{array}{cc} S & H \end{array} \\ \begin{bmatrix} .2 & .8 \end{bmatrix} \end{array}$$

隐变量状态(在不同时刻)的转移矩阵

$$P(x_t | x_{t-1}) = \begin{array}{c|cc} & S & H \\ \hline S & .9 & .1 \\ H & 0 & 1 \end{array}$$

经典HMM的状态转移矩阵非时变

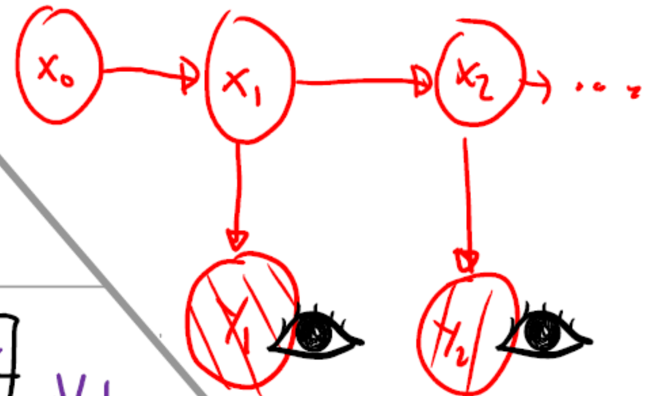
(在一个时刻)心情指导行为的矩阵

$$P(y_t | x_t) = \begin{array}{c|cccc} & W & S & C & F \\ \hline S & 0.3 & 0.3 & 0.2 & 0.2 \\ H & 1 & 0 & 0 & 0 \end{array}$$

emmison matrix.

$\forall t$

非时变



$P(y_1|x_1) = P(y_2|x_2)$

for 2 time steps

各分布均已知

$$P(x_0, x_1, x_2, y_1, y_2) = P(x_0) P(y_1|x_1) P(y_2|x_2) P(x_1|x_0) P(x_2|x_1)$$

当得到联合分布的概率图表示后，利用其进行inference.如下.

6.2.2 HMM的一个经典query P(Xt|Y1:t)的动态规划求解，即 Optimal Filtering

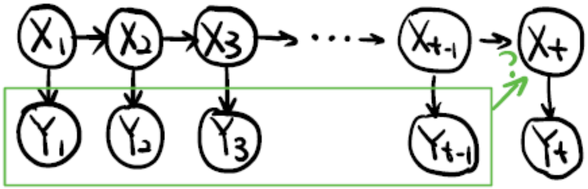
在HMM中，最常见的query是基于所有历史观测 $Y_{1\sim n}$ 来推断当前隐变量状态 X_n

Our goal is to compute, ***for all t****, the posterior (aka* ***filtering****) distribution:*

$$P(X_t \mid Y_{1:t}) = P(X_t \mid Y_1, Y_2, \ldots, Y_t)$$

求解这个inference有专门的名字，叫OF.

We derive a ***recursion*** 递归 *to compute* $P(X_t \mid Y_{1:t})$ *assuming that we have as input* $P(X_{t-1} \mid Y_{1:t-1})$. *The recursion has two steps:* ***prediction*** *and* ***Bayesian update****.*



STEP 1:
Prediction

$$P(A|C) = \sum_B P(AB|C)$$

First, we compute the state prediction: $P(X_t | Y_{1:t-1})$

$$P(AB|C) = P(A|BC)P(B|C)$$

已知 1:t-1 的观测，预测 t 时刻隐~~状~~态 X_t.

加一项 X_{t-1}，再积掉，等于没加一样

$$P(x_t|y_{1:t-1}) = \sum_{x_{t-1}\in\{H,S\}} P(x_t, x_{t-1}|y_{1:t-1})$$

计算的理论依据.

$$= \sum_{x_{t-1}} P(x_t|x_{t-1}, \cancel{y_{1:t-1}})\, P(x_{t-1}|y_{1:t-1})$$

概率图的假设前提：如果有父结点出现，则 X_t 与 $y_{1:t-1}$ 独立。即，认为父结点 X_{t-1} 涵盖了所有关联.

$$= \sum_{x_{t-1}} \underbrace{P(x_t|x_{t-1})}_{\text{转移矩阵，已知}}\ \underbrace{P(x_{t-1}|y_{1:t-1})}_{\text{递归假设中，已知}}$$

由此，就完成了预测，即基于 1:t-1 的观测 $y_{1:t-1}$，预测 t 时刻隐状态 X_t.

Bayes rule revision

$$P(A|BC) = P(A|CB)$$

$$= \frac{P(ABC)}{P(BC)} = \frac{P(B|AC)P(A|C)P(C)}{P(B|C)\,P(C)}$$

$$= \frac{P(B|AC)P(A|C)}{P(B|C)}$$

STEP 2:
Bayes update

Second, we use Bayes rule to obtain $\boldsymbol{P(X_t \mid Y_{1:t})}$

$$P(x_t \mid y_{1:t}) = P(\overset{A}{x_t} \mid \overset{B}{y_t}, \overset{C}{y_{1:t-1}})$$

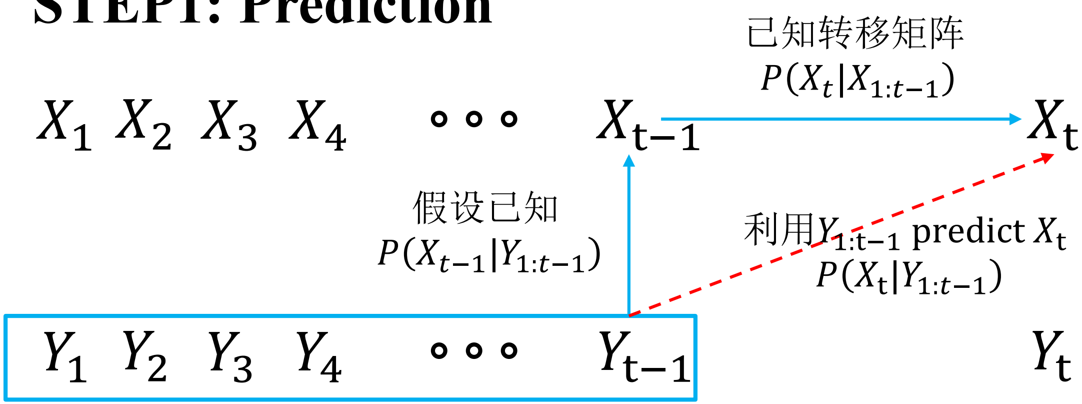
$$= \frac{P(y_t \mid x_t, \cancel{y_{1:t-1}})\, P(x_t \mid y_{1:t-1})}{\sum_{x_t} P(y_t \mid x_t, y_{1:t-1})\, P(x_t \mid y_{1:t-1})}$$

emmison matrix prediction

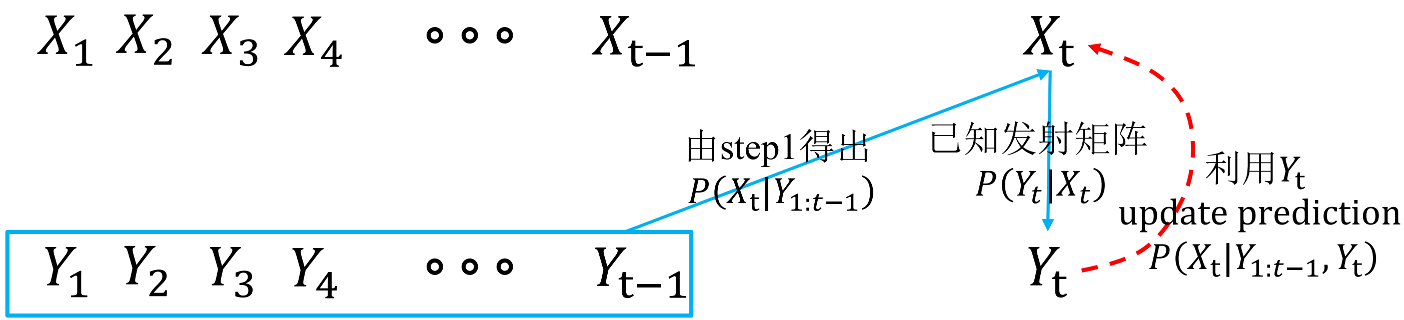
$$= \frac{P(y_t \mid x_t)\, P(x_t \mid y_{1:t-1})}{\sum_{x_t} P(y_t \mid x_t)\, P(x_t \mid y_{1:t-1})}$$

6.2.3 Optimal filtering分步图解

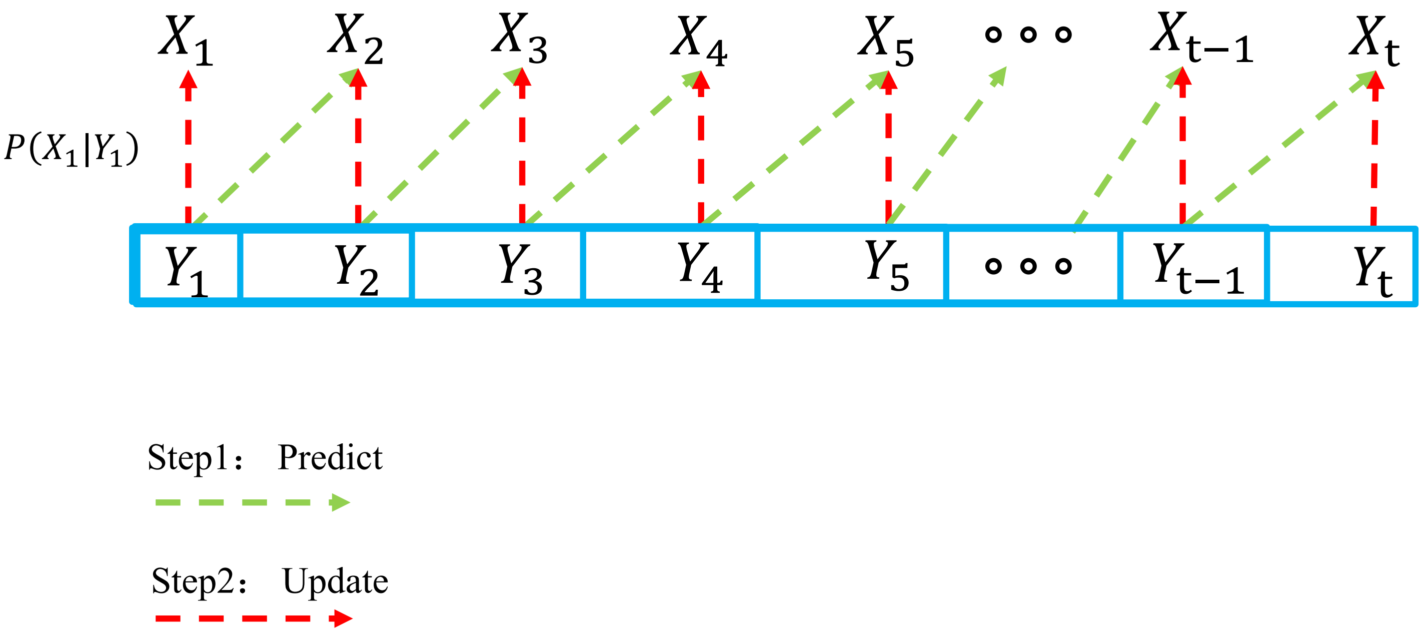
STEP1: Prediction



STEP2: Update



6.2.4 Optimal filtering整体流程



HMM algorithm

$\boxed{\text{Example.}}$

Assume $P(X_{t-1}|Y_{1:t-1}) =$

X_{t-1}: S	H
.2	.8

$P(X_t | X_{t-1}) =$

X_{t-1} \ X_t	S	H
S	1	0
H	0.1	0.9

$P(Y_t|X_t) =$

X_t \ Y_t	W	S	C	F
S	.1	.2	.3	.4
H	0.9	0	0	0.1

Then, we can predict the state X_t (at time t) given the observations $Y_1, Y_2, \cdots Y_{t-1}$. Specifically

$$P(X_t = H | Y_{1:t-1}) = \sum_{X_{t-1}} P(X_t = H | X_{t-1}) P(X_{t-1}|Y_{1:t-1})$$

$$= P(X_t = H | X_{t-1} = S) P(X_{t-1} = S | Y_{1:t-1}) + P(X_t = H | X_{t-1} = H) P(X_{t-1} = H | Y_{1:t-1})$$

$$= (0)(0.2) + (0.9)(0.8) = 0.72$$

$$P(X_t = S | Y_{1:t-1}) = 1 - P(X_t = H | Y_{1:t-1}) = 0.28$$

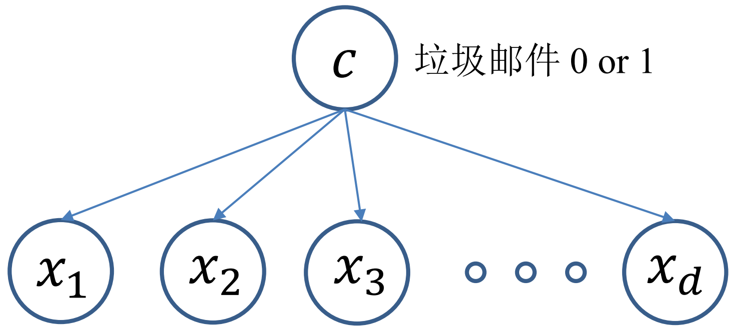
If we observe $Y_t = W$, The Bayes update yields:

$$P(X_t = H | Y_{1:t-1}, Y_t = W) = \frac{P(Y_t = W | X_t = H) P(X_t = H | Y_{1:t-1})}{P(Y_t = W | X_t = H) P(X_t = H | Y_{1:t-1}) + P(Y_t = W | X_t = S) P(X_t = S | Y_{1:t-1})}$$

$$= (0.9)(0.72) \,/\, [(0.9)(0.72) + (0.1)(0.28)]$$

7. Naïve Bayes

对于一个现实场景，如果变量之间的依赖关系可以简化为如下的概率图，即，“各个变量 x_i $(i = 1\sim d)$ 仅与变量c有关，而 x_i 之间相互独立”，那么，这是一个朴素贝叶斯问题。



中文所有30000个单词 “我” “旅游” “发票” “低价”

每一封邮件相当于一个长度为30000的向量，每个位置表示对应单词出现频率 (4, 1, 10, ... , 2)

注：朴素贝叶斯的假设将问题简化了。因为在一封垃圾邮件中，各个单词之间是有关联的，例如，“高价. 收. 发票” “恭喜. 你. 中奖”

由概率图，联合分布为：

$$P(c, x_1, x_2, x_3, \ldots, x_d) = P(c)\, P(x_1|c)\, P(x_2|c)\, P(x_3|c) \ldots P(x_d|c)$$

由此，我们可以求，一封邮件$\vec{x} = (x_1, x_2, x_3, \ldots, x_d)$是垃圾邮件的概率：$P(\mathrm{c}|\vec{x})$

$$P(\mathrm{c}|\vec{x}) = \frac{P(\vec{x}, c)}{P(x)} = \frac{P(c)\, P(x_1|c)\, P(x_2|c)\, P(x_3|c) \ldots P(x_d|c))}{P(x)}$$

$$P(c1|\vec{x}) = \frac{P(\vec{x},c1)}{P(x)} = \frac{\boxed{P(c1)}\ \boxed{P(x_1|c1)}\ P(x_2|c1)\ P(x_3|c1)\ \ldots\ P(x_d|c1))}{\boxed{P(x)}}$$

$$P(c0|\vec{x}) = \frac{P(\vec{x},c0)}{P(x)} = \frac{P(c0)\ P(x_1|c0)\ P(x_2|c0)\ P(x_3|c0)\ \ldots\ P(x_d|c0))}{P(x)}$$

邮件$\vec{x}$的类别将取决于两者的大小，即

$$h_{nb}(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\max}\ P(c) \prod_{i=1}^{d} P(x_i \mid c)\ , \qquad (7.15)$$

这就是朴素贝叶斯分类器的表达式.

均可以从训练集中得到：
$P(c1)$垃圾邮件在训练集中的频率；
$P(x_1|c1)$在垃圾邮件中，单词x_1出现的频率；
$P(x)$不需要求解；

也就是说，求这些概率时，就是用的频率学派的数个数counting。

关于朴素贝叶斯的具体例子、拉普拉斯修正、半朴素贝叶斯，见教材。

无论是朴素贝叶斯还是半朴素贝叶斯，其都是利用训练集来估计出 需要的概率。而估计这些概率的方式，其实就是数个数。

贝叶斯分类器是一种基于概率计算的分类方式。与我们通常意义的优化算法是不同的。