CSCI567 Machine Learning (Fall 2017)

Prof. Fei Sha

U of Southern California

Lecture on Nov. 7, 2017

Outline

1. Administration
2. Review of last lecture
3. Graphical models

Outline

1. Administration
2. Review of last lecture
3. Graphical models

Schedule change

- Quiz 2 is this Thursday (for most of you).
- Please do not forget your Homework 4 Programming Component.

Outline

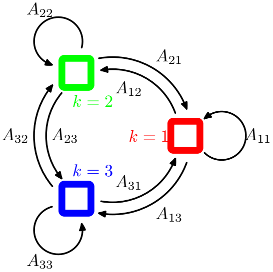
1. Administration
2. Review of last lecture
3. Graphical models

A Markov process

Evolving states form a Markov chain

state transition diagram

Ex: for 3 possible states



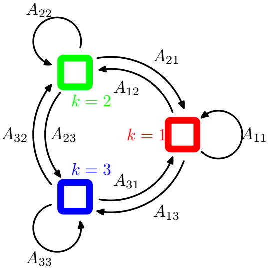
Transition probability matrix

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

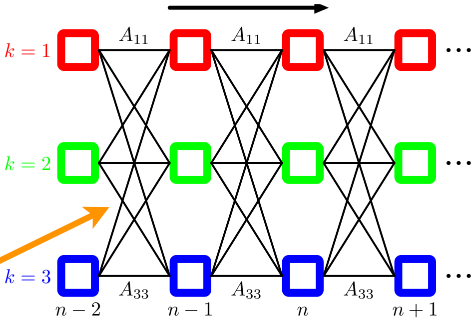
$$A_{ij} \geq 0 \qquad \sum_j A_{ij} = 1$$

Lattice/Trellis

Unfolding state transition



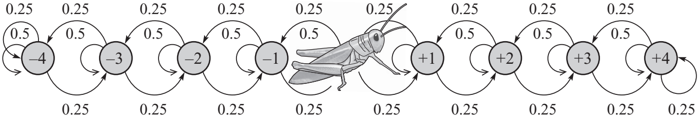
time



each "slices" is repeated

each column represents the state variable at time (n-2), (n-1), n and (n+1)

Grasshopper's move as Markov chain



If the grasshopper keeps hopping, where it would be?

states (ie, location x): 0, 1, 2, 3, 4, -1, -2, -3, -4

transition: P(i →i) = 0.5, P(i → i+1) = 0.25, P(i → i-1) = 0.25

initial probability: $\pi_0(x)$ = {0.9, 0.05, 0,0,0, 0.05,0,0,0,0}

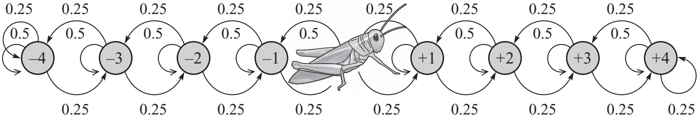
a probabilistic distribution over all location P(x)

How to get P(x) at time t?

previous location

$$P_t(x) = \sum_{x'} P_t(x, x') = \sum_{x'} P(x|x')P_{t-1}(x')$$

Grasshopper, where are you?



Infer where it is at any time t

given a distribution over initial positions, computing $P(s_t)$ is trivial

we cannot see where it is exactly: jumping too fast!

But can we hear where it is?

hearing grasshopper sing

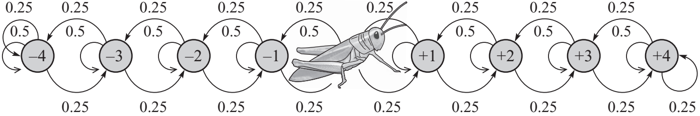
At time t

loud: probably close to me (position 3 or 4?)

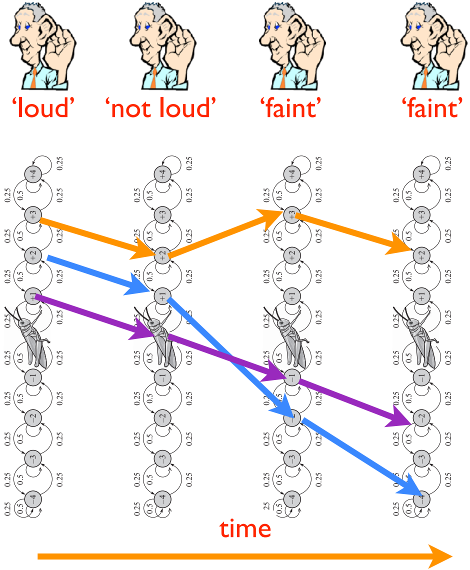
not loud: not close to me (position 1 or 2 or -1 or -2 or 0?)

faint: probably far to me (position -3 or -4?)

Our ears are not radar: so our guessing might be a bit off



let us get more information



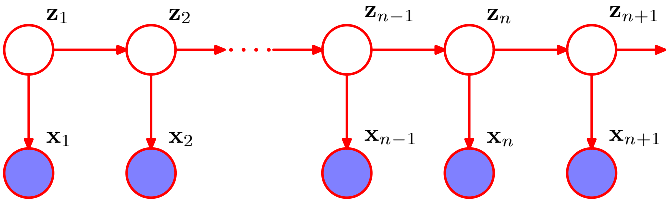
Which path is more likely?

need to integrate two sets of probabilistic information

state transition

observation (hearing)

Formally



Hidden Markov model definition

states: as before (denoted with s_n, s_t, or z_n, z_t), but hidden (hollow circles)

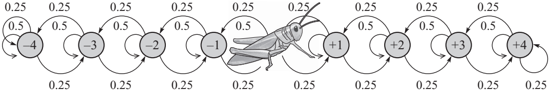
observations: x_t or $\mathbf{x}_t$

define a joint distribution

observation model

$$P(\{x_n\},\{z_n\}) = \pi_0(z_1)\prod_{n=2}^{N} p(z_n|z_{n-1})\prod_{n=1}^{N} p(x_n|z_n)$$

Example: grasshopper



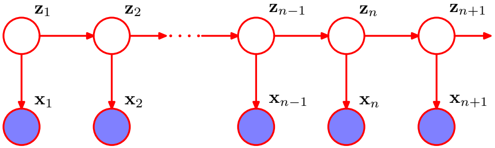
$$p(x_n|z_n)$$

	loud	not so	faint
-4	0	0.2	0.8
-3	0	0.3	0.7
-2	0	0.5	0.5
-1	0	0.6	0.4
0	0	0.7	0.3
1	0	0.8	0.2
2	0.1	0.8	0.1
3	0.7	0.3	0
4	0.8	0.2	0

$$p(z_n|z_{n-1})$$

	-4	-3	-2	-1	0	1	2	3	4
-4	0.75	0.25	0	0	0	0	0	0	0
-3	0.25	0.5	0.25	0	0	0	0	0	0
-2	0	0.25	0.5	0.25	0	0	0	0	0
-1									
0									
1									
2									
3									
4	0	0	0	0	0	0	0	0.25	0.75

Inference problems in HMMs



Marginal

$$P(x_1, x_2, \ldots, x_N)$$

Filtering

$$P(z_n | x_1, x_2, \ldots, x_n)$$

Smoothing

$$P(z_n | x_1, x_2, \ldots, x_T)$$

Most likely path

$$P(z_1, z_2, \ldots, z_T | x_1, x_2, \ldots, x_T)$$

Other types and applications of HMMs

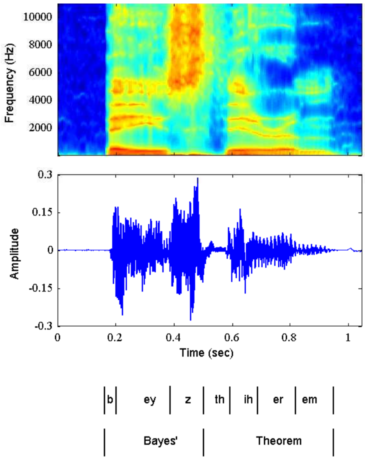
Discrete HMMs

state and observations are discrete:

grasshopper: state (finite grids), observations (‘loud’, ‘not loud’, ‘faint’)

Discrete state but continuous observation HMMs

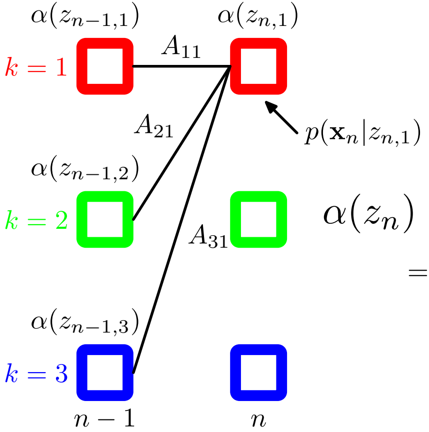
eg. automatic speech recognition



Continuous state and continuous observations

eg. Kalman filtering (used in control and signal processing)

The forward message

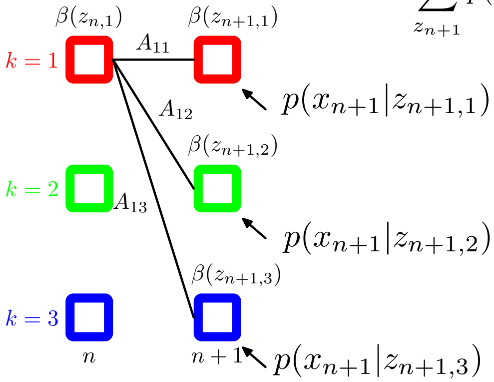


$$\alpha(z_n) = p(x_1, x_2, \ldots, x_n, z_n)$$
$$= p(x_n|z_n) \sum_{z_{n-1}} p(z_n|z_{n-1}) \alpha(z_{n-1})$$

This is the same as in the previous lecture $\alpha_t(j)$

the backward message

$$\begin{aligned}\beta(z_n) &= p(x_{n+1}, x_{n+2}, \ldots, x_N | z_n) \\ &= \sum_{z_{n+1}} p(x_{n+1}|z_{n+1}) p(z_{n+1}|z_n) \beta(z_{n+1})\end{aligned}$$



This is the same as in the previous lecture $\beta_t(j)$

Most likely path $P(z_1, z_2, \ldots, z_T | x_1, x_2, \ldots, x_T)$

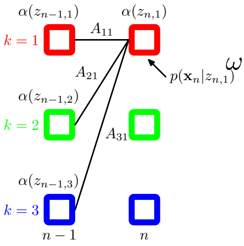
This is called Viterbi decoding

this will tell us about where the grasshopper is likely to be at different time

Replace the forward message from "sum" to "max"

which value of

z_{n-1} made its max



$$\omega(z_n) = p(x_n|z_n) \max_{z_{n-1}} p(z_n|z_{n-1})\omega(z_{n-1})$$

Do not forget your trackback table

time n	time n-1
$z_{n,1}$	$z_{n-1,2}$
$z_{n,2}$	$z_{n-1,1}$
$z_{n,3}$	$z_{n-1,3}$

This is the same as in the previous lecture $\delta_t(j)$

Marginals, filtering and smoothing

Marginals

$$P(x_1, x_2, \ldots, x_N) = \sum_{z_N} \alpha(z_N) = \sum_{z_1} \beta(z_1) = \sum_{z_n} \alpha(z_n)\beta(z_n)$$

Filtering

$$P(z_n | x_1, x_2, \ldots, x_n) \propto \alpha(z_n)$$

$$P(z_n | x_{n+1}, x_2, \ldots, x_N) \propto \beta(z_n)$$

Smoothing

$$P(z_n | x_1, x_2, \ldots, x_N) \propto \alpha(z_n)\beta(z_n)$$

Because of the use of forward/backward messages, this procedure is called forward-backward (FB) algorithm

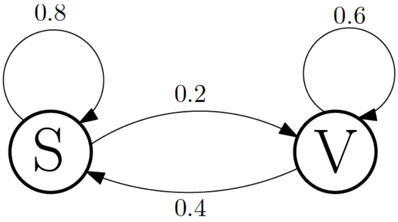
What is the computational complexity?

Example (from Dr. Parisa M.)

Consider the HMM below. In this world, every time step (say every few minutes), you can either be Studying or playing Video games. You're also either Grinning or Frowning while doing the activity.

E	$p(E\|X=S)$
grin	0.5
frown	0.5

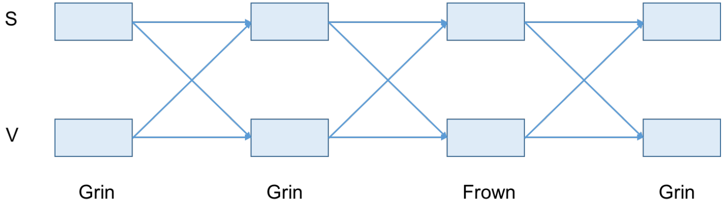
E	$p(E\|X=V)$
grin	0.8
frown	0.2



What are you doing most likely?

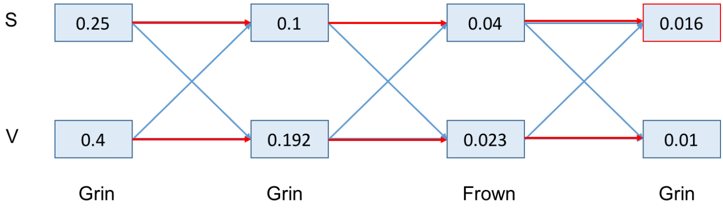
Suppose that we believe that the initial state distribution is 50/50. We observe: Grin, Grin, Frown, Grin. Run the Viterbi algorithm by filling in the values of the lattice below.
What is the most likely path for this sequence of observations?

Trellis



Solution

The final solution is as given in the following figure:



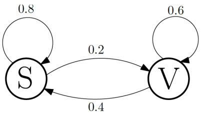
The most likely sequence is SSSS – the following slides detail how you do that.

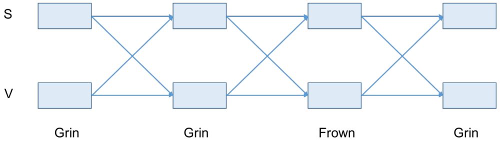
$t = 1$, the initial time

Observation at $t = 1$ is 'Grin'

E	$p(E\|X=S)$
grin	0.5
frown	0.5

E	$p(E\|X=V)$
grin	0.8
frown	0.2





$$\delta_1('S') = p(x_1 =' Grin' | z_1 =' S')\pi(z_1 =' S') = 0.5 \times 0.5 = 0.25 \quad (1)$$
$$\delta_1('V') = p(x_1 =' Grin' | z_1 =' V')\pi(z_1 =' V') = 0.8 \times 0.5 = 0.4 \quad (2)$$

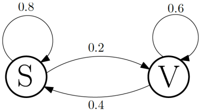
Note. We do not need trace-back table for this initial step.

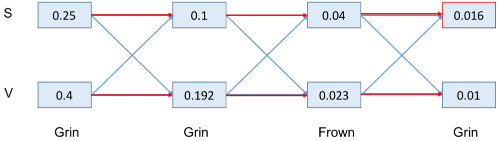
$t = 2$

Observation at $t = 2$ is 'Grin'

E	$p(E\|X = S)$
grin	0.5
frown	0.5

E	$p(E\|X = V)$
grin	0.8
frown	0.2





$$\delta_2('S') = \max\{p(x_2 =' Grin'|z_2 =' S')p(z_2 =' S'|z_1 =' S')\delta_1('S'), \quad (3)$$

$$p(x_2 =' Grin'|z_2 =' S')p(z_2 =' S'|z_1 =' V')\delta_1('V')\} \quad (4)$$

$$= \max\{0.5 \times 0.8 \times 0.25, 0.5 \times 0.4 \times 0.4\} = 0.01 \quad (5)$$

$$\delta_2('V') = \max\{p(x_2 =' Grin'|z_2 =' V')p(z_2 =' V'|z_1 =' S')\delta_1('S'), \quad (6)$$

$$p(x_2 =' Grin'|z_2 =' V')p(z_2 =' V'|z_1 =' V')\delta_1('V')\} \quad (7)$$

$$= \max\{0.04, 0.192\} = 0.192 \quad (8)$$

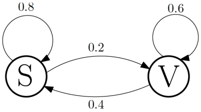
$t = 2$	$t = 1$
S	S
V	V

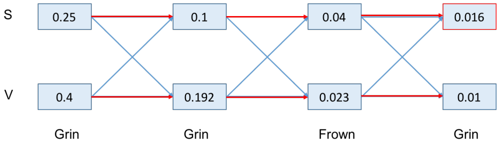
$t = 3$

Observation at $t = 3$ is 'Frown'

E	$p(E\|X = S)$
grin	0.5
frown	0.5

E	$p(E\|X = V)$
grin	0.8
frown	0.2





$$\begin{aligned}
\delta_3('S') &= \max\{p(x_3 =' Frown'|z_3 =' S')p(z_3 =' S'|z_2 =' S')\delta_2('S'), && (9)\\
&\quad p(x_3 =' Frown'|z_3 =' S')p(z_3 =' S'|z_2 =' V')\delta_2('V')\} && (10)\\
&= \max\{0.5 \times 0.8 \times 0.1, 0.5 \times 0.4 \times 0.192\} = 0.04 && (11)\\
\delta_3('V') &= \max\{p(x_3 =' Frown'|z_3 =' V')p(z_3 =' V'|z_2 =' S')\delta_2('S'), && (12)\\
&\quad p(x_3 =' Frown'|z_3 =' V')p(z_3 =' V'|z_2 =' V')\delta_2('V')\} && (13)\\
&= \max\{?, 0.023\} = 0.023 && (14)
\end{aligned}$$

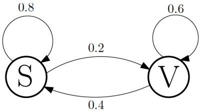
$t = 3$	$t = 2$
S	S
V	V

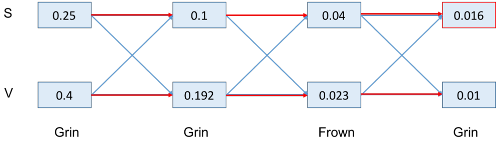
$t = 4$

Observation at $t = 4$ is 'Grin'

E	$p(E \mid X = S)$
grin	0.5
frown	0.5

E	$p(E \mid X = V)$
grin	0.8
frown	0.2





$$\delta_4('S') = \max\{0.016, ?\} = 0.016 \quad (15)$$

$$\delta_4('V') = \max\{?, 0.01\} = 0.01 \quad (16)$$

$t = 4$	$t = 3$
S	S
V	V

Last step, since $\delta_4('S') > \delta_4('V')$, so we choose

$$z_4^* = S$$

Then $z_3^* = S$, then $z_2^* = S$ and then $z_1^* = S$, using the trace-back tables.

All good, but

what if we do not know the model parameters

model parameters: initial distribution, transition model, and observation model

Learning parameters

easy: if we have access to all data, not only the observations but also the hidden states!

what if hidden states are not known to us?

we are dealing with the problem of learning with incomplete data!

Estimate parameters with complete data

Suppose that we didn't know the emission probabilities or transition probabilities for this HMM. Instead, we had to estimate them from data. Consider the following data set:

```
                               1 1 1 1 1 1 1 1 1 1 2
 time: 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
       ---------------------------------------
state: S S V V V S S S S S V S V V S V S S V V
  obs: G F G G F F F F G F G G G G F G F F G G
```

Based on this data, estimate the emission and the transition probabilities for this HMM.

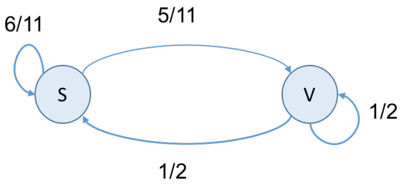
Solution

```
                           1 1 1 1 1 1 1 1 1 1 2
 time: 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
       ---------------------------------------
state: S S V V V S S S S S V S V V S V S S V V
  obs: G F G G F F F F G F G G G G F G F F G G
```

Solution

```
                                   1 1 1 1 1 1 1 1 1 1 2
time: 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
      ---------------------------------------
state: S S V V V S S S S S V S V V S V S S V V
  obs: G F G G F F F F G F G G G G F G F F G G
```

$S \to S : 6,\ S \to V : 5,\ V \to S : 4,\ V \to V : 4$



And the emissions: $S \to g : 3,\ S \to f : 8,\ V \to g : 8,\ V \to f : 1$

EM solution

Step 0 Random guess a θ^0; set t= 0

Step 1 (E-Step)

Compute following posterior probabilities

$$\gamma_n = p(z_n | \boldsymbol{X}, \boldsymbol{\theta}^t)$$

$$\xi_n = p(z_{n-1}, z_n | \boldsymbol{X}, \boldsymbol{\theta}^t)$$

can be intuitively seen pseudo-# of occurences

initial distribution

s	
1	
2	

transition

	1	2
1		
2		

Step 2 (M-step)

Do following update

$$\pi_0(k) \propto \gamma_1(k)$$

$$A_{jk} \propto \sum_{n=2}^{N} \xi_n(j,k)$$

observation

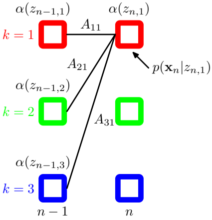
s	1	2	3
1			
2			

Step 3 t = t+1; Back to Step 1 until convergence

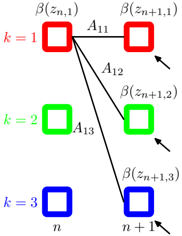
A small details

$$\xi_n = p(z_{n-1}, z_n | \boldsymbol{X}, \boldsymbol{\theta}^t)$$

$$\propto \alpha(z_{n-1}) p(z_n | z_n - 1) p(x_n | z_n) \beta(z_n)$$



why this is true?



Formally

$$\pi_0(k) = \frac{\gamma_1(k)}{\sum_{k'} \gamma_1(k')} = \frac{p(z_1 = k|\boldsymbol{X}, \boldsymbol{\theta}^t)}{\sum_{k'} p(z_1 = k'|\boldsymbol{X}, \boldsymbol{\theta}^t)}$$

$$A_{jk} = p(z_{n-1} = j, z_n = k) = \frac{\sum_{n=2}^{N} p(z_{n-1} = j, z_n = k|\boldsymbol{X}, \boldsymbol{\theta}^t)}{\sum_{k'} \sum_{n=2}^{N} p(z_{n-1} = j, z_n = k'|\boldsymbol{X}, \boldsymbol{\theta}^t)} \tag{17}$$

$$= \frac{\sum_{n=2}^{N} \xi_n(j,k)}{\sum_{k'} \sum_{n=2}^{N} \xi_n(j,k')} \tag{18}$$

How to update observation models?

$$p(x = j | z = i) = ?$$

Can you guess?

What if we have multiple observations

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M$$

How do we estimate parameters? *Can you guess?*

Outline

1 Administration

2 Review of last lecture

3 Graphical models

Graphical models

Bayes nets

Probabilistic distribution represented with directed acyclic graphs (DAGs)

Markov networks

Probabilistic distribution represented with undirected graphs

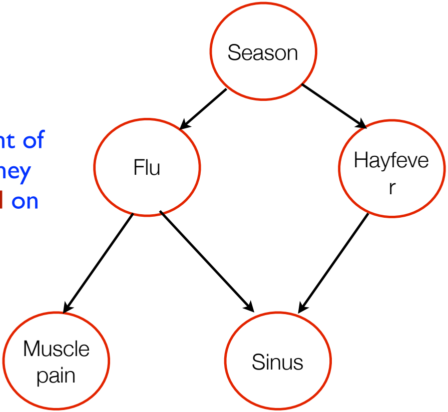
Exploring structures

Draw links between variables

- indicate dependencies, but more importantly, encode independencies
- Ex: Flu and Hayfever are independent of each other in any given season; ie, they independently occur **conditioned** on season

This is an example of Bayes networks

- Directed acyclic graphs
- Compact representation of joint distribution



The key concept

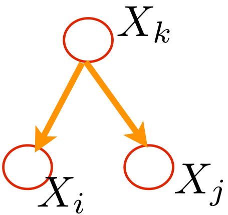
conditional independence

$$X_i \perp\!\!\!\perp X_j \mid X_k$$

allows us to write

$$\begin{aligned} p(X_i, X_j, X_k) &= p(X_i|X_j, X_k)p(X_j, X_k) \\ &= p(X_i|X_k)p(X_j|X_k)p(X_k) \end{aligned}$$

Representing it graphically



Thus, to factorize

a N-term joint distribution

$$P(X_1, X_2, \ldots, X_N) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \cdots P(X_N|X_1, X_2, \ldots, X_{N-1})$$

we need only a subset of terms

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathcal{S}_i)$$

a subset of (N-1) variables

How this is going to help us?

Factorization and conditional independence

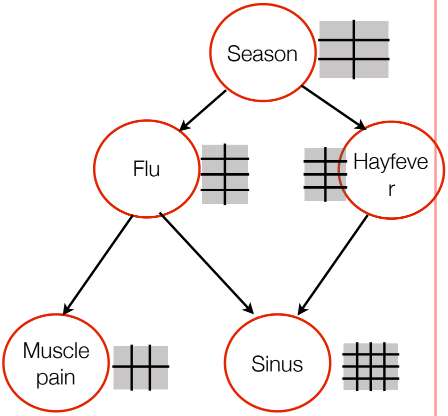
P(Season = fall, Flu = true, Muscle pain = true, Sinus = false, Hayfever = false) = P(Season = fall) *

P(Flu = true | Season = fall)
P(Hayfever= false | Season = fall) *

P(Muscle pain = true | Flu = true)

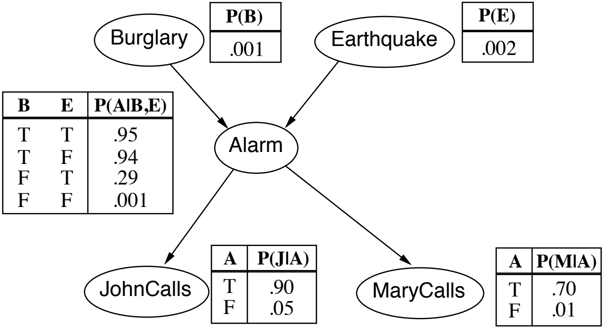
P(Sinus = true | Flu = true, Hayfever = false)

Total # of parameters for 5 random variables is?



More examples

The classical earthquake, alarm, burglary, phonecall example



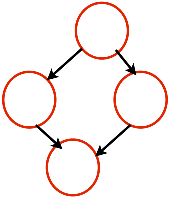
Formal definition of Bayesian networks

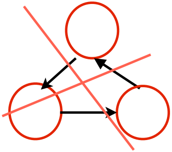
Structure (graph $\mathcal{G}$)

Vertex: random variable

Edge: directed, child vertex depends on parent

No "directed" loop: directed acyclic graph





Conditional probabilities distributions (CPD)

$$P(X_i|\mathbf{Pa}_{X_i})$$

for every vertex

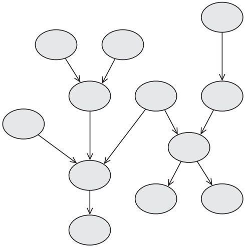
$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathbf{Pa}_{X_i})$$

also referred as CPT (cond. prob. table) with discrete variables

Semantics of Bayesian networks

The "syntax" view

Factorizing joint distribution with respect to graph structure



What are the properties can we infer from the structure?

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | \mathbf{Pa}_{X_i})$$

Semantics: local Markov property

$$X_i \perp \mathbf{NonDescendants}_{X_i} | \;\; \mathbf{Pa}_{X_i}$$

The two views are equivalent

Factorization → Local Markov Properties

If a distribution P factorizes according to the graph, then the distribution satisfies the local Markov properties (ie, local conditional independencies)

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathbf{Pa}_{X_i})$$

$$\Updownarrow$$

$$X_i \perp \mathbf{NonDescendants}_{X_i} | \ \mathbf{Pa}_{X_i}$$

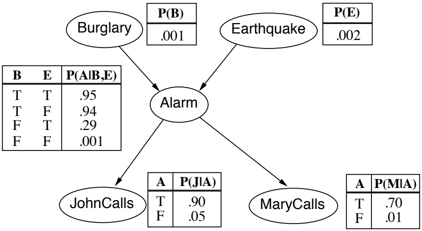
Local Markov Properties

If a distribution P satisfies local Markov properties implied in the graph, then the distribution factorizes according to the graph.

Examine the local Markov properties

$$X_i \perp \mathbf{NonDescendants}_{X_i} \mid \ \mathbf{Pa}_{X_i}$$

?



Examine the local Markov properties

$$X_i \perp \mathbf{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i}$$

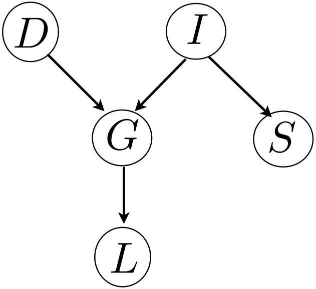
$$L \perp I, D, S \mid G$$

$$S \perp D, G, L \mid I$$

$$G \perp S \mid D, I$$

$$I \perp D$$

$$D \perp I, S$$



Note:
we constructed the graph with factorization in mind. But we are arriving at a set of independencies statements which are intuitively right. Namely, Factorization implies local Markov properties.

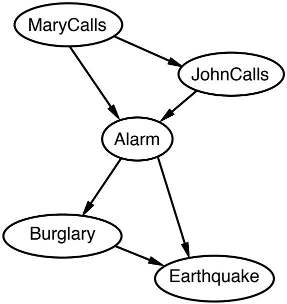
How to construct Bayesian network

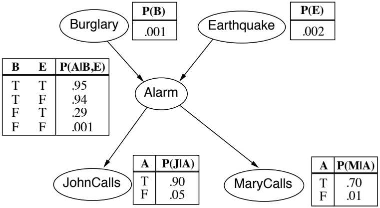
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1})$

This choice of parents guarantees the global semantics:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1}) && \text{(chain rule)}\\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i)) && \text{(by construction)}\end{aligned}$$

Different order gives different network





How to use Bayesian networks?

Once knowledge is encoded

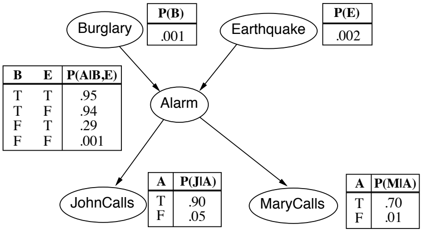
- we can query the network, ie, ask questions, ie, doing (probabilistic) inference
- Let us see a few different types of inference problem..

Causal reasoning

How likely John calls if there is a burglary?

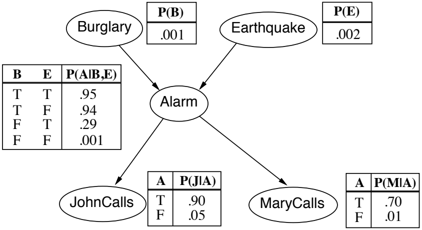
a naive approach

a better approach



Diagnostic/evidential reasoning

John calls, what is the probability of "burglary"?



P(B)
.001

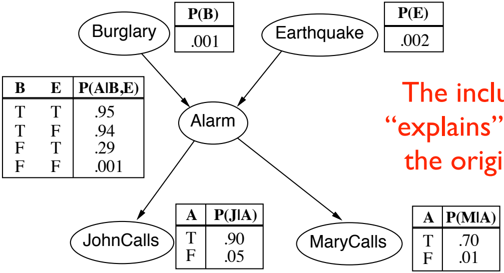
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

explaining away



The inclusion of an evidence "explains" the effect and makes the original cause less likely!

What is P('Burglary'== true | 'alarm== true')?

= 0.376

What is P('Burglary'==true | 'alarm == true' & Earthquake == 'true')?

= 0.003

Maybe the graph can tell us more?

More independence

$$X \longrightarrow Y \longrightarrow Z$$

A Bayesian network structure implies more independence than local Markov properties.

(local Markov property) $X \perp Z \mid Y$

$$X \longrightarrow Y \longrightarrow W \longrightarrow Z$$

(local Markov property) $X, Y \perp Z \mid W$

How about this guy? $\longrightarrow$ $X \perp Z \mid Y$

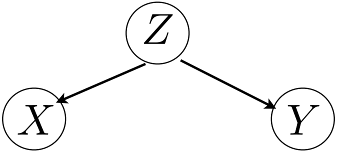
Simple cases

Indirect causal effect

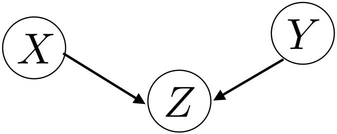
Indirect evidential effect

What are the independencies?

common cause



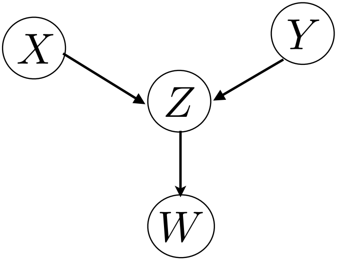
common effect
(v-structure)



More v-structure

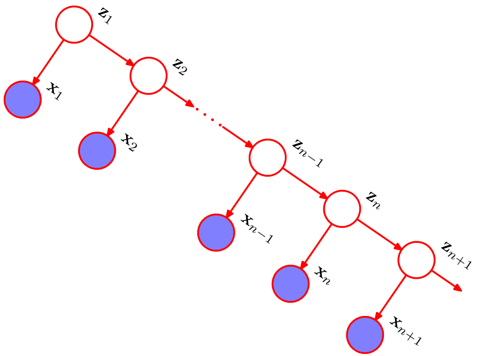
$$X \perp Y$$

How about? $X \perp Y \mid W$



Intuition: knowing W helps us to know Z, namely, as if Z is known when evaluating the independence between X and Y

But we have seen this structure before!



Application: topic model (LDA)