Brief Introduction to Graphical Models + Bayesian Networks

↳ Representation

- Probabilistic graphical models = graphs ~ compact representation of joint probability distributions!

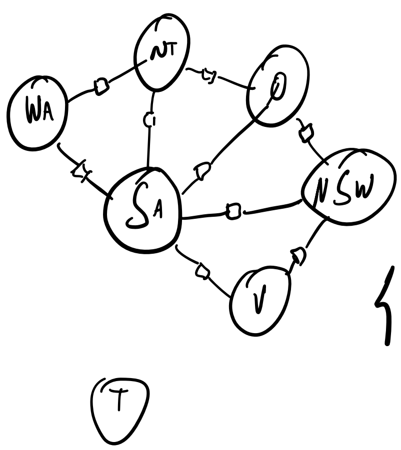
  - nodes ≡ random variables
  - (lack of) arcs ≡ conditional independence assumptions

- Undirected Graphical Models (≡ Markov Random Fields, MRF) ≡ Markov Networks

  └ independence: A and B are conditionally independent given a third set C

  IF

  all paths between A and B are separated by a node C



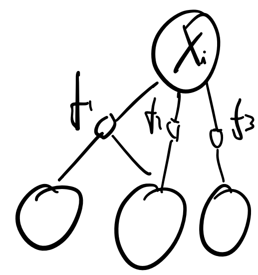
$\{WA, NT, SA, Q, NSW, V\} = A$

&

$\{T\} = B$

are independent

(there are no edges between A and B ≡ $A \perp\!\!\!\perp B$)



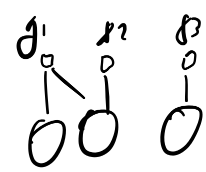
Conditioning...

└ condition $X_i = v$

└ remove X_i and all factors that depend on X_i (f_1, f_2, f_3)

└ add $g_j(x) = f_j(x \cup \{X_i = v\})$

⇓



A and B are conditionally independent given C, IF conditioning on C produces a graph in which A and B are independent

⇓

- Directed Graphical Models (≡ Bayesian Networks)

Disadvantages

- more complicated notion of conditional independence

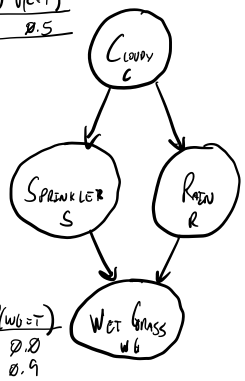
Advantages

- indication of cause-effect (helpful guide to construct the graph)

  A "causes" B

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

T - TRUE
F - FALSE



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(WG=F)	P(WG=T)
F	F	1.0	0.0
F	T	0.1	0.9
T	F	0.1	0.9
T	T	0.01	0.99

CONDITIONAL PROBABILITY DISTRIBUTION (CPD)

each row must sum to 1

$\sum_i P(WG_i \mid R=r, S=s) = 1$

$$P(C,S,R,W) = P(C) \cdot P(S|C) \cdot P(R|S,C) \cdot P(W|R,S,C)$$

Joint Probability of all nodes

Chain Rule

$(P(A,B) = P(B|A)\,P(A)$

conditional independence relationships

$$P(C,S,R,W) = P(C) \cdot P(S|C) \cdot P(R|C) \cdot P(W|R,S)$$

{ R independent of S, given C
W independent of C, given its parents R and S

- models deterministic relations ⇒ easier to learn/fit the data

"causes"

Cloudy → Rain

Cloudy → Sprinkler

Most "simplest" definition of conditional independence in Bayesian networks:

NODE INDEPENDENT OF ITS ANCESTORS GIVEN ITS PARENTS,

IF ANCESTOR – PARENT RELATION IS WITH RESPECT TO SOME FIXED TOPOLOGICAL ORDER

↳ Inference

- We observe the fact that grass is wet $(W=T)$ ⟶ 2 possible causes: it is raining OR sprinkler is on → which is more likely?

⇓

Bayes rule ⟶ transpose the patterns

$P(R=T \mid W=T) = ?$

1. Compute the joint probability $P(C,S,R,W)$

$$P(C,S,R,W) = P(C) \cdot P(S|C) \cdot P(R|C) \cdot P(W|R,S)$$

2. Marginalize all variables we do not need

C	S
F	F
F	T
T	F
T	T

⇓

$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T \mid W=T) = \frac{\sum_C \sum_S P(C,S,R=T,W=T)}{\sum_C \sum_S \sum_R P(C,S,R,W=T)} =$$

$$= \frac{\overset{(F,F,T,T)}{0.5 \times 0.5 \times 0.2 \times 0.9} + \overset{(F,T,T,T)}{0.5 \times 0.5 \times 0.2 \times 0.99} + \overset{(T,F,T,T)}{0.5 \times 0.9 \times 0.8 \times 0.9} + \overset{(T,T,T,T)}{0.5 \times 0.1 \times 0.8 \times 0.99}}{\dots} = \frac{0.4581}{0.6471} = 0.708$$

vs

C	R
F	F
F	T
T	F
T	T

$$P(S=T \mid W=T) = \frac{P(S=T, W=T)}{P(W=T)} = \frac{\sum_C \sum_R P(C,S=T,R,W=T)}{\sum_C \sum_S \sum_R P(C,S,R,W=T)} =$$

$$= \frac{\overset{(F,T,F,T)}{0.5 \times 0.5 \times 0.8 \times 0.9} + \overset{(F,T,T,T)}{0.5 \times 0.5 \times 0.2 \times 0.99} + \overset{(T,T,F,T)}{0.5 \times 0.1 \times 0.2 \times 0.9} + \overset{(T,T,T,T)}{0.5 \times 0.1 \times 0.8 \times 0.99}}{0.6471} = \frac{0.2781}{0.6471} = 0.430$$

normalizing constant (= likelihood of the data)!

- Top-down vs bottom-up
  - └ diagnostic → evidence on an effect ⤻ infer most likely cause
  - how causes generate effects
- Conditional independence on Bayes nets ⟶ BAYES BALL

└ hidden nodes → values not known!

└ observed nodes

A and B conditionally independent given C

IF ONLY IF there is no way for a ball to get from A to B in the graph

↳ Temporal Models = DYNAMIC BAYESIAN NETWORKS (DBN)

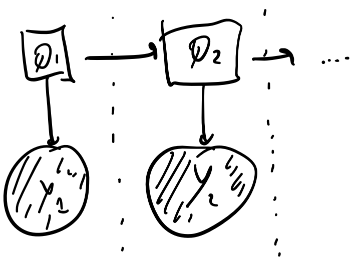
generalize HIDDEN MARKOV MODELS (HMM) and LINEAR DYNAMICAL SYSTEMS (LDS)

↳ by representing hidden (and observed) states in terms of state variables

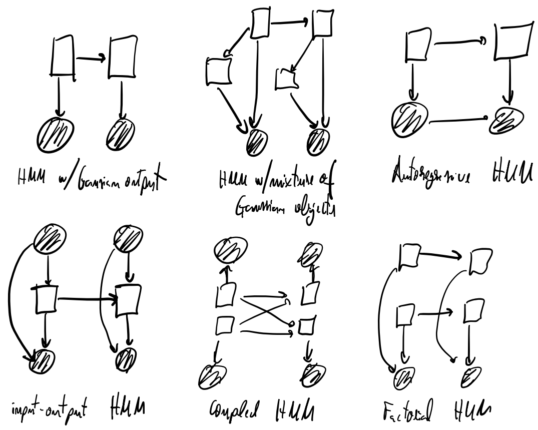
- HIDDEN MARKOV MODELS (HMM)
  └ 1 discrete hidden node + 1 discrete / continuous observed node per slice

□ → discrete

○ → continuous



└→ time slices

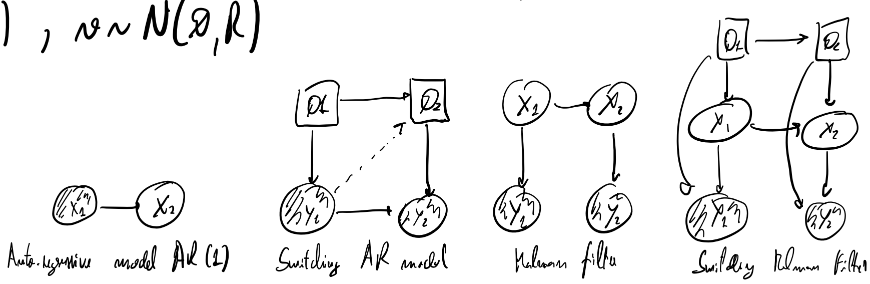


- LINEAR DYNAMICAL SYSTEMS (LDSs) and KALMAN FILTERS → all nodes assumed to have linear Gaussian distributions

$$X_{t+1} = A x_t + w(t), \quad w \sim N(0, Q)$$

$$Y_t = C x_t + v(t), \quad v \sim N(0, R)$$

→ Kalman filter = way of doing online filtering of this model



INFERENCE

→ graphical model specifies the complete joint probability distribution (JPD) over all the variables

→ with JPD, can answer all possible inference queries by MARGINALISATION (summing up all irrelevant variables)

→ however, JPD $\sim O(2^n)\Big|_{n:\#\text{nodes}}$ ⊕ only assuming 2 states

↳ Variable elimination

- use factored representation of JPD to do marginalisation efficiently!

⇓

"push sums" in as far as possible ≡ VARIABLE ELIMINATION ALGORITHM

$$P(W=w) = \sum_C \sum_S \sum_R P(C=c_i, S=s_j, R=r_k, W=w)$$

$$= \sum_C \sum_S \sum_R P(C=c_i) \cdot P(S=s_j \mid C=c_i) \cdot P(R=r_k \mid C=c_i) \cdot P(W=w \mid S=s_j, R=r_k)$$

$$= \sum_C P(C=c_i) \cdot \underbrace{\sum_S P(S=s_j \mid C=c_i) \underbrace{\sum_R P(R=r_k \mid C=c_i) \cdot P(W=w \mid S=s_j, R=r_k)}_{T_1(c_i, s_j, w)}}_{T_2(c_i, w)}$$