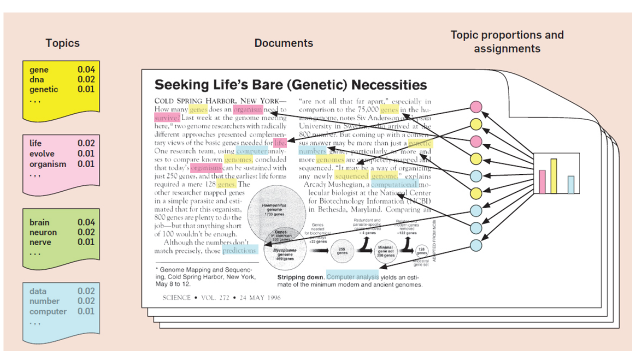
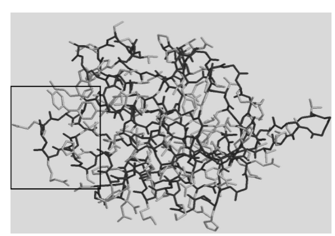
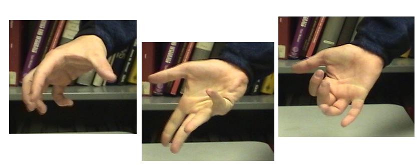
Bayesian Networks – Representation

CSE 446: Machine Learning
Emily Fox
University of Washington
March 6, 2017

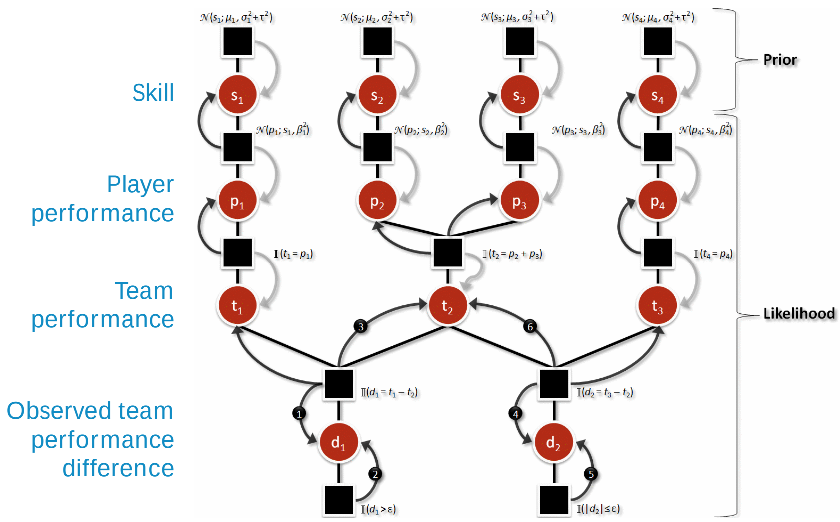
Learning from structured data





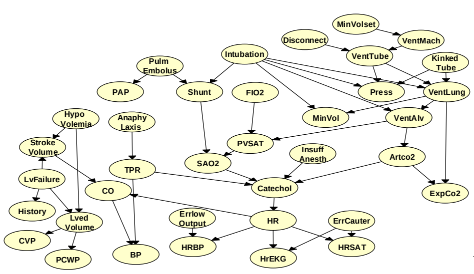


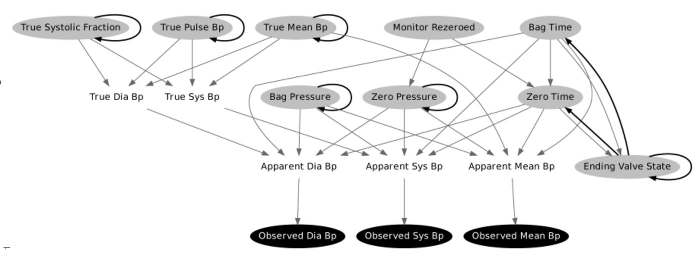
TrueSkill: A Bayesian Skill Rating System



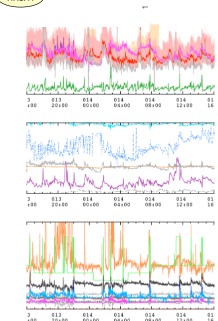
Herbrich et al., 2007

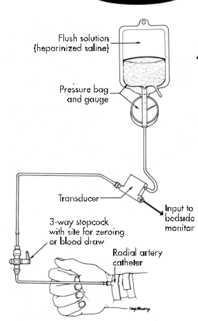
ICU Monitoring





Beinlich et al., 1989

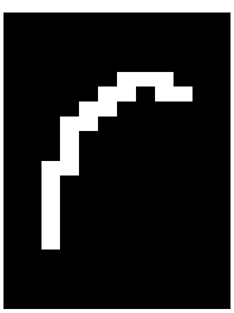


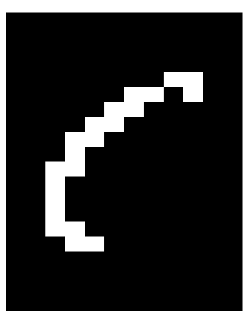


Aleks, Russell, et al., 2008

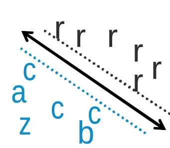
Digging in:
Learning with and without context/structure

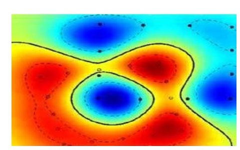
Without context: Handwriting recognition





Character recognition,
e.g., kernel SVMs





Without context: Webpage classification

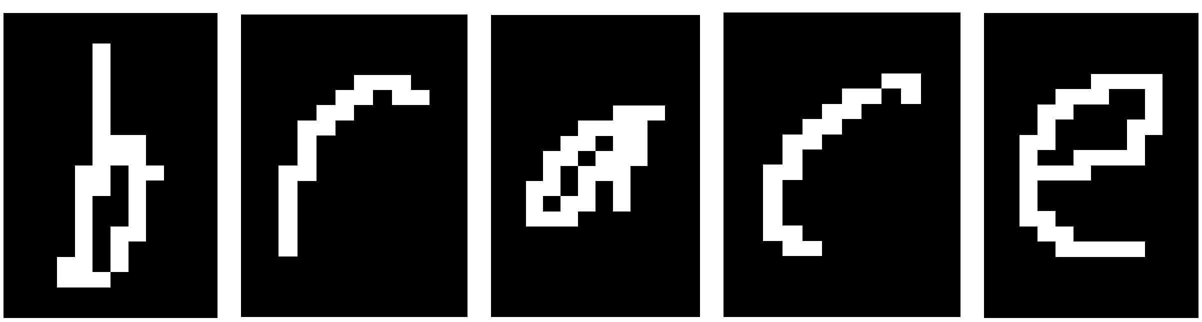
Company website

University website

Personal website

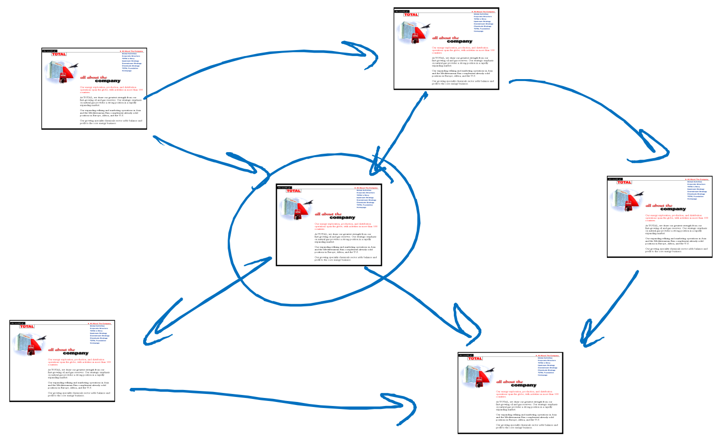
…

With context: Handwriting recognition



"c" more likely to come after "a" than another "a"

With context: Webpage classification



company pages tend to point to each other

Modeling structured relationships via Bayesian networks

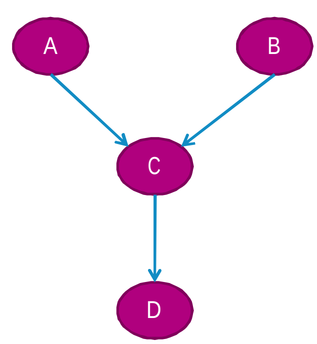
Today – Bayesian networks

- Provided a huge advancement in AI/ML
- Generalizes naïve Bayes and logistic regression
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

Bayesian network representation

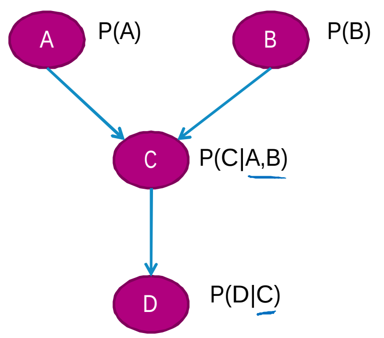
Compact representation of a probability distribution.

Directed Acyclic Graph



Vertices: Random Variables
Edges: Conditional dependencies
"probabilistic relationships"

Bayesian network probability factorization



One **CPT** (conditional probability table) for each variable

P(variable | parents of variable)

implies the factorization:

$$P(\mathbf{X}) = \prod_{j=1}^{d} P(\mathbf{X}[j] \mid \text{parents}(\mathbf{X}[j]))$$

node

parents

joint:

P(A,B,C,D) = P(A) P(B) P(C|A,B) P(D|C)

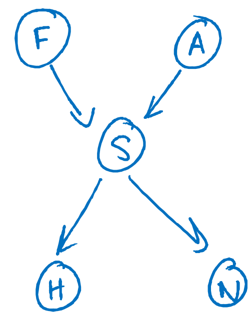
13

CSE 446: Machine Learning

What a Bayesian network represents (in detail) and what does it buy you?

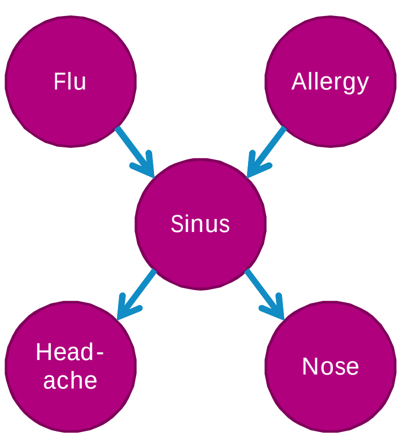
Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?



(Not a true "causal model")

Possible queries



- Inference ← given some evidence / observable

  $P(F=t \mid N=t)$

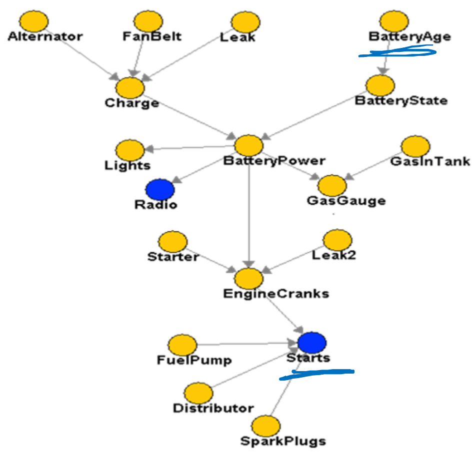
  ↖ infer state of some variable
- Most probable explanation

  $\max_{f,a,s,h} P(f,s,a,h \mid N=t)$
- Active data collection

  What variable should I observe next?

  H=? S=?

CarStarts? Bayesian network



- 18 binary attributes

  2^{18} probabilities

- Inference
  - P(BatteryAge|Starts=f) $= \dfrac{P(BA, S=f)}{P(S=f)} \propto P(BA, S=f)$

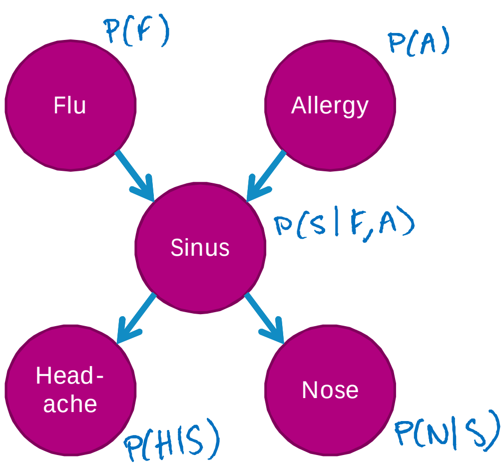
$$= \sum_{\underbrace{a,f,l,c,\ldots}_{16}} P(a,f,l,c,\ldots,S=f)$$

2^{16}

(everything other than BatteryAge and Starts)

- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution – A preview

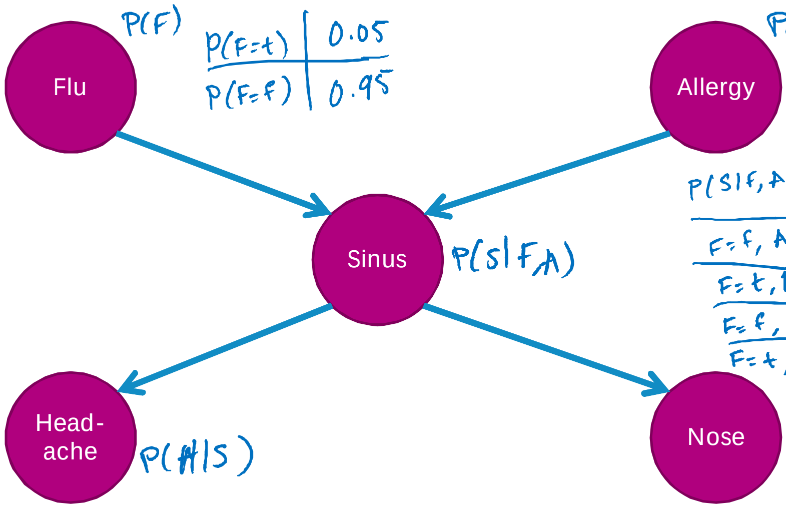


$\overbrace{P(F,A,S,H,N)}^{2^5 = 32 \text{ terms} \ldots}$ 31 params

$= P(F)P(A)P(S|F,A)P(H|S)P(N|S)$

(will see later why this fact holds)

What are these probabilities? Conditional probability tables (CPTs)

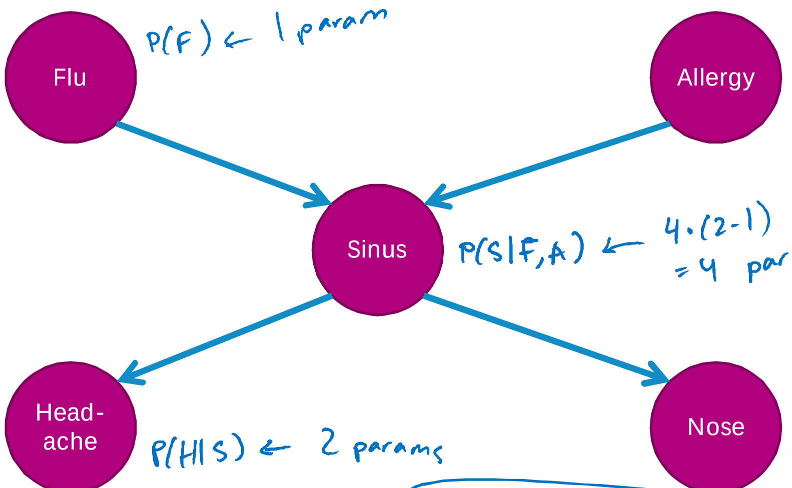


P(F)	
P(F=t)	0.05
P(F=f)	0.95

P(A)	
t	0.2
f	0.8

P(S|F,A)	S=t	S=f
F=f, A=f	0.1	0.9
F=t, A=f	0.7	0.3
F=f, A=t	0.4	0.6
F=t, A=t	0.8	0.2

Number of parameters



P(F) ← 1 param

P(A) ← 1 param

P(S|F,A) ← 4·(2-1) = 4 params

P(H|S) ← 2 params

P(N|S) ← 2 params

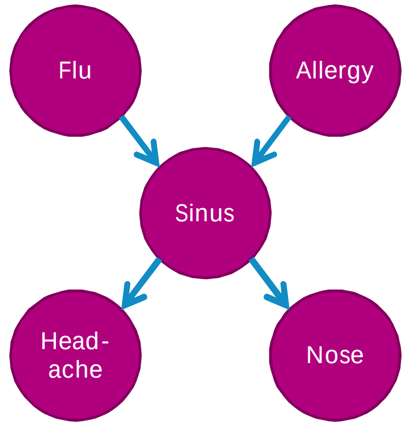
10 total params

P(F, A, S, H, N) → 31 params

10 < 31

⇒ - more bias
- less flexible
+ need less data to learn
+ more accurate on smaller datasets

Factorization speeds up inference



Exploit distributivity:

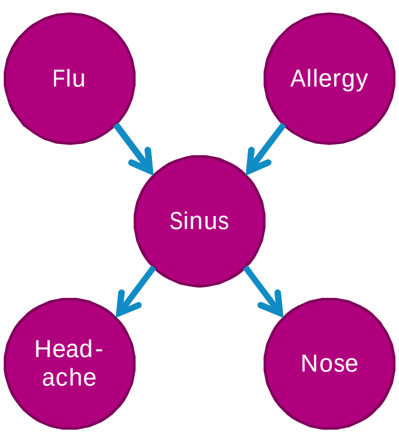
$$P(F=x_F|N=t) \propto \sum_{x_A,x_S,x_H} P(F=x_F, A=x_A, S=x_S, H=x_H, N=t)$$

$$= \sum_{x_A,x_S,x_H} P(F=x_F)P(A=x_A)P(S=x_S \mid F=x_F, A=x_A)P(H=x_H \mid S=x_S)P(N=t \mid S=x_S)$$

$$= P(F=x_F)\sum_{x_A} P(A=x_A)\sum_{x_S} P(S=x_S \mid F=x_F, A=x_A)P(N=t \mid S=x_S)\sum_{x_H} P(H=x_H \mid S=x_S)$$

variable elimination algorithm

Key: Independence assumptions



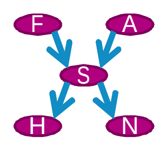
$F \perp N \mid S$

- Flu only "causes" Nose through Sinus
- If you tell N=t, this changes prob. of F, but if I first tell you S=t, N doesn't affect prob. of F

Knowing sinus separates variables from each other

Marginal and conditional independence

(Marginal) Independence



- Flu and Allergy are (marginally) independent

$F \perp A$

$P(A,F) = P(A)P(F)$

$\Updownarrow$

$P(A|F) = P(A)$

$P(F)$

Flu = t	0.2
Flu = f	0.8

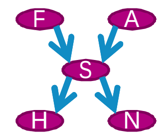
$P(A)$

Allergy = t	0.4
Allergy = f	0.6

$P(A,F)$

	Flu = t	Flu = f
Allergy = t	$0.4 \times 0.2 = 0.08$	0.4×0.8
Allergy = f	0.8×0.2	0.8×0.6

Conditional independence



- Flu and Headache are not (marginally) ind.

  $P(H=t \mid F=t) \neq P(H=t)$

- Flu and Headache are independent given Sinus infection

  $P(H=t \mid F=t, S=t) = P(H=t \mid S=t)$

- More generally:

  $X \perp Y \mid Z$

  $P(X,Y \mid Z) = P(X \mid Z)\,P(Y \mid Z)$

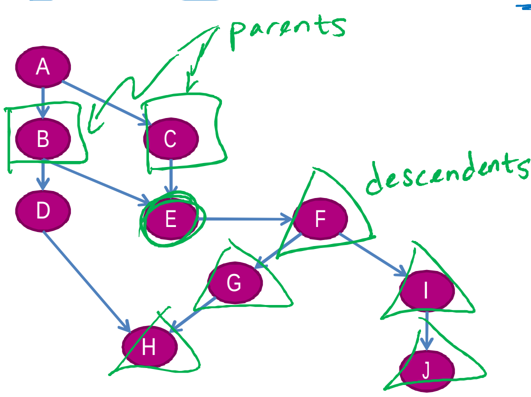
  $\Updownarrow$

  $P(X \mid Y,Z) = P(X \mid Z)$

Conditional independence statements encoded by Bayesian networks

What is a Bayes net assuming?

Local Markov Assumption: A variable X is independent of its non-descendents given its parents



$E \perp A \mid B,C$

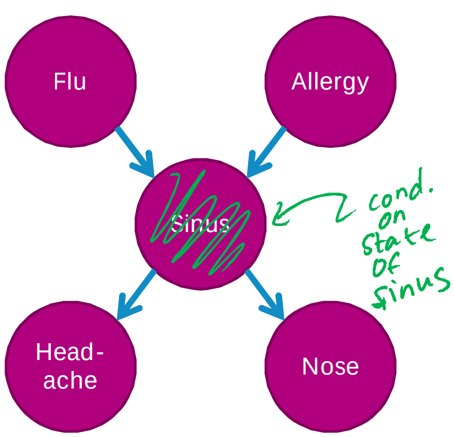
$E \perp D \mid B,C$

$F \perp B \mid E$

Allows you to read off some simple conditional independence relationships

Explaining away example

Local Markov Assumption:
A variable X is independent of its non-descendents given its parents



$F \perp A$ (cond. ind. given ϕ = marg. ind.)

$F \perp A \mid S$?? don't know

$P(F=t \mid A=t, S=t) \stackrel{?}{=} P(F=t \mid S=t)$

Suppose $P(F=t \mid S=t)$ is high but $P(F=t \mid S=t, A=t)$ is lower because $A=t$ explains away sinus inflamation.

Naïve Bayes revisited

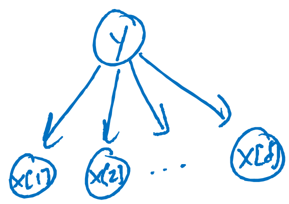
Local Markov Assumption:
A variable X is independent of its non-descendents given its parents

$X[1] \perp X[2],\dots,X[d] \mid y$

↑ true for any feature and set of remaining features

$$P(y, X[1],\dots,X[d]) = P(y)\prod_{j=1}^{d} P(X[j] \mid y)$$

Naive Bayes



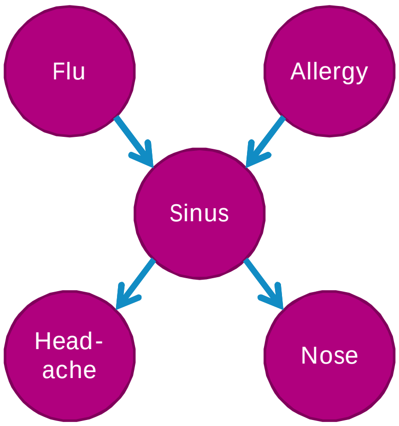
Local Markov assumption of this graph:

$X[j] \perp X[1],\dots,X[j-1], X[j+1],\dots,X[d] \mid y$

← Naive Bayes!

Factorization of the joint distribution

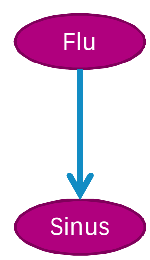
Joint distribution



Why can we decompose?
Markov Assumption!

The chain rule of probabilities

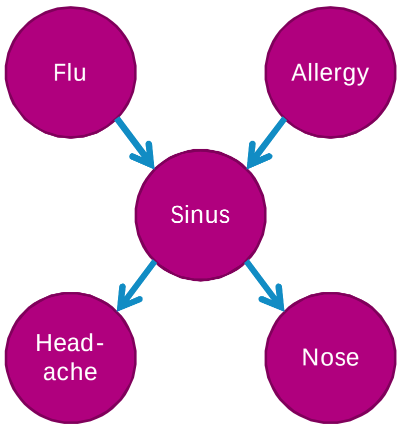
- P(A,B) = P(A)P(B|A)



- More generally:
  - $P(\mathbf{X}[1],...,\mathbf{X}[d]) = P(\mathbf{X}[1])\ P(\mathbf{X}[2]|\mathbf{X}[1])\ \ldots\ P(\mathbf{X}[d]|\mathbf{X}[1],..., \mathbf{X}[d-1])$

Chain rule & joint distribution

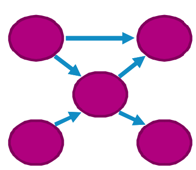
Local Markov Assumption:
A variable X is independent of its non-descendents given its parents



Order of expansion matters! Use topological order

The Representation Theorem – Joint distribution to BN

Bayes net



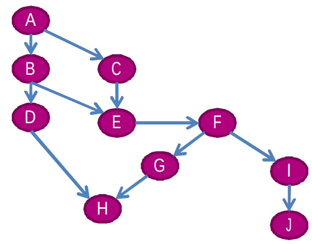
Encodes independence assumptions

If cond. ind. in Bayes net are subset of cond. ind. in P

Joint distribution factorizes:

$$P(\mathbf{X}) = \prod_{j=1}^{d} P(\mathbf{X}[j] \mid \text{parents}(\mathbf{X}[j]))$$

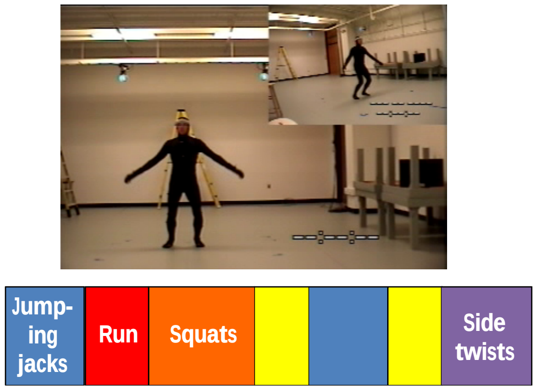
Bayesian networks recap



- Representation benefits
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
  - Lower variance parameter estimates from limited data
- Inference benefits
  - Efficient computation of P(X|e) (i.e., fast probabilistic inference)
  - Involves variable elimination algorithms
- Other important topics
  - Structure learning: What graph structure to use?
  - Understanding how evidence can be incorporated and how this changes conditional independence statements (d separation)

Hidden Markov models:
A Bayesian network for time series

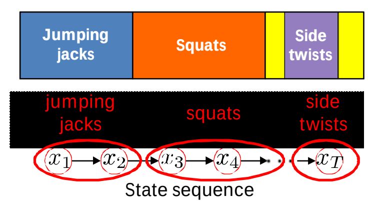
Example: Motion Capture Segmentation



37 ©2017 Emily Fox CSE 446: Machine Learning

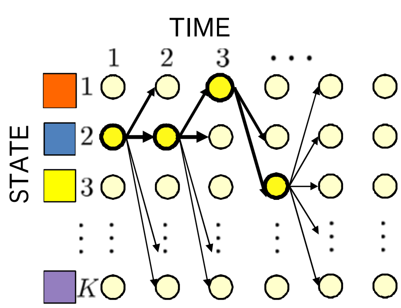
Hidden Markov model

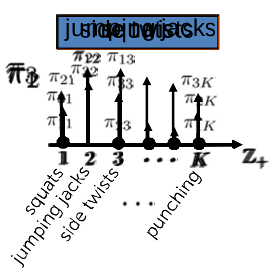
Tutorial: Rabiner, Proc. IEEE 1989

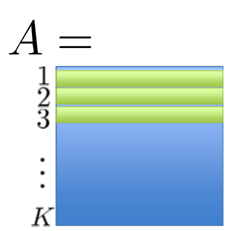


Markov transition dynamics:

$$\Pr(x_t = \square \,|\, x_{t-1} = \square) = A_{\square\square}$$



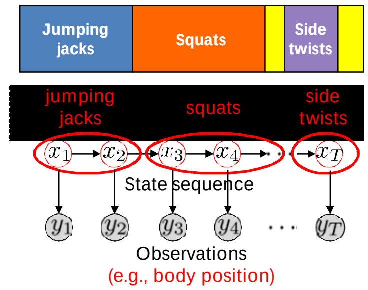




38 ©2017 Emily Fox CSE 446: Machine Learning

Hidden Markov model

Tutorial: Rabiner, Proc. IEEE 1989



Markov transition dynamics:

$$\Pr(x_t = \blacksquare \mid x_{t-1} = \blacksquare) = A_{\blacksquare\blacksquare}$$

Conditionally independent emissions:

$$\Pr(y_t \mid x_t = \blacksquare) = N(\mu_{\blacksquare}, \Sigma_{\blacksquare})$$

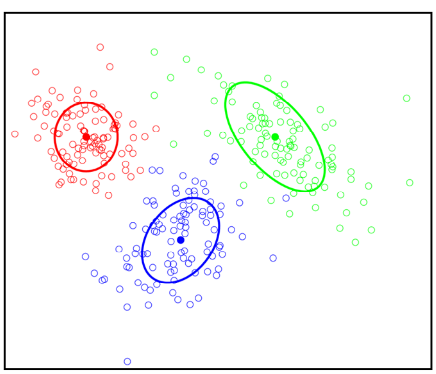
Joint distribution factorization:

Latent Markov chain structure enables

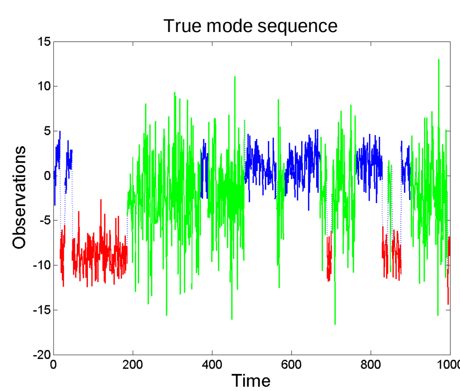
- Efficient computation of marginals $p(x_t \mid y_1, \ldots, y_T)$ via forward-backward alg.
- Most-probable sequence via Viterbi
- Parameter learning via Baum-Welch (EM for HMMs)

GMMs vs. HMMs

Gaussian mixture model

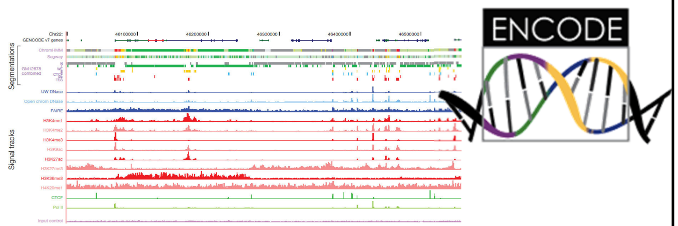


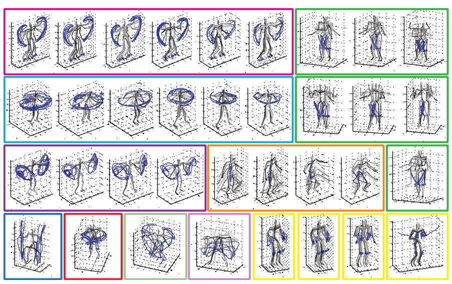
Hidden Markov model

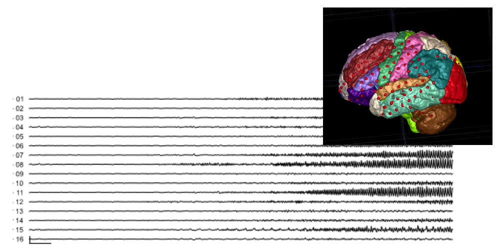


HMM applications

Example applications:
- Parsing EEG recordings
- Discovering behaviors in videos
- Speech segmentation
- Volatility regimes in financial time series
- Genomics
- ...





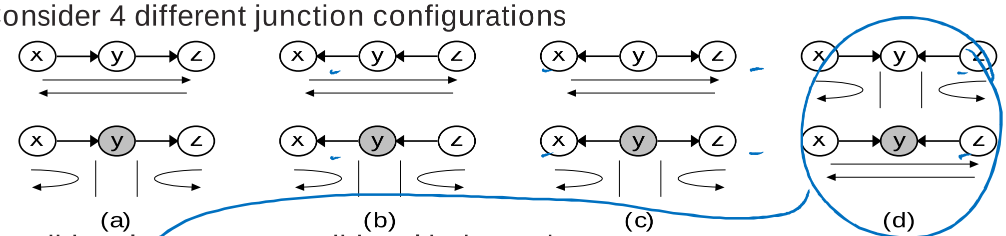


Incorporating evidence: Bayes ball algorithm for analyzing conditional independencies

OPTIONAL

Conditional independence in Bayes nets

- Consider 4 different junction configurations



- Conditional versus unconditional independence:

$p(x,y,z) = p(x)p(z)p(y|x,z) \overset{\text{int. over } y}{\Rightarrow} p(x,z) = p(x)p(z) \Rightarrow x \perp\!\!\!\perp z$

$p(x,z|y) \propto p(x,y,z) = p(x)p(z)p(y|x,z)$

$\neq p(x|y)p(z|y) \leftarrow x \not\perp\!\!\!\perp z \mid y$

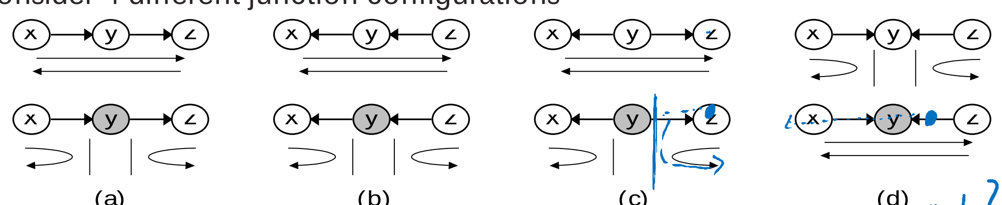
"explaining away": x= earthquake z = burglar , y= car alarm

ind. apriori

If alarm (y=1) , an increase in earthquake p(x|y) , means p(z|y) lower

Bayes ball algorithm

- Consider 4 different junction configurations



yes!

$z \perp\!\!\!\perp x | y$?

→ No!

- Bayes ball algorithm:

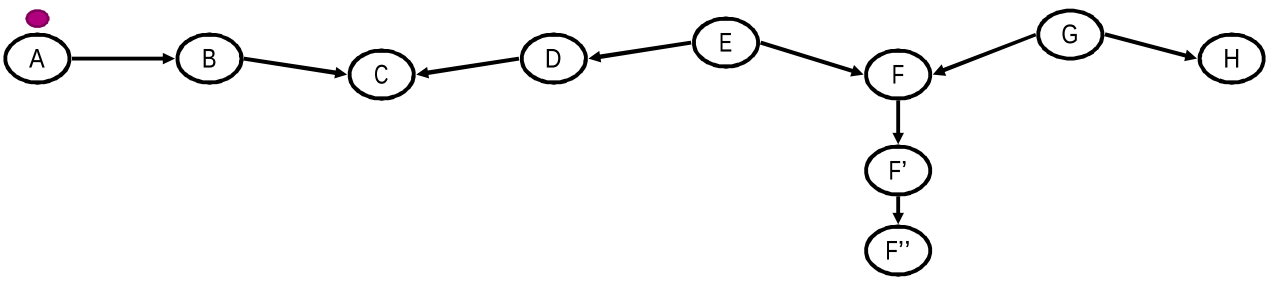
start ball at one end or other.

If ball passes to a node (straight arrows) then,
not cond./marg. ind.

If ball bounces back (walls + curved arrows)
the nodes are cond./marg. ind.

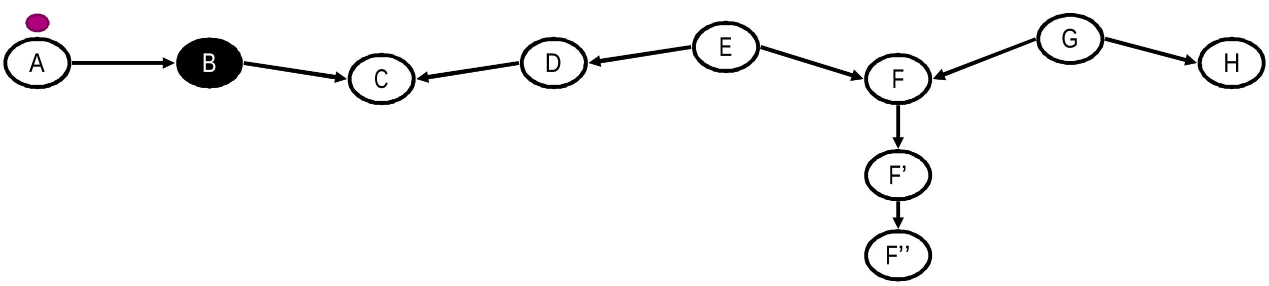
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



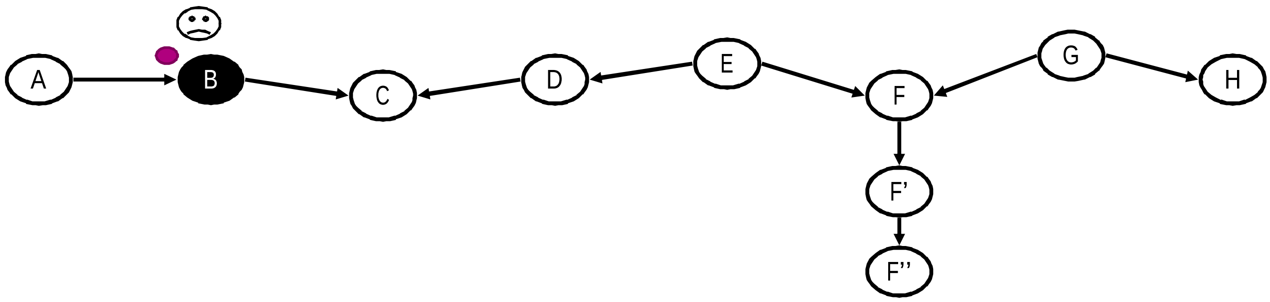
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



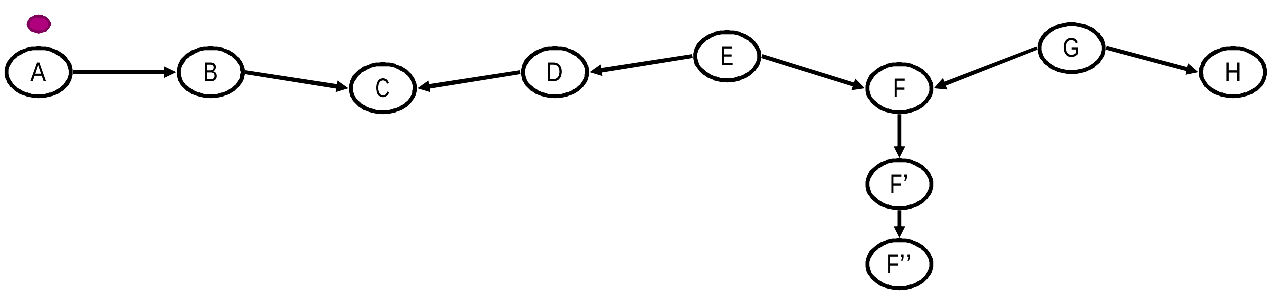
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



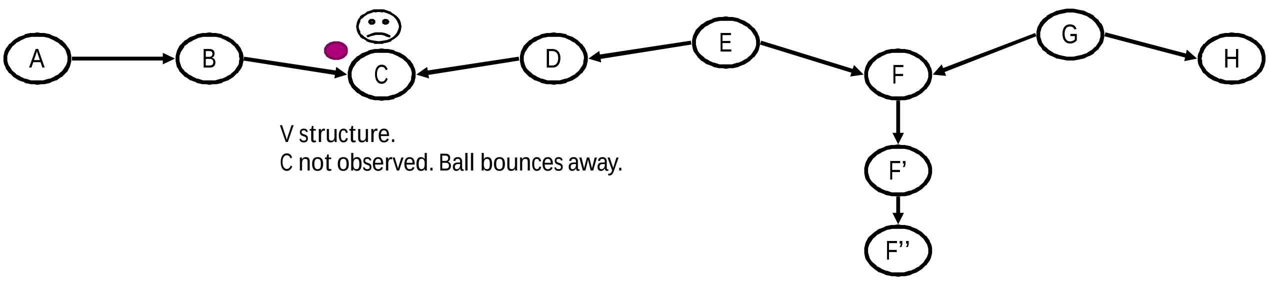
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



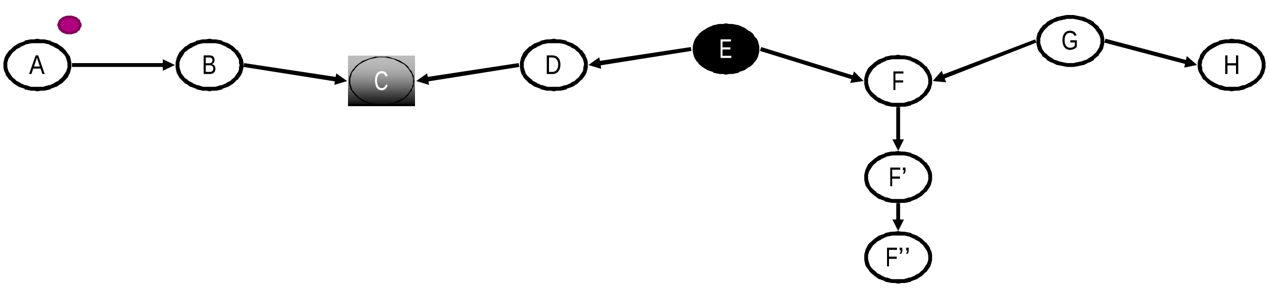
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



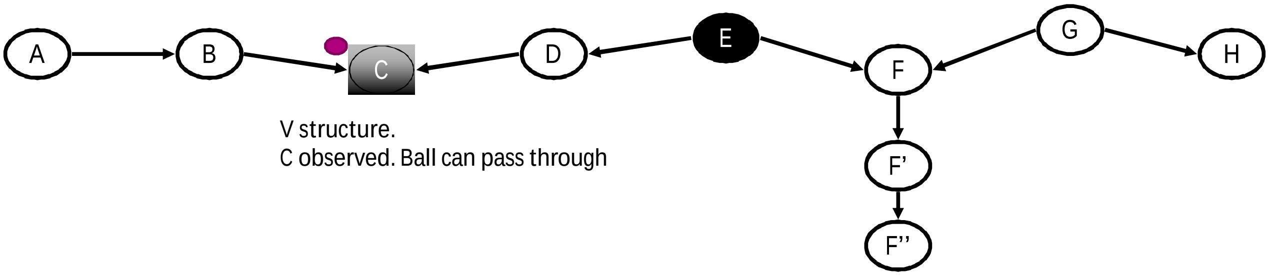
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



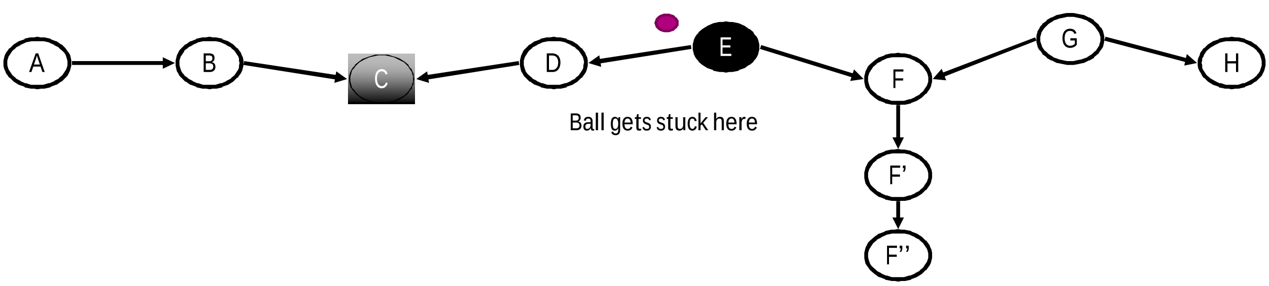
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



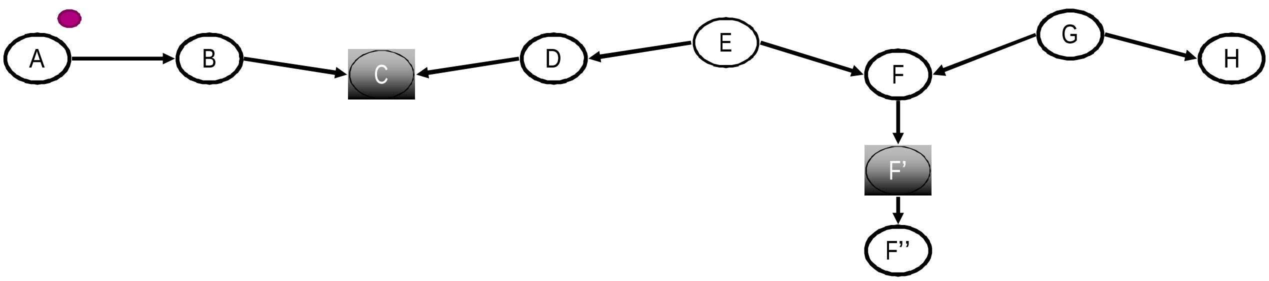
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



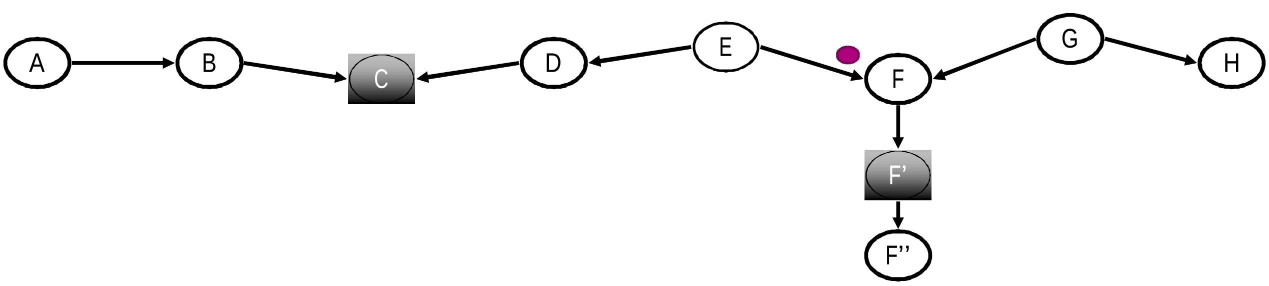
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



V structure.
Descendent of F observed.
Ball can pass through

Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H