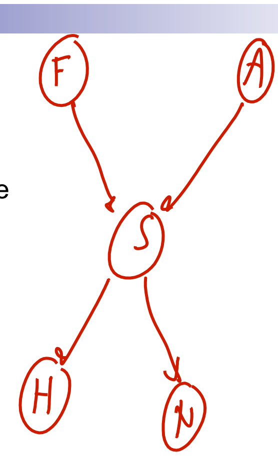
Bayesian Networks – Representation

Machine Learning – CSE446
Carlos Guestrin
University of Washington

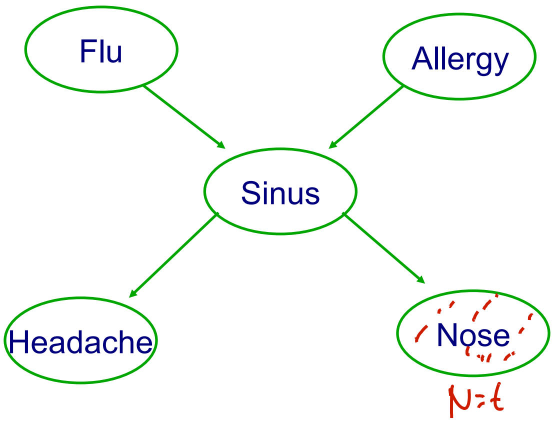
May 29, 2013

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?

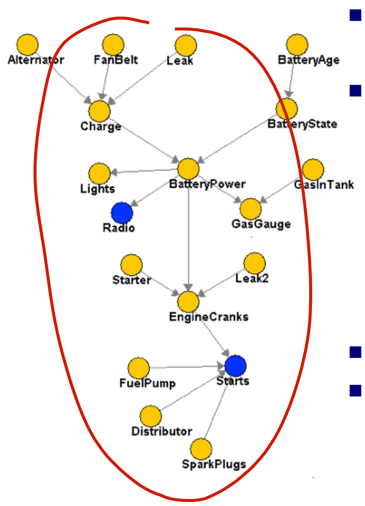


Possible queries



- Inference
  $P(F=t \mid N=t)$
- 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=?

Car starts BN

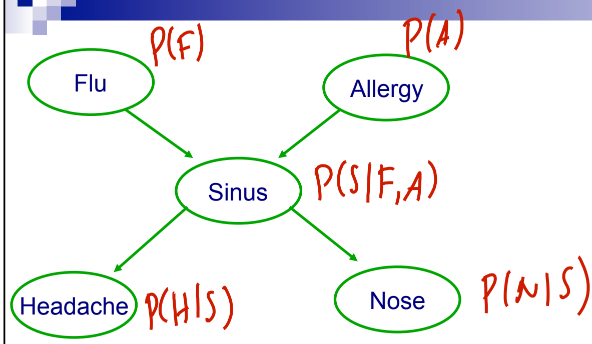


- 18 binary attributes
  2^{18} probabilities
- Inference
  - P(BatteryAge|Starts=f)

$$P(BA \mid S=f) = \sum_{\underbrace{a,f,l,c,\ldots}_{16}} P(a,fb,l,\ldots,R=f,\ldots,S=f) \leftarrow 2^{16}$$

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

Factored joint distribution - Preview



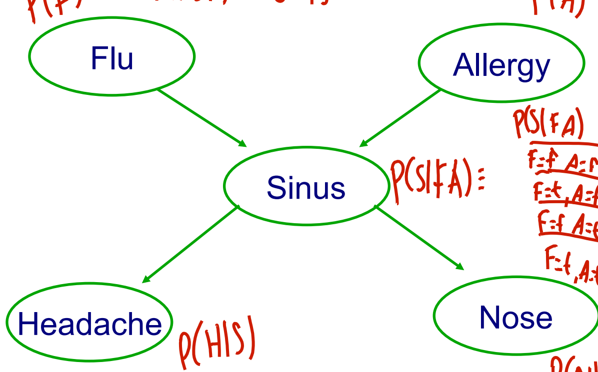
$$\underbrace{P(F,A,S,H,N)}_{\substack{2^5 = 32 \text{ terms} \\ 31 \text{ params}}} = P(F)\,P(A)\,P(S|F,A)\,P(H|S)\,P(N|S)$$

What about probabilities? Conditional probability tables (CPTs)

$P(F)$: $P(F=t) = 0.05$, $P(F=f) = 0.95$

$P(A)$:

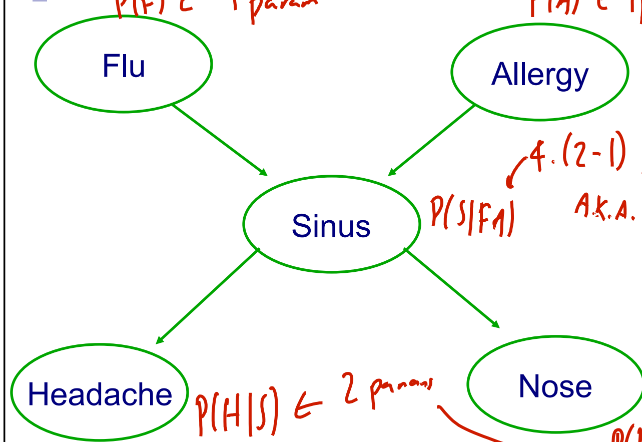
t	0.2
f	0.8



$P(S|F,A)$:

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



$P(F) \leftarrow 1$ param

$P(A) \leftarrow 1$ params

$P(S|FA) \leftarrow 4\cdot(2-1)$ params A.K.A. 4

$P(H|S) \leftarrow 2$ params

$P(N|S)$

Total: 10 params

$P(FASHN)$

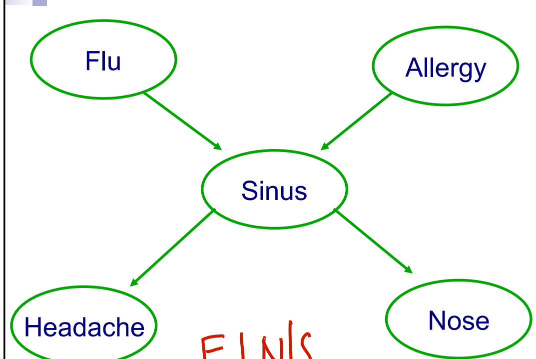
$\rightarrow 31$ params

$10 < 31$

$\Rightarrow$

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

Key: Independence assumptions



$F \perp N | S$

Flu only "causes" Nose through Sinus

if you tell $N=t$ changes prob of F, but if I first tell you $S=t$, N doesn't affect prob. of Flu

Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A$

$\underbrace{P(A,F) = P(A)P(F)}$

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

$P(F)$

Flu = t	.2
Flu = f	.8

$P(A)$

Allergy = t	.4
Allergy = f	.6

$P(F,A)$	Flu = t	Flu = f
Allergy = t	.4×.2 = 0.08	.4×.8
Allergy = f	.6×.2	.8×.6

Marginally independent random variables

- **Sets** of variables **X**, **Y**
- X is independent of Y if
  - *P* ⊢ (**X**=**x**⊥**Y**=**y**), ∀**x**∈Val(**X**), **y**∈Val(**Y**)

entails

distribution $P(X=x, Y=y) = P(X=x)\,P(Y=y) \quad \forall x, \forall y$

- Shorthand:
  - **Marginal independence:** *P* ⊢ $\underline{(\mathbf{X} \perp \mathbf{Y})}$

- **Proposition:** *P* statisfies (**X** ⊥ **Y**) if and only if
  - P(**X**,**Y**) = P(**X**) P(**Y**)

$P(X|Y) = P(X)$ equivalent

Conditional independence

- Flu and Headache are not (marginally) independent

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

- Flu and Headache are independent given Sinus infection $P(H=t \mid S=t) = P(H=t \mid S=t, F=t)$

- More Generally: $X \perp Y \mid Z$

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

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

Conditionally independent random variables

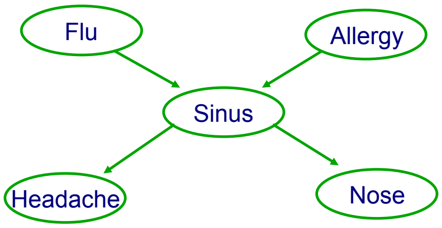
- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if
  - *P* ⊢ (**X**=**x**⊥**Y**=**y**|**Z**=**z**), ∀**x**∈Val(**X**), **y**∈Val(**Y**), **z**∈Val(**Z**)

  $P(X=x \mid Y=y, Z=z) = P(X=x \mid Z=z)$
- Shorthand:
  - **Conditional independence:** *P* ⊢ (**X** ⊥ **Y** | **Z**)
  - For *P* ⊢ (**X** ⊥ **Y** |∅), write P ⊢ (**X** ⊥ **Y**)

- **Proposition:** *P* statisfies (**X** ⊥ **Y** | **Z**) if and only if
  - P(**X**,**Y**|**Z**) = P(**X**|**Z**) P(**Y**|**Z**)

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

The independence assumption

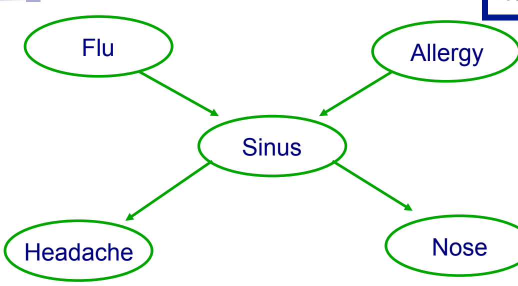


Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parent

	F	A	S	H	N
non. descendants	A	F	FA	FAN	FAH
implies	$F \perp A$	$A \perp F$	$S \perp FA \mid FA$ ⇒ nothing	$H \perp \{F, A, N\} \mid S$	$N \perp \{F, A, H\} \mid S$

Explaining away

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parents



$F \perp A$

$F \perp A \mid S$??

don't know

$P(F=t \mid A=t, S=t) \neq P(F=t \mid S=t)$

No!!

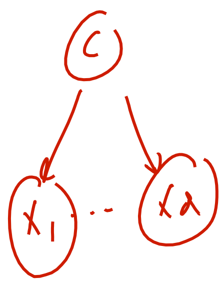
Suppose $P(F=t \mid S=t)$ is high
but $P(F=t \mid S=t, A=t)$ is lower
because A=t explain away sinus infection

Naïve Bayes revisited

$X_1 \perp \{X_2 \ldots X_d\} | C$

$P(C, X_1, \ldots X_d) = P(C) \prod P(X_i | C)$

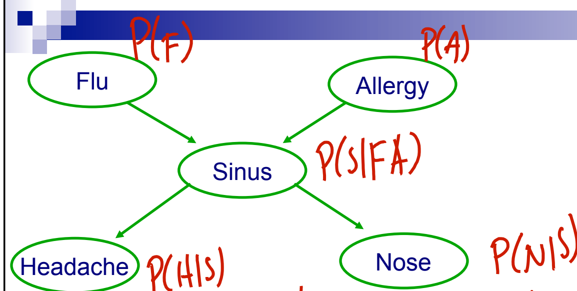
Local Markov Assumption:
A variable X is independent of its non-descendants given its parents



$X_1 \perp \{X_2 .. X_d\} | C$

← naive Bayes!!

Joint distribution

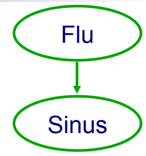


why can you write this way?

$\underbrace{P(FASHN)}_{\text{some general distribution}} = \underbrace{P(A)\,P(F)\,P(S|FA)\,P(H|S)\,P(N|S)}_{\text{few parameters}}$

Why can we decompose? Markov Assumption!

The chain rule of probabilities



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

For any dist:

$P(F,S) = P(F)\ P(S|F)$ — for any ordering

$P(F,A,S) = P(F)\ P(A|F)\ P(S|F,A)$

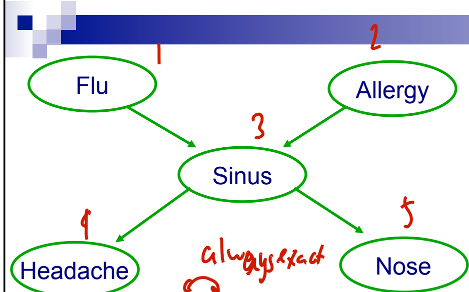
- More generally: any ordering over variables
  - $P(X_1,\ldots,X_n) = P(X_1)\ P(X_2|X_1)\ \ldots\ P(X_n|X_1,\ldots,X_{n-1})$

$P(X_3|X_2,X_1)$

©Carlos Guestrin 2005-2013 **17**

Chain rule & Joint distribution

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



proof by example: works for all BNs !! :)

always exact

$P(FASHN) = P(F)\ \underbrace{P(A|F)}_{P(A)}\ P(S|FA)\ \underbrace{P(H|SFA)}_{P(H|S)}\ \underbrace{P(N|SFAH)}_{P(N|S)}$

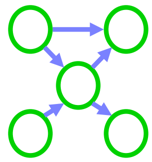
order of chain rule expansion matters a lot !! : use topological order

$A \perp F \Rightarrow P(A|F) = P(A)$	$H \perp \{F A\} | S$ $P(H|SFA) = P(H|S)$	$P(N|SFAH) = P(N|S)$ $N \perp \{FAH\} | S$

©Carlos Guestrin 2005-2013 **18**

The Representation Theorem – Joint Distribution to BN

BN:



Encodes independence assumptions var indep of non-descendants given its parents

there is also a converse

If conditional independencies in BN are subset of conditional independencies in *P*

Obtain

Joint probability distribution:

$$P(X_1, \ldots, X_n) = \underbrace{\prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)}_{\text{product over CPTs}}$$

Two (trivial) special cases

Edgeless graph

X_1 X_2 X_3

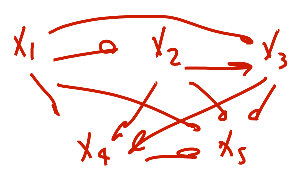
X_4 X_5

$X_i \perp \{\text{everybody}\} \mid \emptyset$

all vars independent

structure learning

Fully-connected graph



$X_i \perp \underbrace{\{\text{non-descendants}\}}_{\text{and not parents}} \mid$ parents $X_1 \ldots X_{i-1}$

$\emptyset$

no vars independents

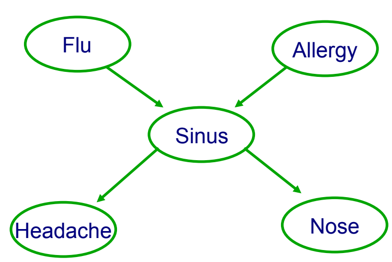
Bayesian Networks – (Structure) Learning

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 31, 2013

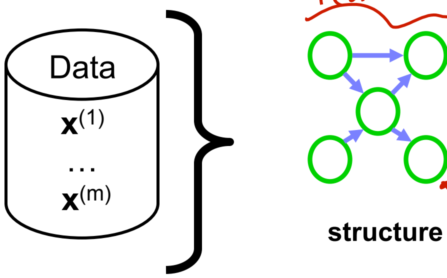
Review

- Bayesian Networks
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
- Fast probabilistic inference
  - As shown in demo examples
  - Compute P(X|e)
- Today
  - Learn BN structure



Learning Bayes nets

learn



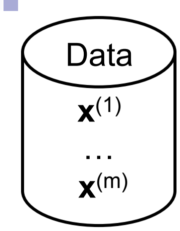
\+ CPTs – $P(X_i | \mathbf{Pa}_{X_i})$

parameters

maximizing likelihood of data

MLE $P(D | \theta_G, G)$

Learning the CPTs

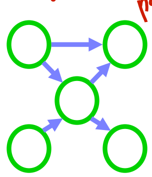


For each discrete variable X_i

$P(S=t | A=t) \stackrel{MLE}{=} \frac{\text{Count}(S=t, A=t)}{\text{Count}(A=t)}$

given we know params

$P(X_i = x_i | Pa_{X_i} = u) \stackrel{MLE}{=} \frac{\text{Count}(X=x_i, Pa_{X_i} = u)}{\text{Count}(u)}$

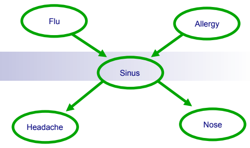


Small subtlety: Count(u) = 0, or small …

need Smoothing / AKA regularization / AKA Bayesian learning

MLE: $P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$

Information-theoretic interpretation of maximum likelihood 1



m data points , n variables

- Given structure, log likelihood of data:

$$\log \underset{G}{\text{maximized}} P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) \overset{iid}{=} \log \prod_{j=1}^{m} P(x_1^{(j)} \cdots x_n^{(j)} \mid \theta_G, G)$$

params structure

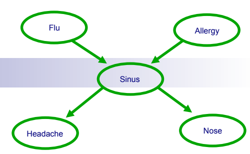
$$= \log \prod_{j=1}^{m} \prod_{i=1}^{n} P(x_i^{(j)} \mid Pa_{x_i,G} = u_i^{(j)})$$

$$= \underbrace{\sum_{j=1}^{m} \sum_{i=1}^{n} \log P(x_i^{(j)} \mid Pa_{x_i,G} = u_i^{(j)})}_{\text{max over } G}$$

$x_i^{(j)}$ ← data points (F=t, A=f, S=t ....)

i ← variables F,A,S,H,N

Information-theoretic interpretation of maximum likelihood 2



- Given structure, log likelihood of data:

$$\log \underset{G}{\max} P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \overbrace{\sum_{j=1}^{m} \sum_{i=1}^{n}}^{\text{flip}} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_{i,G}} = \cancel{\mathbf{x}^{(j)}[\mathbf{Pa}_{X_i}]}\, u_i^{(j)}\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_i^{(j)} \mid Pa_{x_i,G} = u_i^{(j)})$$

e.g. $\sum_{j=1}^{m} \log P(h^{(j)} \mid s^{(j)})$

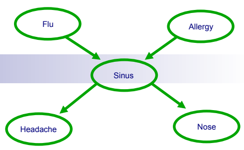
$= (\text{Count}(H=t, S=t) \log P(H=t \mid S=t)$
$+ \text{Count}(H=t, S=f) \log P(H=t \mid S=f)$
$+ \text{Count}(H=f, S=t) \log P(H=f \mid S=t)$
$+ \text{Count}(H=f, S=f) \log P(H=f \mid S=f)$

$$= \sum_{i=1}^{n} \sum_{x_i} \sum_{u_i} \text{Count}(X_i = x_i, Pa_{x_i,G} = u_i) \log P(x_i \mid Pa_{x_i,G} = u_i)$$

$$= m \sum_{i=1}^{n} \underbrace{\sum_{x_i} \sum_{u_i} P(x_i, Pa_{x_i,G} = u_i) \log P(x_i \mid Pa_{x_i,G} = u_i)}_{-H(X_i \mid Pa_{x_i,G})}$$

$$P(x_i, u_i) \overset{MLE}{=} \frac{\text{Count}(X_i = x_i, Pa_{x_i,G} = u_i)}{m}$$

Information-theoretic interpretation of maximum likelihood 3



- Given structure, log likelihood of data:

$$\max_{G} \log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \sum_{x_i, \mathbf{Pa}_{x_i,\mathcal{G}}} \hat{P}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) \log \hat{P}(x_i \mid \mathbf{Pa}_{x_i,\mathcal{G}})$$

$$\max_G = -m \sum_i H(X_i \mid Pa_{X_i,G}) \equiv \min_G m \sum_{i=1}^n H(X_i \mid Pa_{X_i,G})$$

$$\equiv \max_G \underbrace{m \sum_{i=1}^n I(X_i, Pa_{X_i,G})}_{\text{information theoretic criteria over } G \Rightarrow \text{max mutual info between vars \& parents}} \underbrace{- m \sum_{i=1}^n H(X_i)}_{\text{just a constant WRT } G}$$

also measures how dependent vars are

For DTs:

$$H(A \mid B) = -\sum_a \sum_b P(a,b) \log P(a \mid b)$$

if X_i is highly "correlated" with parents $H(X_i \mid Pa_{X_i,G})$ is low $\Rightarrow$ good G

Mutual information

$$I(A,B) = H(A) - H(A \mid B)$$

Decomposable score

- Log data likelihood

constant

$$\max_G \log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- example of a Decomposable score:
  - Decomposes over families in BN (node and its parents)
  - Will lead to significant computational efficiency!!!
  - Score(G : D) = $\sum_i^n$ FamScore(X_i|$\mathbf{Pa}_{X_i}$: D)

e.g. $= \sum_{i=1}^n I(X_i, Pa_{X_i,G})$

but there are many others