Bayesian Networks – Representation

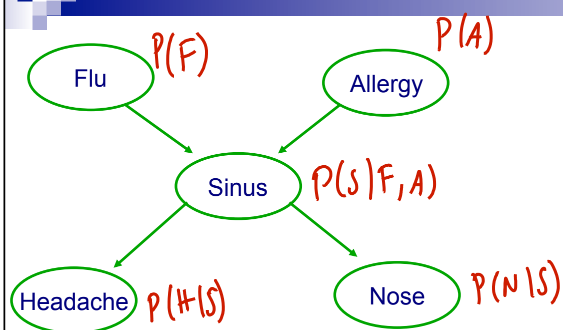
Machine Learning – CSE546

Carlos Guestrin

University of Washington

November 20, 2014

Factored joint distribution - Preview



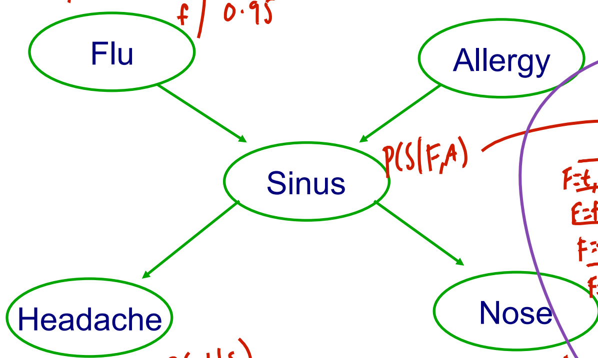
$$\underbrace{P(F,A,S,H,N)}_{2^5-1=31 \text{ params}} = P(F)\,P(A)\,P(S|F,A)\,P(H|S)\,P(N|S)$$

What about probabilities? Conditional probability tables (CPTs)

$P(F)$ ✓

F	P(F)
t	0.05
f	0.95

A	
t	0.2
f	0.8



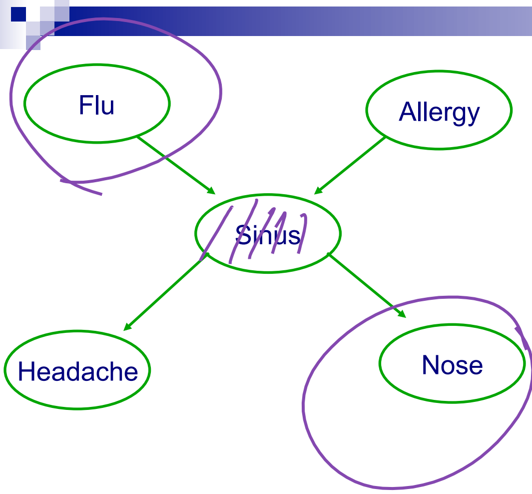
$P(S|F,A)$

	S=t	S=f
F=t, A=t	0.8	0.2
F=f, A=t	0.6	0.4
F=t, A=f	0.5	0.5
F=f, A=f	0.1	0.9

$P(H|S)$

$P(N|S)$

Key: Independence assumptions



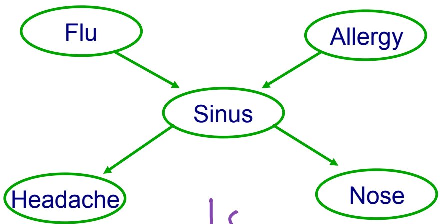
Flu "causes" Nose

Flu only "causes" Nose through Sinus

if N=t, changes Prob F=t
but if I tell you first S=t
N=t doesn't influence prob F=t

Knowing sinus separates the variables from each other

The independence assumption



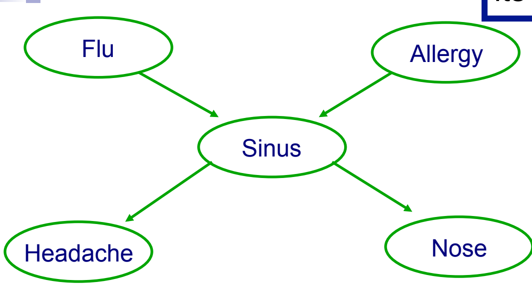
Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parents

$H \perp N \mid S$

	F	A	S	H	N
non descendants	A	F	FA	FAN (S)	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



$F \perp A$

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

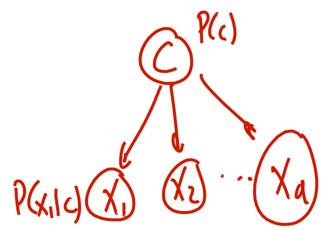
$P(F=t \mid S=t) \neq P(F=t \mid S=t, A=t)$
$>$

Naïve Bayes revisited

$\forall_i \; X_i \perp \{X_2, X_3, \ldots X_d\} \mid C$

$P(C, X_1, \ldots X_d) = P(C) \prod_i P(X_i \mid C)$

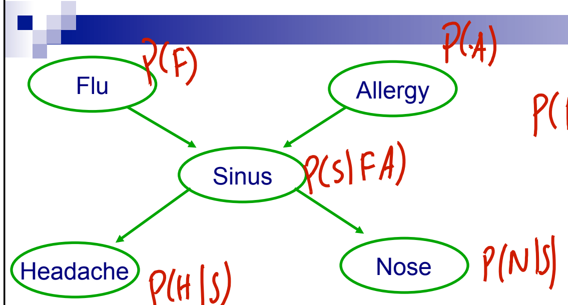
Local Markov Assumption:
A variable X is independent of its non-descendants given its parents



$P(C, X_1, \ldots X_d) = P(C)\, P(X_1 \mid C)\, P(X_2 \mid C) \ldots$

← Naïve Bayes !! ☺

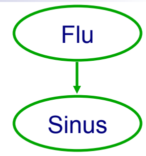
Joint distribution



$P(F, A, S, H, N) = P(A)\, P(F)\, P(S \mid FA)\, P(H \mid S)\, P(N \mid S)$

Why can we decompose? Markov Assumption!

The chain rule of probabilities



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

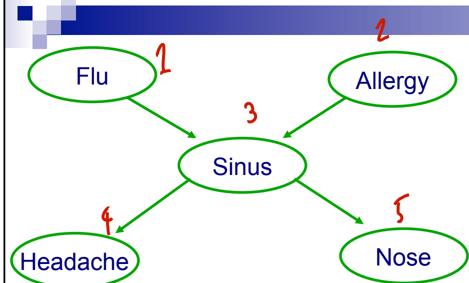
For any dist

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

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

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

Chain rule & Joint distribution



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

Proof by example!, Holds for any BN

$$P(FASHN) = P(F)\,P(A|F)\,P(S|FA)\,P(H|FAS)\,P(N|FASH)$$

$$= P(F)\,P(A)\,P(S|FA)\,P(H|S)\,P(N|S) \quad !! \quad \text{qed}$$

order matters!

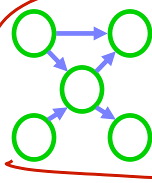
$P(FASHN) = P(H)\,P(N|H)$ . . .

would not get

follow topological order

$A \perp F \Rightarrow P(A|F) = P(A)$ | $H \perp \{FA\} | S \Rightarrow P(H|FAS) = P(H|S)$ | $N \perp \{FAH\} | S \Rightarrow P(N|FASH) = P(N|S)$

The Representation Theorem – Joint Distribution to BN

BN:  **Encodes independence assumptions**

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

Obtain

Joint probability distribution:

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

Two (trivial) special cases

Edgeless graph

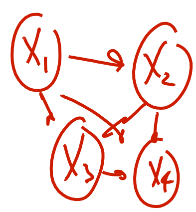
X_1 X_2 X_3 . . .

$X_i \perp$ all others

fewest params

high bias

Fully-connected graph



no independence

structure learning

most params

high variance

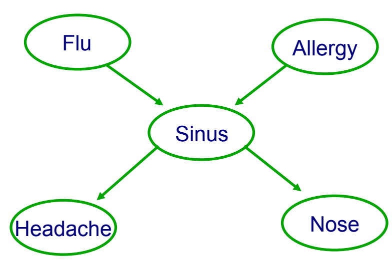
Bayesian Networks – (Structure) Learning

Machine Learning – CSE546
Carlos Guestrin
University of Washington

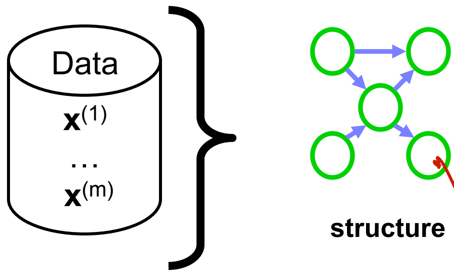
November 20, 2014

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



$+$

CPTs – $P(X_i | \mathbf{Pa}_{X_i})$

parameters

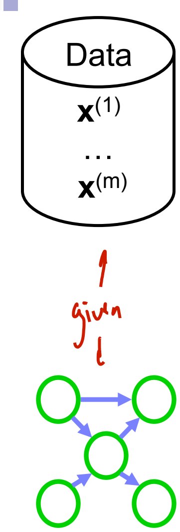
Max likelihood Data

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

Learning the CPTs

var with no parents

$P(X_i = x_i) = \frac{Count(X_i = x_i)}{\# m}$



For each discrete variable X_i

$$P(S=t \mid A=t, F=f) \overset{MLE}{=} \frac{Count(S=t, A=t, F=f)}{Count(A=t, F=f)}$$

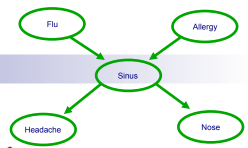
$$P(X_i = x_i \mid Pa_{X_i} = u) \overset{MLE}{=} \frac{Count(X_i = x_i, Pa_{X_i} = u)}{Count(Pa_{X_i} = u)}$$

Subtlety: $Count(Pa_{X_i} = u) = 0$ or very small

→ add smoothing / AKA regularization / AKA Bayesian prior

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



- Given structure, log likelihood of data:

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

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

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

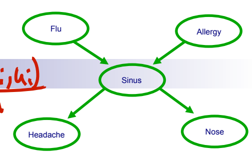
graph defines who parents are

$x_i^{(j)} \leftarrow$ data point $(F=f, A=t, S=t, H=t, N=f)$

Variable F, A, S, H, N

$P(S^{(j)} \mid P(S), G^{(j)})$

Information-theoretic interpretation of maximum likelihood 2



$$\hat{P}(x_i, u_i) = \frac{Count(x_i, u_i)}{m}$$

- Given structure, log likelihood of data:

$$\max_G \log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = u_i^{(j)}\right)$$

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

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

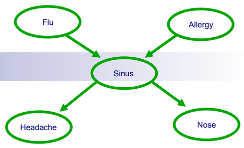
$= Count(H=t, S=t) \log P(H=t \mid S=t)$
$+ Count(H=f, S=t) \log P(H=f \mid S=t)$
$+ Count(H=t, S=f) \log P(H=t \mid S=f)$
$+ Count(H=f, S=f) \log P(H=f \mid S=f)$

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

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

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

Information-theoretic interpretation of maximum likelihood 3



- Given structure, log likelihood of data:

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

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

$$\equiv \max_G \cancel{m} \underbrace{\sum_{i=1}^{n} \hat{I}(X_i, Pa_{X_i,G})}_{\text{Information Theoretic interpretation of MLE for } G} - \underbrace{m \sum_{i=1}^{n} \hat{H}(X_i)}_{\text{doesn't depend on } G}$$

$\Rightarrow \max_G \equiv$ choosing parents with max mutual info with vars

in DTs $I(A,B) = H(A) - H(A|B)$

if $A \perp B$ $I(A,B) = 0$

if A very dependent with B, $I(A,B)$ high

$\hat{I}(X_i, Pa_{X_i,G}) = \hat{H}(X_i) - \hat{H}(X_i \mid Pa_{X_i,G})$

Decomposable score

- Log data likelihood

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

- Decomposable score:
  - Decomposes over families in BN (node and its parents)
  - Will lead to significant computational efficiency!!!
  - Score($G : D$) = $\hat{\sum}_{i=1} \underbrace{\text{FamScore}(X_i|\mathbf{Pa}_{X_i} : D)}_{\hat{I}(X_i, Pa_{X_i,G})}$

How many trees are there?

Nonetheless – Efficient optimal algorithm finds best tree

A → B, B → C, B → D

A→B→C→D

A→B, A → C, B → D

HMM is a tree

A → B, A → C, A → D ← Naive Bayes

every var has at most one parent

For n vars, how many possible trees?

O(n log n)

2

exhaustive search is impossible

$I(A,B) = I(B,A)$

Scoring a tree 1: equivalent trees

$$\max \log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i) = \max_{G} \sum_{i=1}^{n} \hat{I}(X_i, \mathrm{Pa}_{X_i,G})$$

A → B → C | A ← B ← C | A ← B → C ← all are trees

score = $0 + I(B,A) + I(C,B)$ | same | score = $I(B,C) + I(A,B)$ | same | score = $I(A,B) + I(B,C)$

same skeleton

$A \perp C \mid B$ | $A \perp C \mid B$ | $A \perp C \mid B$

represent same distribution

no true for all edge directions: A → B ← C $A \perp C \mid \emptyset$

not a tree, because B has 2 parents

Scoring a tree 2: similar trees

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

$A \to B \to C$ score $= I(A,B) + I(B,C)$

$B \leftarrow A \to C$ score $= I(A,B) + I(A,C)$

in general for trees : $\max_G \text{score} = \max_G \sum_{(i,j)\in E} I(X_i, X_j)$

Chow-Liu tree learning algorithm 1

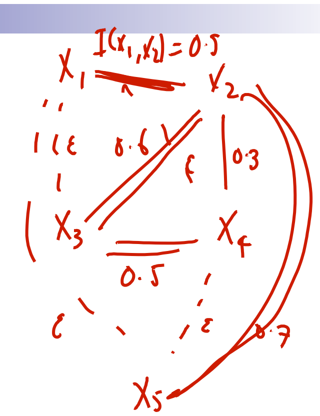
- For each pair of variables X_i,X_j
  - Compute empirical distribution:

$$\hat{P}(x_i, x_j) = \frac{\text{Count}(x_i, x_j)}{m}$$

  - Compute mutual information:

$$\hat{I}(X_i, X_j) = \sum_{x_i,x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

- Define a graph
  - Nodes $X_1,\ldots,X_n$
  - Edge (i,j) gets weight $\hat{I}(X_i, X_j)$



D run max spanning tree ← complexity is about $O(E \log E)$

$O(n^2 \log n)$

Chow-Liu tree learning algorithm 2

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

- Optimal tree BN
  - Compute maximum weight spanning tree
  - Directions in BN: pick any node as root, breadth-first-search defines directions