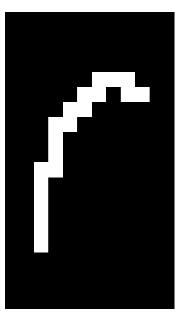
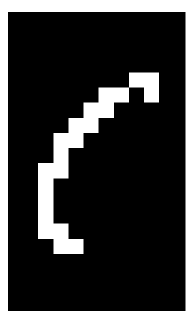
Bayesian Networks – Representation

Machine Learning – CSE546
Carlos Guestrin
University of Washington

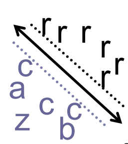
November 25, 2013

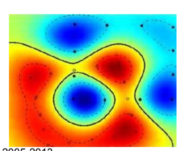
Handwriting recognition





Character recognition, e.g., kernel SVMs





Webpage classification

→ Company home page

vs

Personal home page

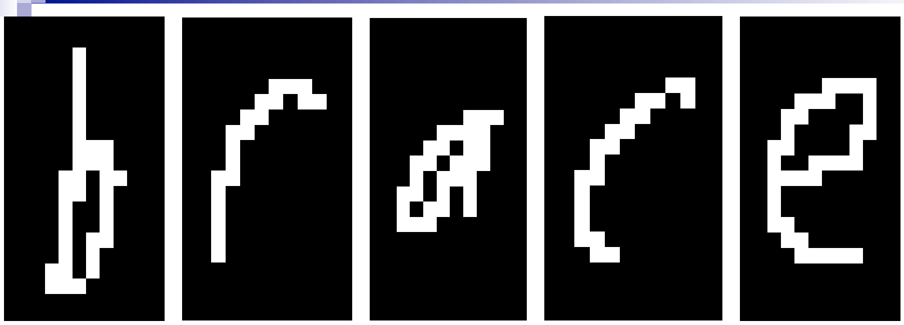
vs

University home page

vs

…

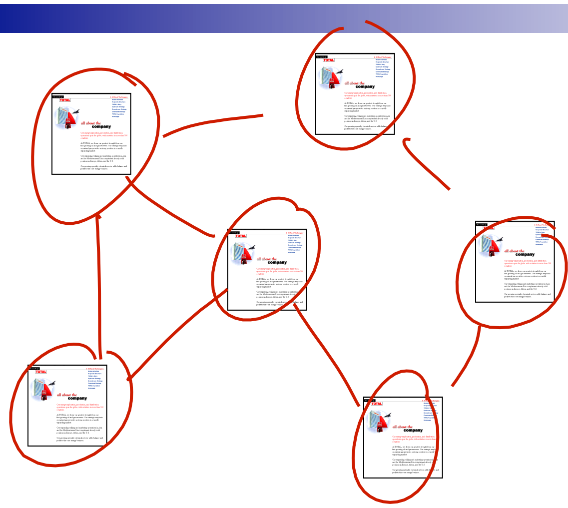
Handwriting recognition 2



"c" comes after an "a" much more

often than after a "z"

Webpage classification 2

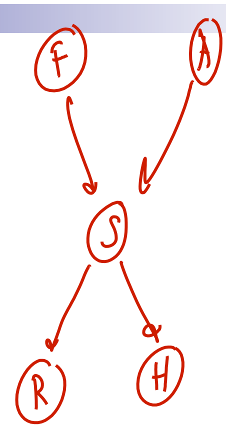


Today – Bayesian networks

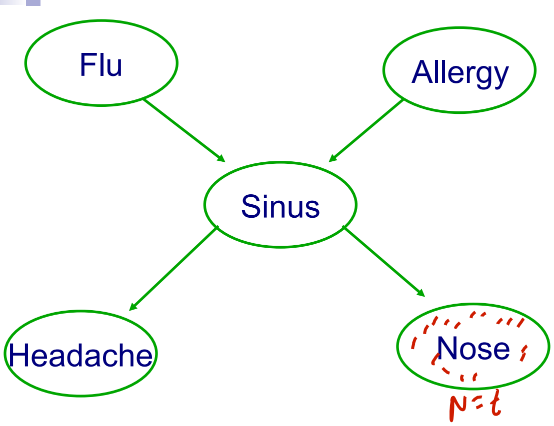
- One of the most exciting advancements in statistical AI in the last decades
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

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

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

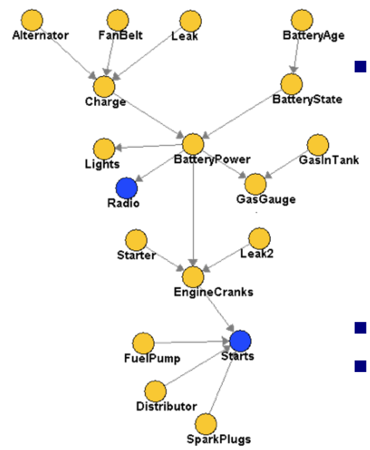
- Most probable explanation

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

- Active data collection

  given N=t, which test should I perform

Car starts BN



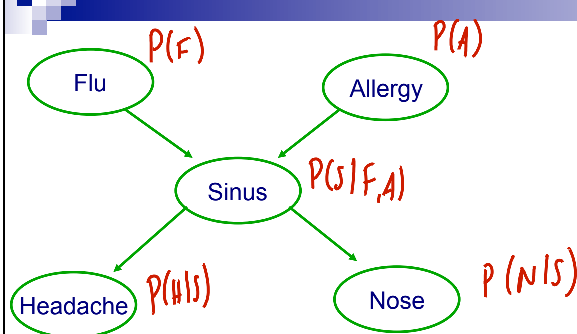
- 18 binary attributes

  2^{18} possibilities
- Inference
  - P(BatteryAge|Starts=f)

$$P(BA \mid Starts=f) = \sum_{a,f,l,c,bs\ldots} P(a,fb,l,c,\ldots BA \mid Starts=f)$$

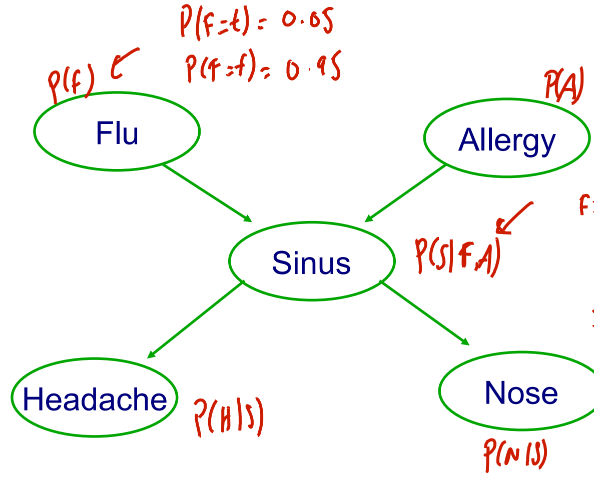
- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution - Preview



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

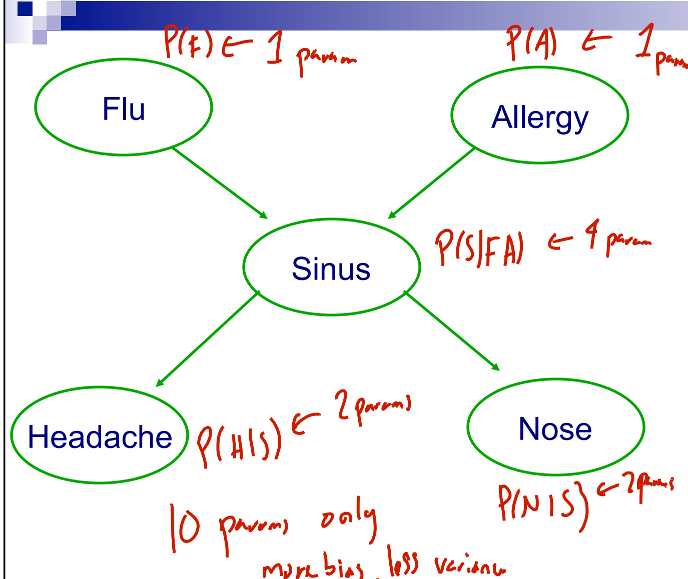
What about probabilities?
Conditional probability tables (CPTs)



A	
t	0.2
f	0.8

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

Number of parameters

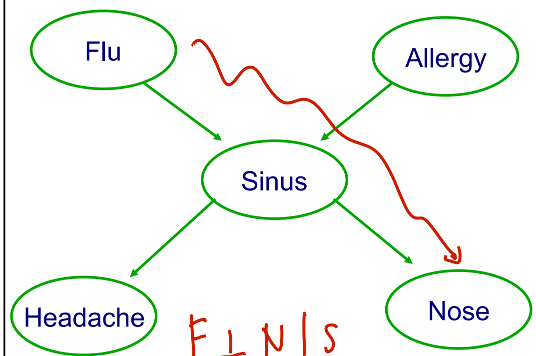


$\underbrace{P(A\,F\,S\,H\,N)}$

$2^5 - 1$ params

$= 31$ params

Key: Independence assumptions



$F \perp N \mid S$

Flu "causes" Nose through Sinus

if $N=t$, "changes" prob F from $P(F)$

if first tell you $S=f$, then $N=t$, prob of F is unchanged from $P(F \mid S=f)$

Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A$

$\Downarrow$

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

OR

$P(A \mid F) = P(A)$

Flu = t	.2
Flu = f	.8

Allergy = t	.4
Allergy = f	.6

	Flu = t	Flu = f
Allergy = t	$.4 \times .2 = 0.08$	$.4 \times .8$
Allergy = f	$.6 \times .2$	$.8 \times .6$

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

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

    $\Updownarrow\ P(X|Y) = P(X)$

Conditional independence

- Flu and Headache are ~~not~~ (marginally) independent

  $P(H=t \mid F=t) \neq P(H=t) \qquad F \not\perp H$
- Flu and Headache are independent given Sinus infection $\quad F \perp H \mid S$

  $P(H=t \mid S=t) = P(H=t \mid S=t, F=t)$
- More Generally: $\quad X \perp Y \mid Z$

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

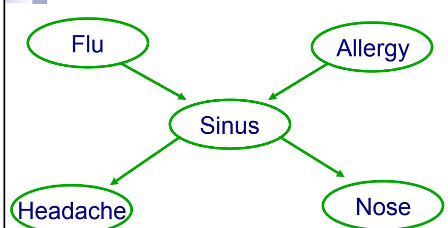
  $\Updownarrow$

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

Conditionally independent random variables

- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if
  - $P \vdash (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}), \forall \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$
- Shorthand:
  - **Conditional independence:** $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \varnothing)$, write $P \vdash (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** P statisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$

The independence assumption

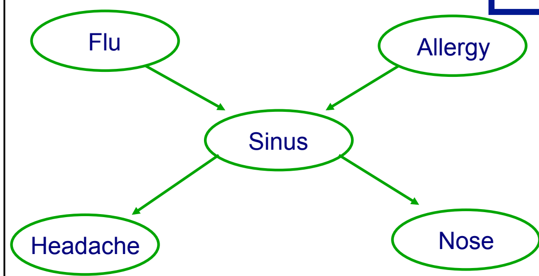


Local Markov Assumption:
A variable X is independent of its non-descendants minus parents given its parents (and only its parents)

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

Explaining away

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



$F \perp A$

$\boxed{F \perp A \mid S \; ???}$ No!

$P(F=t \mid S=t, A=f) \neq P(F=t \mid S=t)$ No

if it's not allergy, prob it's flu

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

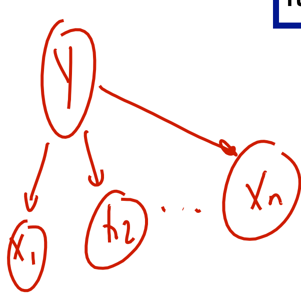
Naïve Bayes revisited

Variables: $Y, X_1, \ldots X_n$

$X_1 \perp \{X_2 \ldots X_n\} \mid Y$

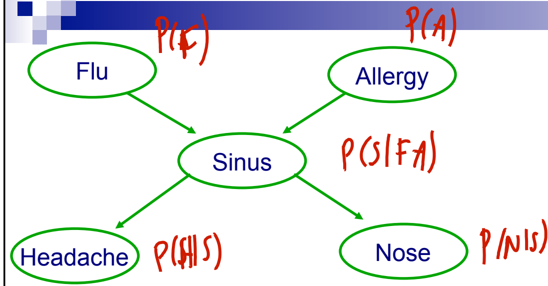
$\vdots$

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



Naive Bayes!! :)

Joint distribution



Why??

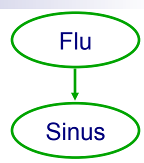
P(A H F S N) = P(F) P(A) P(S|FA) P(H|S) P(N|S)

Why can we decompose? Markov Assumption!

The chain rule of probabilities

no indep assumption

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



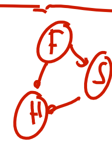
For any distribution:

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

Same as what Bayes Net Says

e.g. P(F S H) = P(F) P(S|F) P(H|SF)

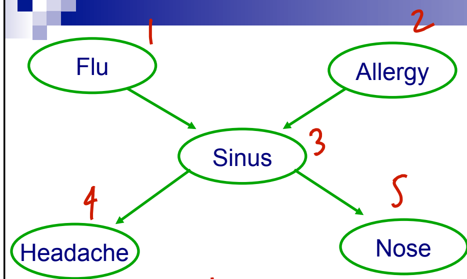
any distribution



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

always true

Chain rule & Joint distribution



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

could have picked another ordering for chain rule, but for proof to work must pick any topological ordering

$$P(F\,A\,S\,H\,N) = P(F)\,\underbrace{P(A|F)}_{P(A)}\,P(S|F,A)\,\underbrace{P(H|SFA)}_{P(H|S)}\,\underbrace{P(N|HSFA)}_{P(N|S)}$$

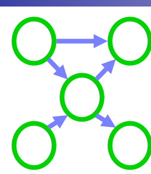
$A \perp F \Rightarrow P(A|F) = P(A)$

$H \perp \{FA\} | S \Rightarrow P(H|SFA) = P(H|S)$

$N \perp \{FAH\} | S \Rightarrow P(N|HSFA) = P(N|S)$

The Representation Theorem – Joint Distribution to BN

BN:



Encodes independence assumptions

there is also a converse theorem

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

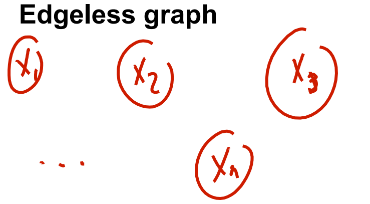
Obtain

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

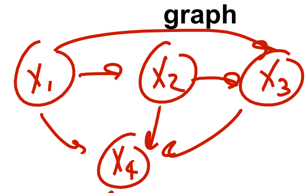


$X_i \perp$ {everybody else}

all variables indep.

all the bias in the world

Fully-connected graph



$X_i \perp \{ \underset{\emptyset}{\text{nobody}} \} \mid$ parents

no independence assumptions

$\Rightarrow$ represent anything

but exponentially many parameters

$\Rightarrow$ lots of variance (no bias)