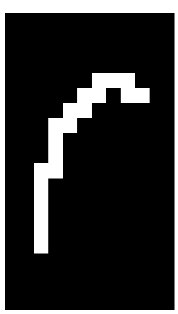
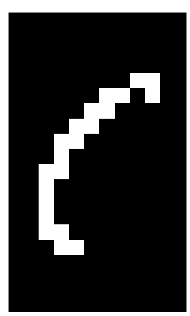
Bayesian Networks – Representation

Machine Learning – CSE446
Carlos Guestrin
University of Washington

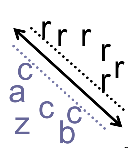
May 29, 2013

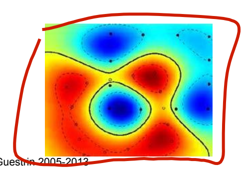
Handwriting recognition





Character recognition, e.g., kernel SVMs





Webpage classification

→ Company home page

vs

Personal home page

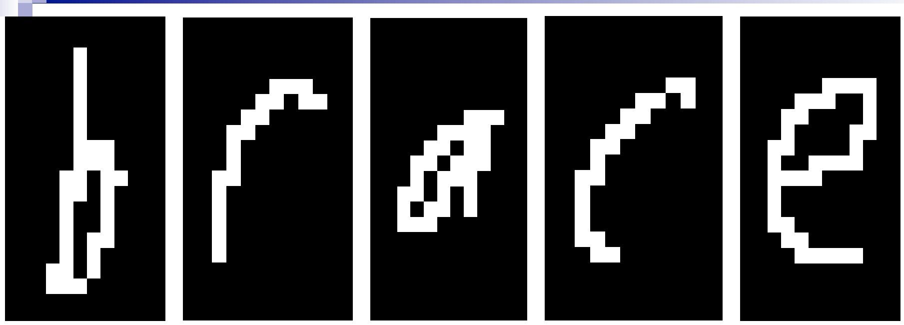
vs

University home page

vs

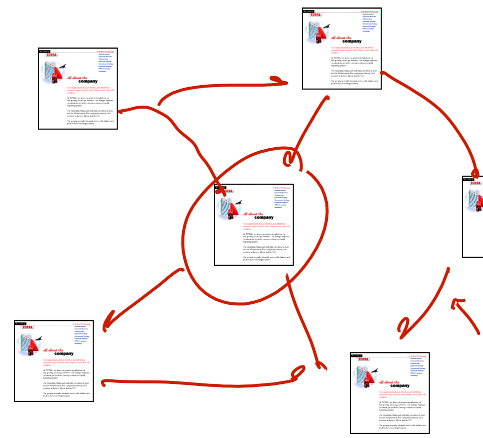
… ;

Handwriting recognition 2



"c" more likely to come after an "a" than an "q"

Webpage classification 2



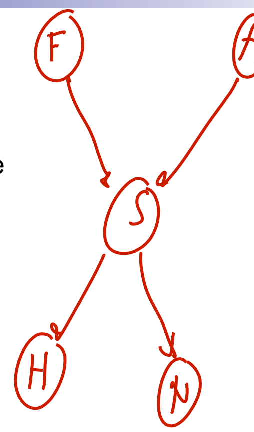
Company pages tend to point to each other

Today – Bayesian networks

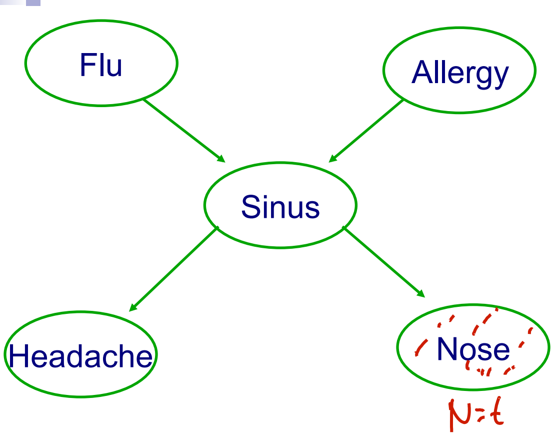
- One of the most exciting advancements in statistical AI in the last decades
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

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

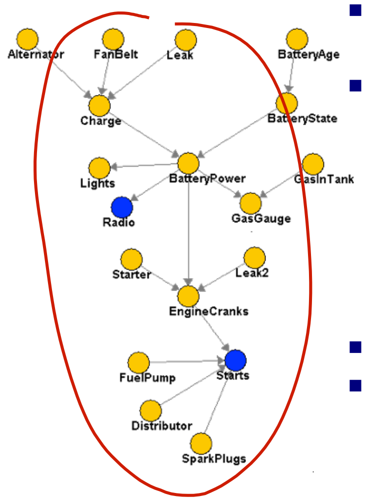
  $\max_{f,a,s,h} P(f,s,a,h \mid N=t)$

- Active data collection

  what variable should I observe next?

  H=?, S=?

Car starts BN



- 18 binary attributes

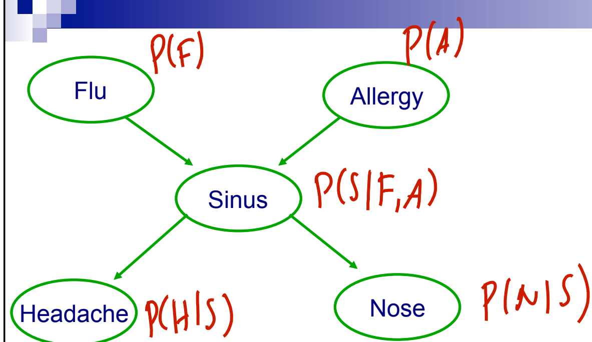
  2^{18} probabilities
- Inference
  - P(BatteryAge|Starts=f)

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

2^{16}

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

Factored joint distribution - Preview



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

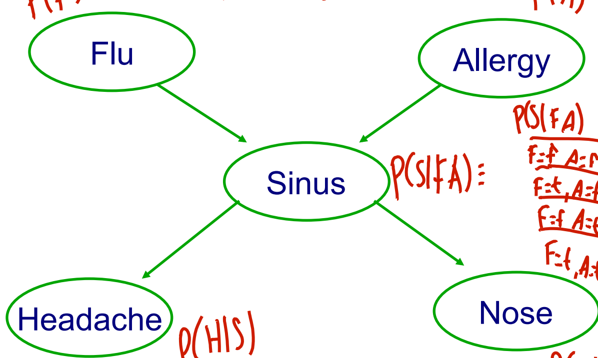
31 params

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



$P(S|FA)$:

$P(S\|FA)$	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

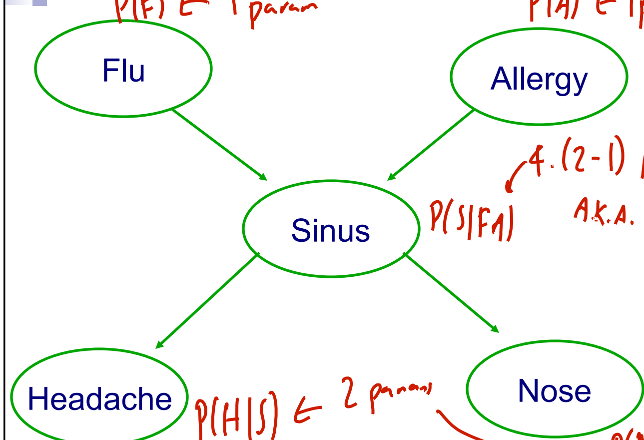
$P(H|S)$

$P(N|S)$

Number of parameters

$P(F)$ ← 1 param

$P(A)$ ← 1 params



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

$P(H|S)$ ← 2 params

$P(N|S)$

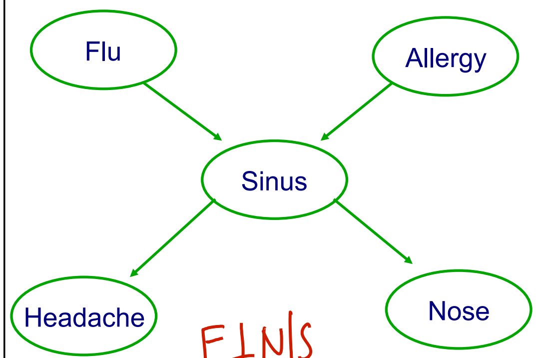
Total: 10 params

$P(FASHN)$ → 31 params

10 < 31

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

Key: Independence assumptions



$F \perp N \mid S$

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

$F \perp A$

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

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

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
  - $P \vdash (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y}), \forall \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y})$ entails

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

- Shorthand:
  - **Marginal independence:** $P \vdash (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** P statisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}) = P(\mathbf{X})\,P(\mathbf{Y})$

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

Conditional independence

- Flu and Headache are not (marginally) independent

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

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

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

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

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

Conditionally independent random variables

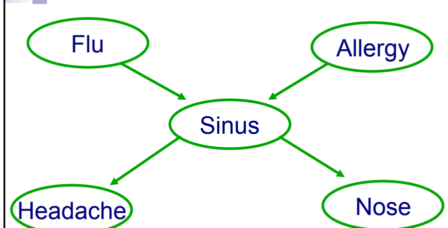
- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if
  - $P \vdash (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} | \mathbf{Z}=\mathbf{z})$, $\forall \mathbf{x} \in Val(\mathbf{X})$, $\mathbf{y} \in Val(\mathbf{Y})$, $\mathbf{z} \in Val(\mathbf{Z})$

    $P(X=x | Y=y, Z=z) = P(X=x | Z=z)$
- Shorthand:
  - **Conditional independence:** $P \vdash (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$
  - For $P \vdash (\mathbf{X} \perp \mathbf{Y} | \emptyset)$, write $P \vdash (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** P statisfies $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$

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

The independence assumption

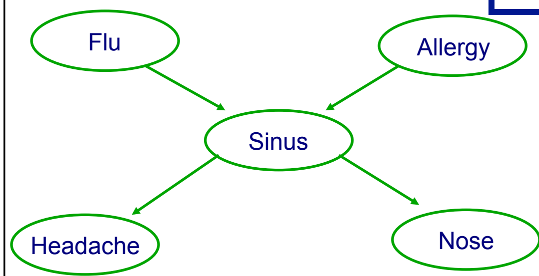


Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parent

	F	A	S	H	N
non.desc	A	F	F A	F A N	F A H
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!!

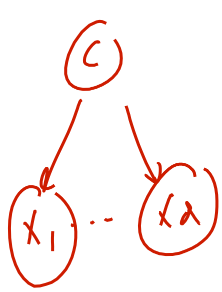
Suppose $P(F=t \mid S=t)$ is high
but $P(F=t \mid S=t, A=t)$ is lower
because $A=t$ explain away sinus infection

Naïve Bayes revisited

$X_1 \perp \{X_2 \dots X_d\} \mid C$

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

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



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

← naive Bayes!!