VC Dimension

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

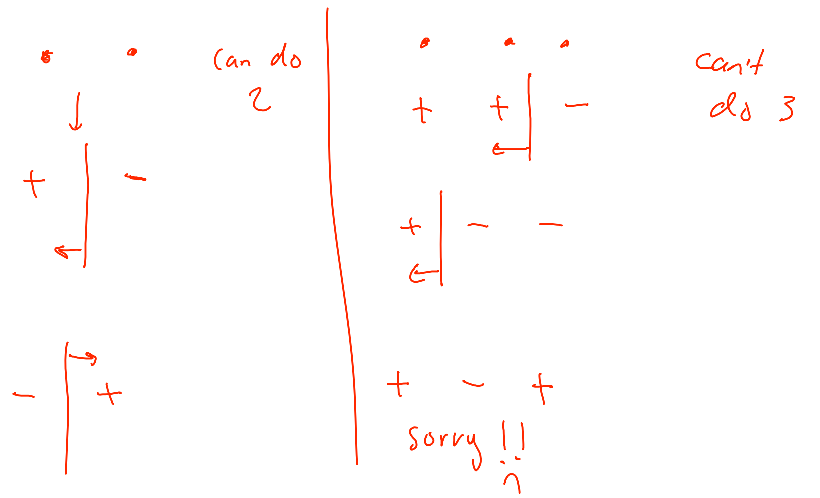
October 29th, 2007

What about continuous hypothesis spaces?

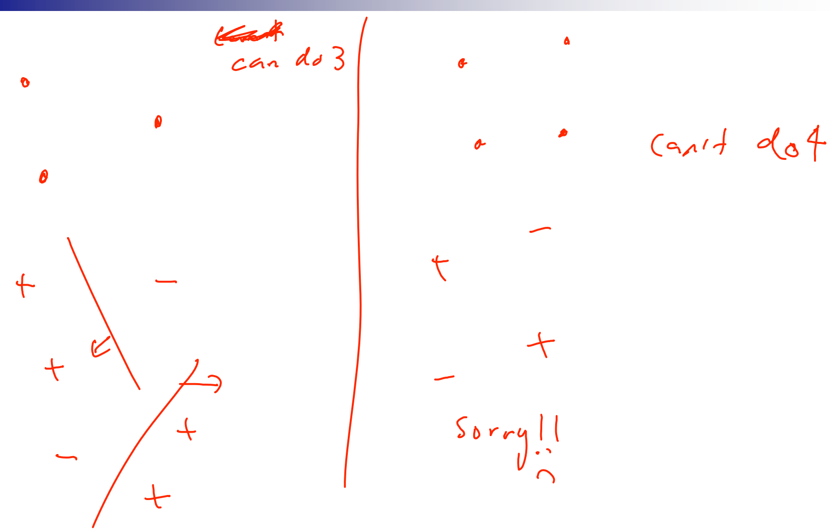
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???
- **As with decision trees, only care about the maximum number of points that can be classified exactly!**

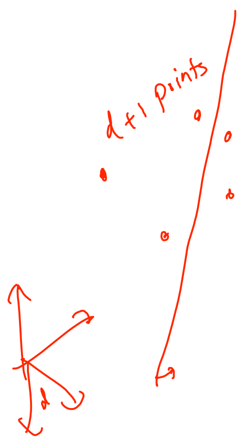
How many points can a linear boundary classify exactly? (1-D)



How many points can a linear boundary classify exactly? (2-D)



How many points can a linear boundary classify exactly? (d-D)



can do d+1 points

how many parameters in a linear classifier in d-dimensions?

$$w_0 + \sum_{i=1}^{d} w_i x_i$$

d+1

PAC bound using VC dimension

- **Number of training points that can be classified exactly is VC dimension!!!**
  - **Measures relevant size of hypothesis space, as with decision trees with k leaves**

e.g., linear classifiers — may be continuous

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

only depend on VC(H)
not on |H|

Shattering a set of points

Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

Definition: a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

VC dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

PAC bound using VC dimension

- **Number of training points that can be classified exactly is VC dimension!!!**
  - **Measures relevant size of hypothesis space, as with decision trees with k leaves**
  - **Bound for infinite dimension hypothesis spaces:**

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

Examples of VC dimension

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term *b*
- Neural networks
  - VC(H) = #parameters
  - Local minima means NNs will probably not find best parameters
- 1-Nearest neighbor?

Another VC dim. example - What can we shatter?

- What's the VC dim. of decision stumps in 2d?

Another VC dim. example - What can't we shatter?

- What's the VC dim. of decision stumps in 2d?

What you need to know

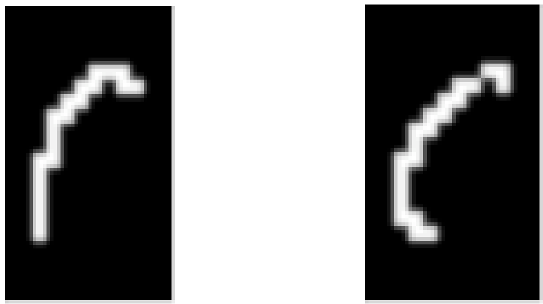
- Finite hypothesis space
  - Derive results
  - Counting number of hypothesis
  - Mistakes on Training data
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – decision trees
  - Infinite case – VC dimension
- Bias-Variance tradeoff in learning theory
- Remember: will your algorithm find best classifier?

Bayesian Networks – Representation

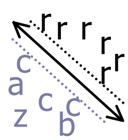
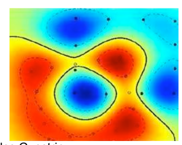
Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

October 29th, 2007

Handwriting recognition



Character recognition, e.g., kernel SVMs

Webpage classification

→ Company home page

vs

Personal home page

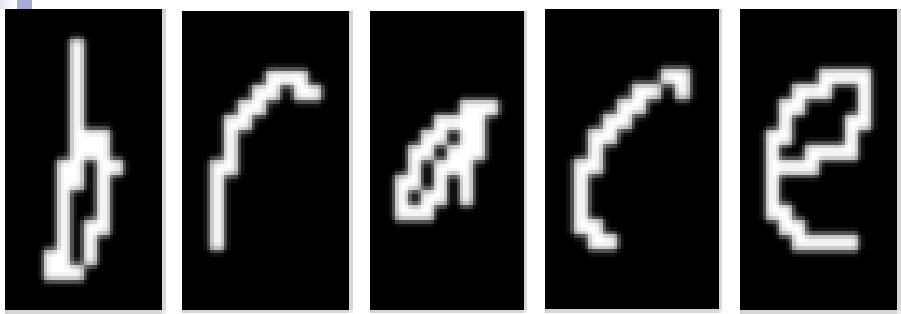
vs

University home page

vs

…

Handwriting recognition 2



©2005-2007 Carlos Guestrin 17

Webpage classification 2

©2005-2007 Carlos Guestrin 18

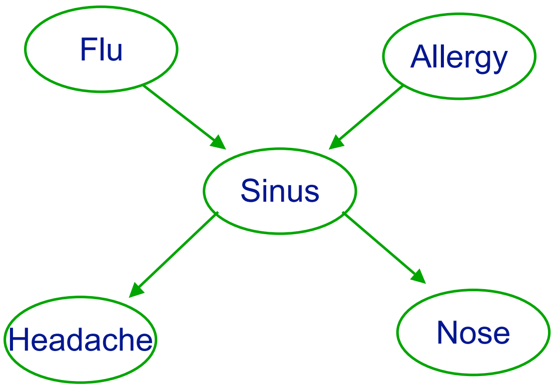
Today – Bayesian networks

- One of the most exciting advancements in statistical AI in the last 10-15 years
- 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?

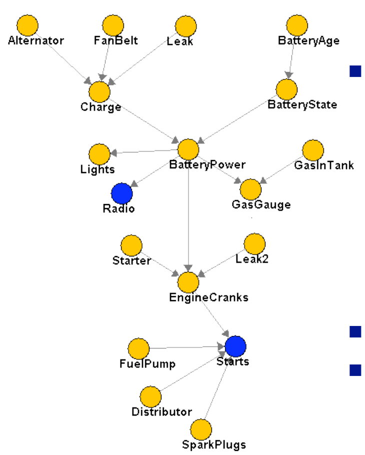
Possible queries



- Inference
- Most probable explanation
- Active data collection

©2005-2007 Carlos Guestrin

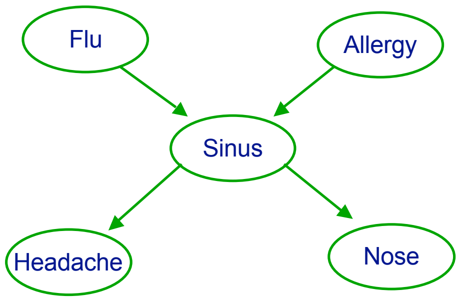
Car starts BN



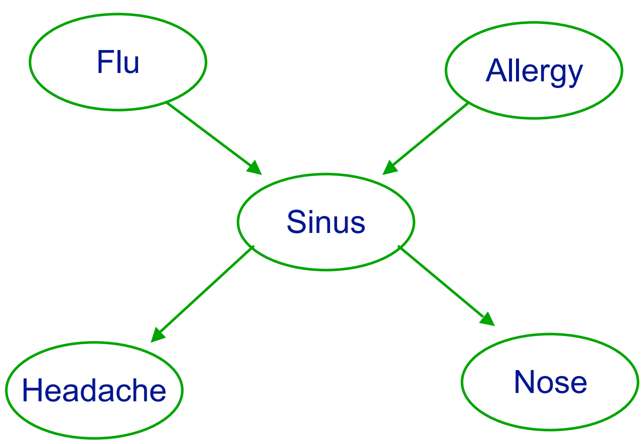
- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f)
- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

©2005-2007 Carlos Guestrin

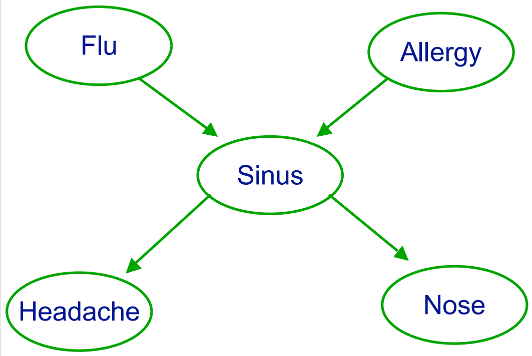
Factored joint distribution - Preview



Number of parameters



Key: Independence assumptions



Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

Flu = t	
Flu = f	

- More Generally:

Allergy = t	
Allergy = f	

	Flu = t	Flu = f
Allergy = t		
Allergy = f		

Marginally independent random variables

- **Sets** of variables **X**, **Y**
- X is independent of Y if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y}), \forall\ \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y})$
- Shorthand:
  - **Marginal independence:** $P \models (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** *P* satisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}) = P(\mathbf{X})\ P(\mathbf{Y})$

Conditional independence

- Flu and Headache are not (marginally) independent
- Flu and Headache are independent given Sinus infection
- More Generally:

Conditionally independent random variables

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

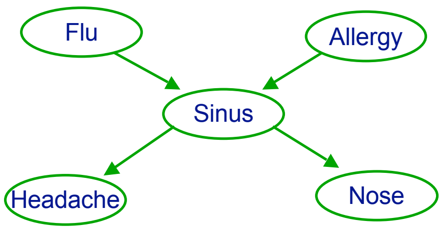
- Shorthand:
  - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$
  - For $P \models (\mathbf{X} \perp \mathbf{Y} | \emptyset)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$

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

Properties of independence

- **Symmetry:**
  - $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X} | \mathbf{Z})$
- **Decomposition:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} | \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$
- **Weak union:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} | \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z},\mathbf{W})$
- **Contraction:**
  - $(\mathbf{X} \perp \mathbf{W} | \mathbf{Y},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} | \mathbf{Z})$
- **Intersection:**
  - $(\mathbf{X} \perp \mathbf{Y} | \mathbf{W},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{W} | \mathbf{Y},\mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} | \mathbf{Z})$
  - Only for positive distributions!
  - $P(\alpha)>0, \forall\alpha, \alpha\neq\emptyset$

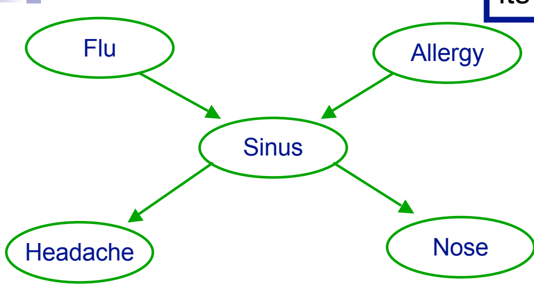
The independence assumption



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

Explaining away

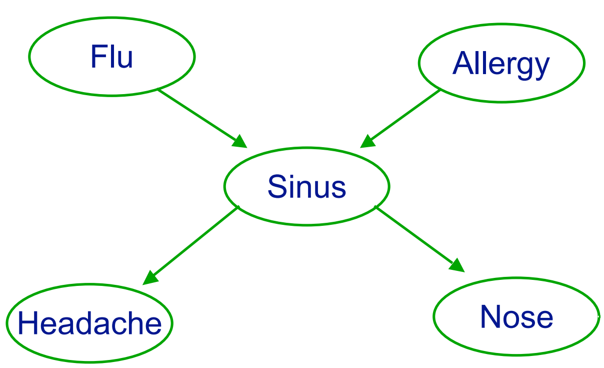
Local Markov Assumption:
A variable X is independent of its non-descendants given its parents



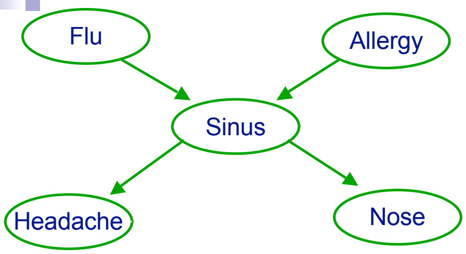
Naïve Bayes revisited

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

What about probabilities?
Conditional probability tables (CPTs)

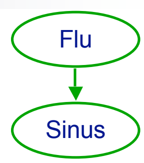


Joint distribution



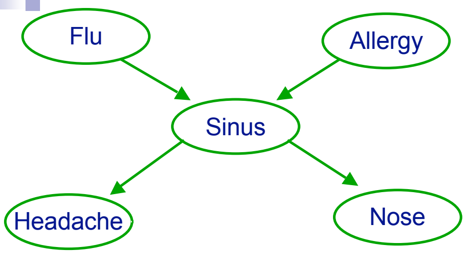
Why can we decompose? Markov Assumption!

The chain rule of probabilities



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

Chain rule & Joint distribution



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

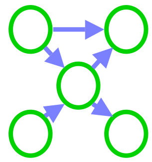
Two (trivial) special cases

Edgeless graph

Fully-connected graph

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:

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

Real Bayesian networks applications

- Diagnosis of lymph node disease
- Speech recognition
- Microsoft office and Windows
  - http://www.research.microsoft.com/research/dtg/
- Study Human genome
- Robot mapping
- Robots to identify meteorites to study
- Modeling fMRI data
- Anomaly detection
- Fault dianosis
- Modeling sensor network data

A general Bayes net

- Set of random variables
- Directed acyclic graph
  - Encodes independence assumptions
- CPTs
- Joint distribution:

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

How many parameters in a BN?

- Discrete variables $X_1, \ldots, X_n$
- Graph
  - Defines parents of X_i, $\mathbf{Pa}_{X_i}$
- CPTs – $P(X_i | \mathbf{Pa}_{Xi})$

Another example

- Variables:
  - B – Burglar
  - E – Earthquake
  - A – Burglar alarm
  - N – Neighbor calls
  - R – Radio report
- Both burglars and earthquakes can set off the alarm
- If the alarm sounds, a neighbor may call
- An earthquake may be announced on the radio

Another example – Building the BN

- B – Burglar
- E – Earthquake
- A – Burglar alarm
- N – Neighbor calls
- R – Radio report

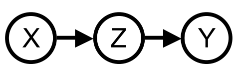
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - ($X_i \perp NonDescendants_{Xi} \mid \mathbf{Pa}_{Xi}$)
- But then we talked about other (in)dependencies
  - e.g., explaining away
- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!

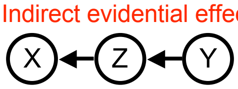
Understanding independencies in BNs – BNs with 3 nodes

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

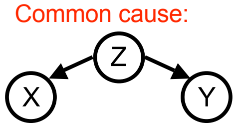
Indirect causal effect:



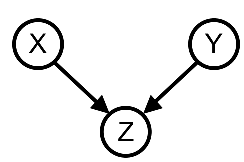
Indirect evidential effect:



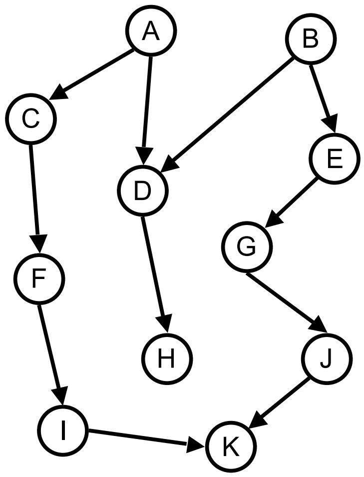
Common cause:



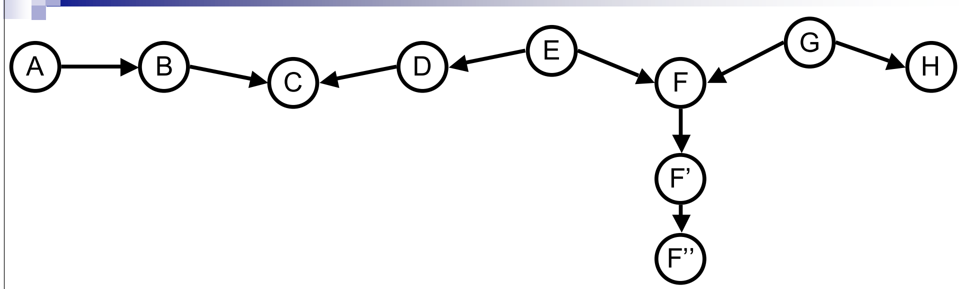
Common effect:



Understanding independencies in BNs – Some examples



An active trail – Example



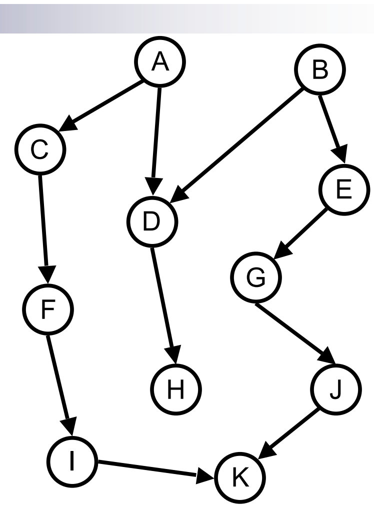
When are A and H independent?

Active trails formalized

- A path $X_1 - X_2 - \cdots - X_k$ is an **active trail** when variables $\boldsymbol{O} \subseteq \{X_1,\ldots,X_n\}$ are observed if for each consecutive triplet in the trail:
  - $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and X_i is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, and X_i **is observed** ($X_i \in \boldsymbol{O}$), or **one of its descendents**

Active trails and independence?

- **Theorem**: Variables $\mathbf{X_i}$ **and** $\mathbf{X_j}$ **are independent given** $\boldsymbol{Z} \subseteq \{X_1,\ldots,X_n\}$ if the is **no active trail** between X_i and X_j when variables $\boldsymbol{Z} \subseteq \{X_1,\ldots,X_n\}$ are observed



The BN Representation Theorem

If conditional independencies in BN are subset of conditional independencies in *P*	Obtain →	**Joint probability distribution:** $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$

Important because:
Every *P* has at least one BN structure *G*

If joint probability distribution: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$	Obtain →	**Then conditional independencies in BN are subset of conditional independencies in *P***

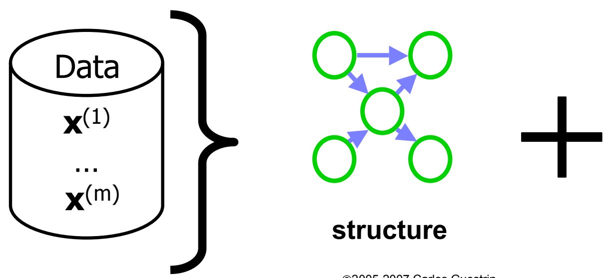
Important because:
Read independencies of *P* from BN structure *G*

"Simpler" BNs

- A distribution can be represented by many BNs:
- Simpler BN, requires fewer parameters

Learning Bayes nets

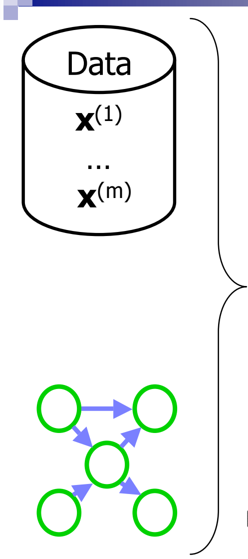
	Known structure	Unknown structure
Fully observable data		
Missing data		



CPTs –
$P(X_i | \mathbf{Pa}_{X_i})$

parameters

Learning the CPTs



For each discrete variable X_i

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

Queries in Bayes nets

- Given BN, find:
  - Probability of X given some evidence, P(X|e)
  - Most probable explanation, $\max_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid e)$
  - Most informative query
- Learn more about these next class

What you need to know

- Bayesian networks
  - A compact **representation** for large probability distributions
  - Not an algorithm
- Semantics of a BN
  - Conditional independence assumptions
- Representation
  - Variables
  - Graph
  - CPTs
- Why BNs are useful
- Learning CPTs from fully observable data
- Play with applet!!! ☺

Acknowledgements

- JavaBayes applet
  - http://www.pmr.poli.usp.br/ltd/Software/javabayes/Home/index.html

