ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

- Bayes Nets
  - (Finish) Parameter Learning
  - Structure Learning

Readings: KF 18.1, 18.3; Barber 9.5, 10.4

Dhruv Batra

Virginia Tech

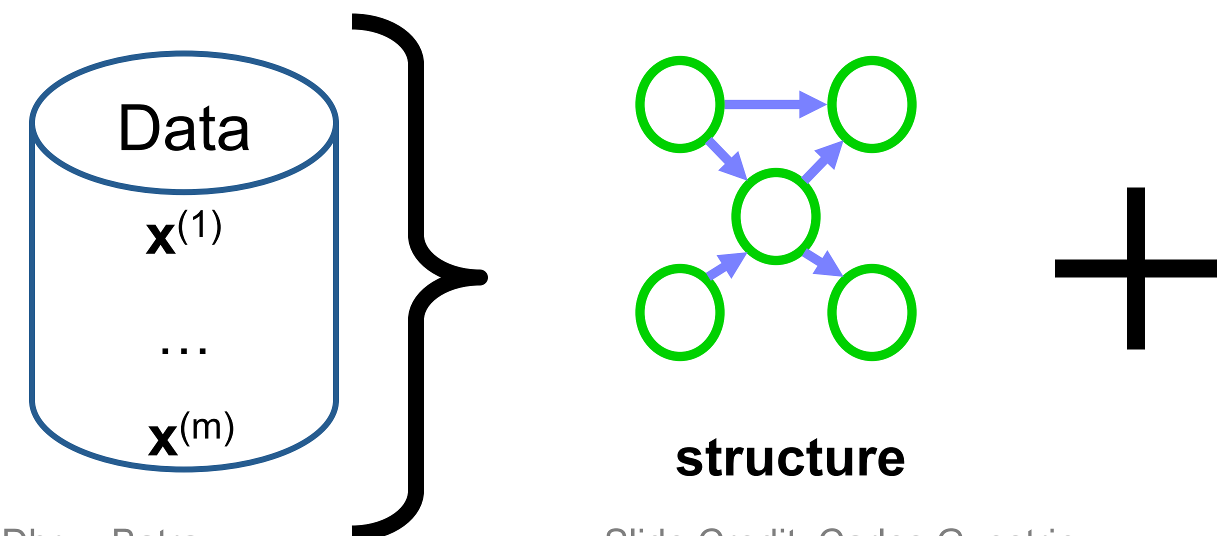
Administrativia

- HW1
  - Out
  - Due in 2 weeks: ~~Feb 17~~, Feb 19, 11:59pm
  - Please please please please start early
  - Implementation: TAN, structure + parameter learning
  - Please post questions on Scholar Forum.

Recap of Last Time

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data	Very easy	Hard
Missing data	Somewhat easy (EM)	Very very hard

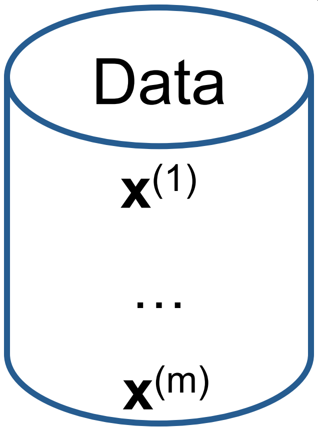


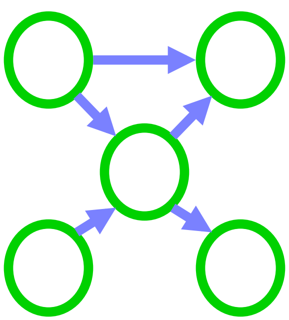
structure

CPTs – $P(X_i | \mathbf{Pa}_{X_i})$

parameters

Learning the CPTs





For each discrete variable X_i

$$\hat{P}_{MLE}(X_i = a \mid \mathrm{Pa}_{X_i} = b) = \frac{\mathrm{Count}(X_i = a, \mathrm{Pa}_{X_i} = b)}{\mathrm{Count}(\mathrm{Pa}_{X_i} = b)}$$

Plan for today

- (Finish) BN Parameter Learning
  - Parameter Sharing
  - Plate notation

- (Start) BN Structure Learning
  - Log-likelihood score
  - Decomposability
  - Information never hurts

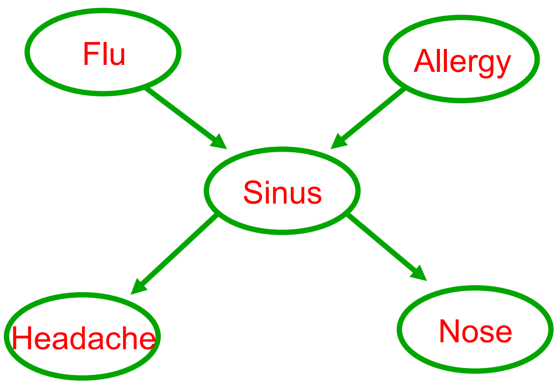
Meta BN

- Explicitly showing parameters as variables

- Example on board
  - One variable X; parameter θ_X
  - Two variables X,Y; parameters θ_X, $\theta_{Y|X}$

Global parameter independence

- **Global parameter independence:**
  - All CPT parameters are independent
  - Prior over parameters is product of prior over CPTs



- **Proposition**: For fully observable data *D*, if prior satisfies global parameter independence, then

$$P(\theta \mid \mathcal{D}) = \prod_i P(\theta_{X_i \mid \mathbf{Pa}_{X_i}} \mid \mathcal{D})$$

Parameter Sharing

- What if $X_1, \ldots, X_n$ are n random variables for coin tosses of the same coin?

Naïve Bayes vs Bag-of-Words

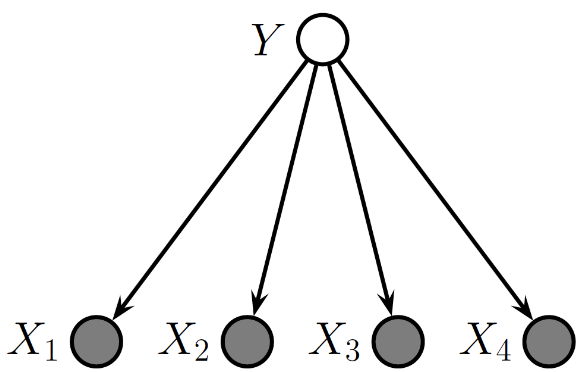
- What’s the difference?

- Parameter sharing!

Text classification

- Classify e-mails
  - Y = {Spam,NotSpam}

- What about the features **X**?
  - X_i represents i^{th} word in document; i = 1 to doc-length
  - X_i takes values in vocabulary, 10,000 words, etc.



$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of Words

- **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - Order of words on the page ignored
  - Parameter sharing

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

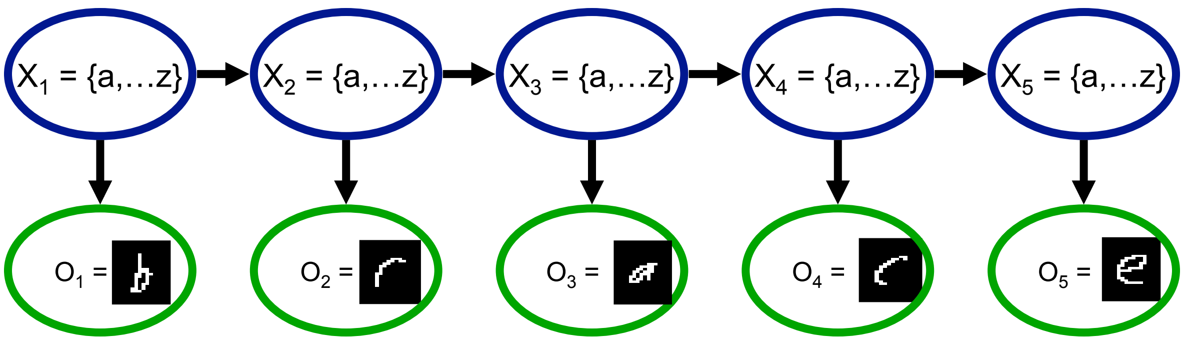
Bag of Words

- **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - Order of words on the page ignored
  - Parameter sharing

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

in is lecture lecture next over person remember room sitting the the the to to up wake when you

HMMs semantics: Details



Just 3 distributions:

$$P(X_1)$$

$$P(X_i \mid X_{i-1})$$

$$P(O_i \mid X_i)$$

N-grams

- Learnt from Darwin's *On the Origin of Species*

Unigrams

1	0.16098	_
2	0.06687	a
3	0.01414	b
4	0.02938	c
5	0.03107	d
6	0.11055	e
7	0.02325	f
8	0.01530	g
9	0.04174	h
10	0.06233	i
11	0.00060	j
12	0.00309	k
13	0.03515	l
14	0.02107	m
15	0.06007	n
16	0.06066	o
17	0.01594	p
18	0.00077	q
19	0.05265	r
20	0.05761	s
21	0.07566	t
22	0.02149	u
23	0.00993	v
24	0.01341	w
25	0.00208	x
26	0.01381	y
27	0.00039	z

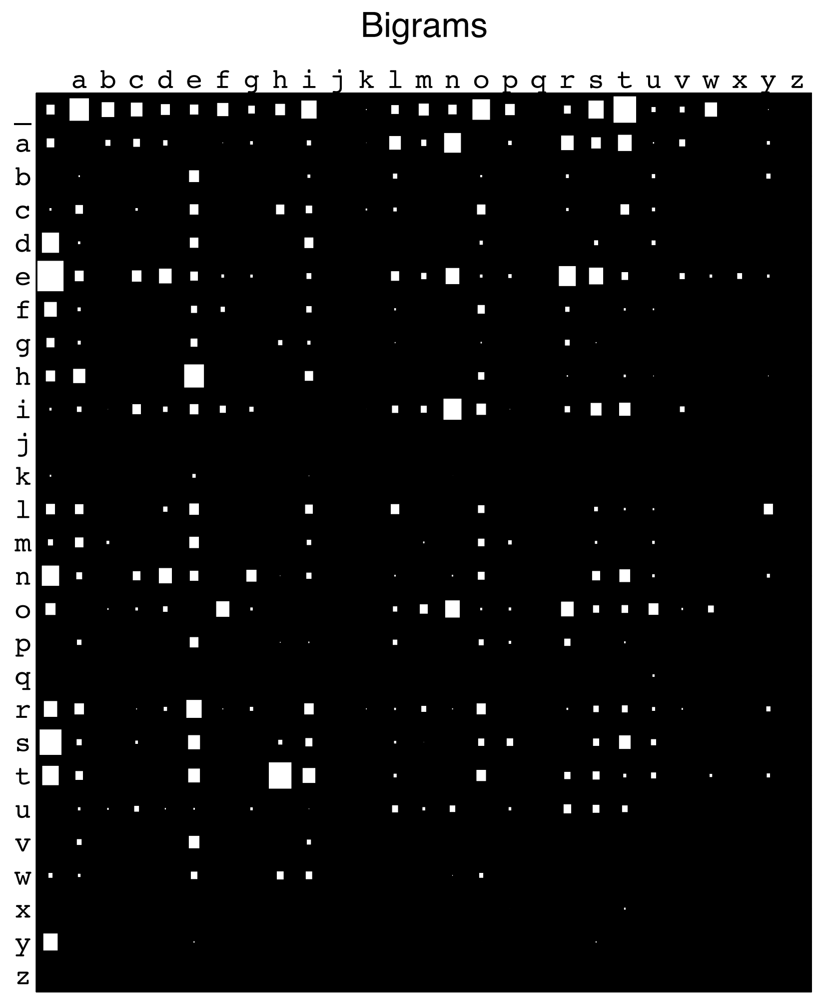
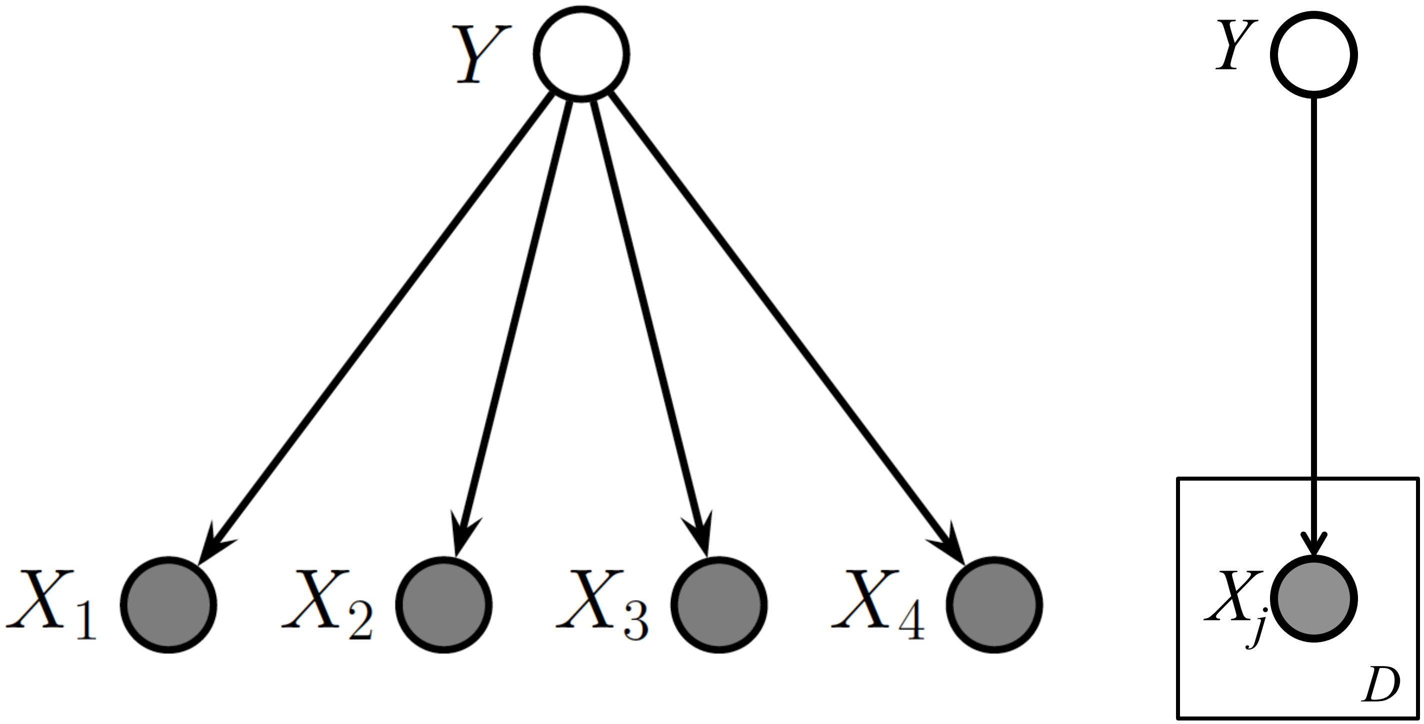


Plate Notation

- $X_1, \ldots, X_n$ are n random variables for coin tosses of the same coin

- Plate denotes replication

Plate Notation



Plates denote replication of random variables

Hierarchical Bayesian Models

- Why stop with a single prior?

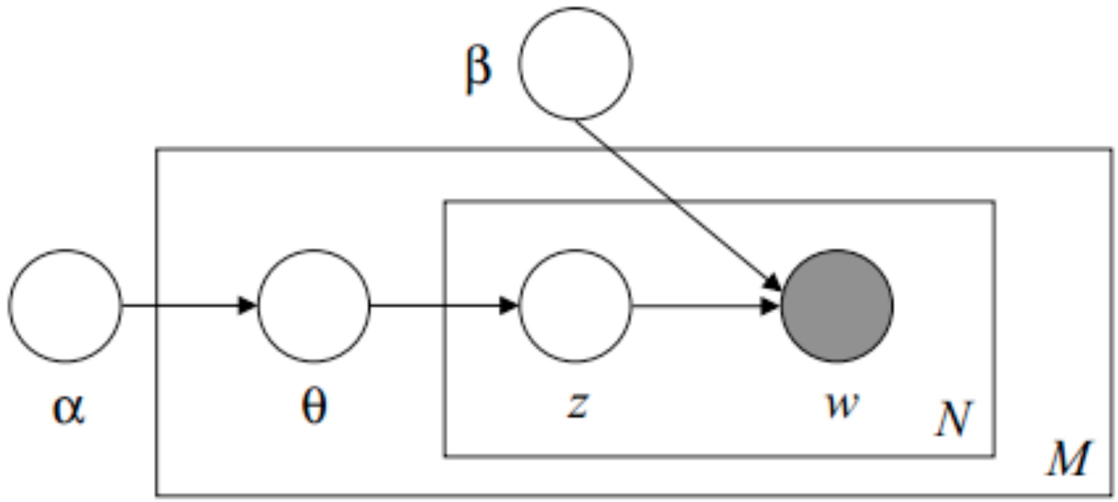


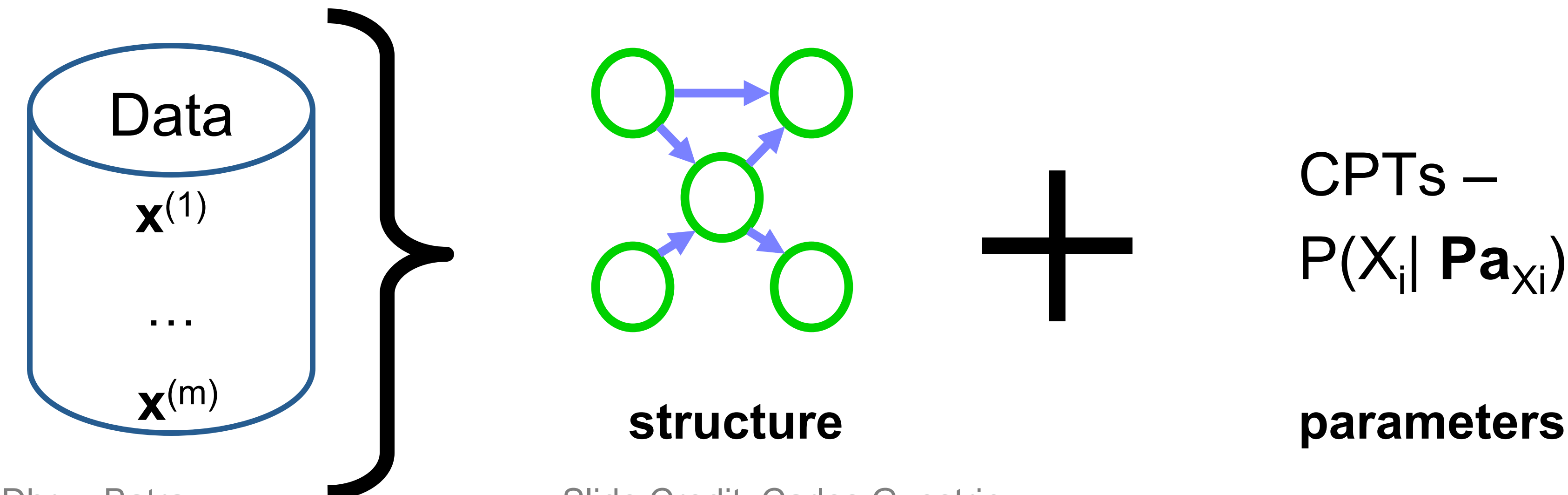
Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

BN: Parameter Learning: What you need to know

- Parameter Learning
  - MLE
    - Decomposes; results in counting procedure
    - Will shatter dataset if too many parents
  - Bayesian Estimation
    - Conjugate priors
    - Priors = regularization (also viewed as smoothing)
    - Hierarchical priors
  - Plate notation
  - Shared parameters

Learning Bayes nets

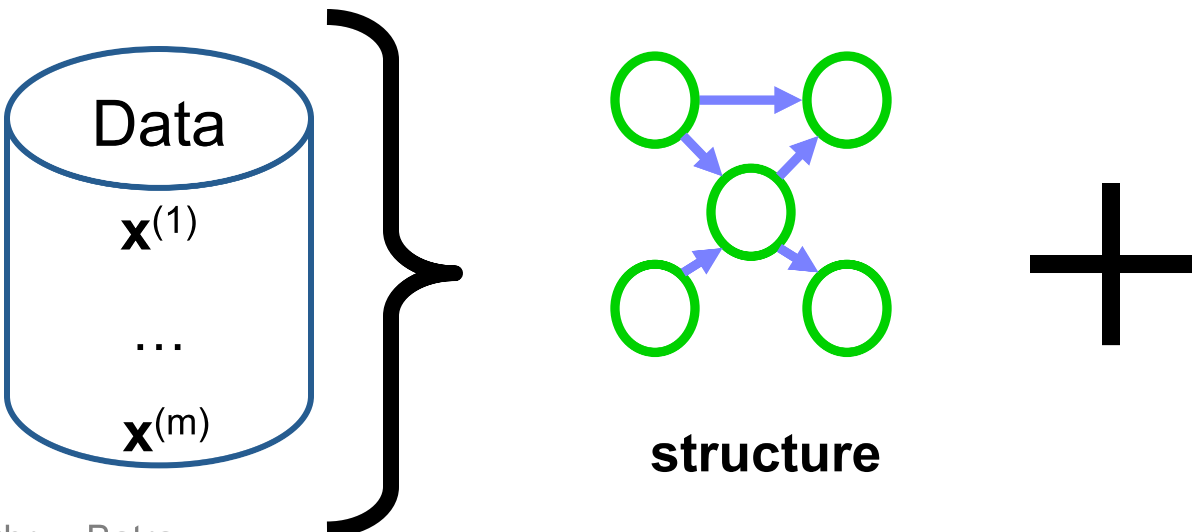
	Known structure	Unknown structure
Fully observable data	Very easy	Hard
Missing data	Somewhat easy (EM)	Very very hard



Goals of Structure Learning

- Prediction
  - Care about a good structure because presumably it will lead to good predictions

- Discovery
  - I want to understand some system

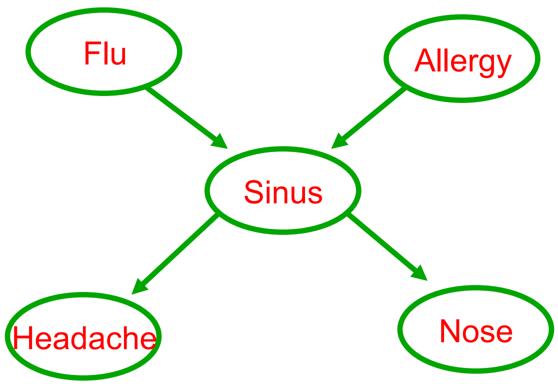


CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

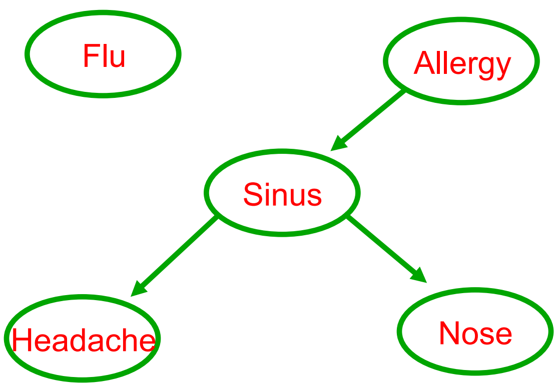
parameters

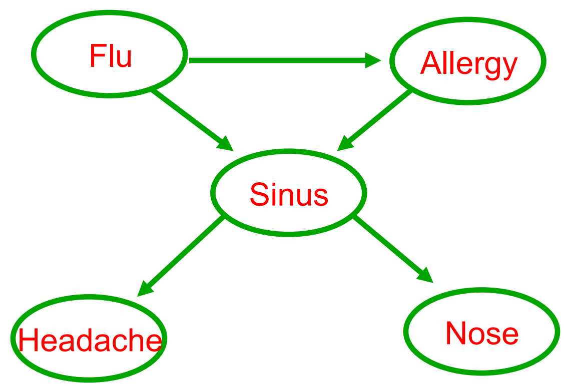
Types of Errors

- Truth:

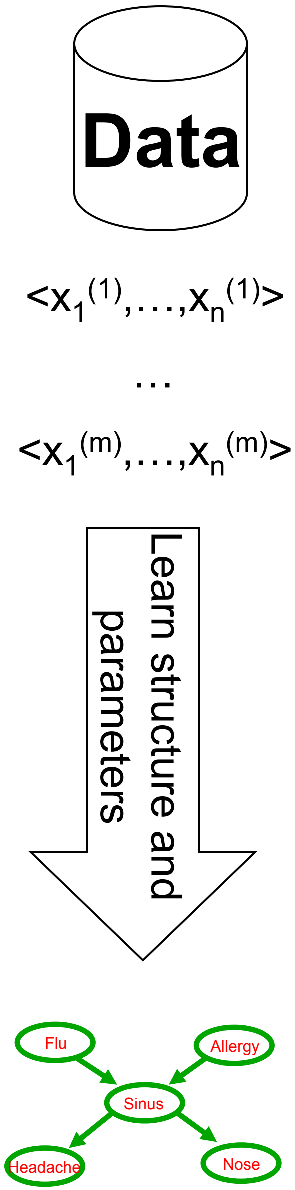


- Recovered:



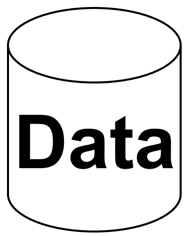


Learning the structure of a BN



- **Constraint-based approach**
  - Test conditional independencies in data
  - Find an I-map
- **Score-based approach**
  - Finding a structure and parameters is a density estimation task
  - Evaluate model as we evaluated parameters
    - Maximum likelihood
    - Bayesian
    - etc.

Score-based approach

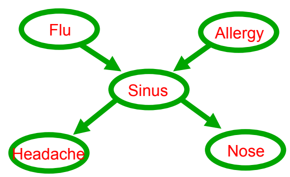


$<x_1^{(1)},\ldots,x_n^{(1)}>$

…

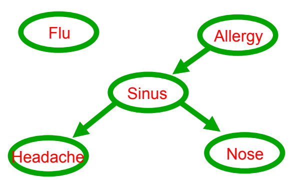
$<x_1^{(m)},\ldots,x_n^{(m)}>$

Possible structures



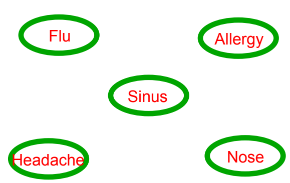
Learn parameters

Score structure -52



Learn parameters

Score structure -60



Learn parameters

Score structure -500

How many graphs?

- N vertices.
- How many (undirected) graphs?
- How many (undirected) trees?

What's a good score?

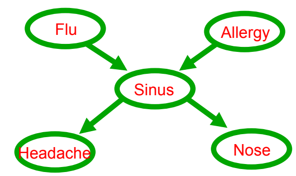
- Score(G) = log-likelihood(G : D, θ_{MLE})

Information-theoretic interpretation of Maximum Likelihood Score

- Consider two node graph
  - Derived on board

Information-theoretic interpretation of Maximum Likelihood Score

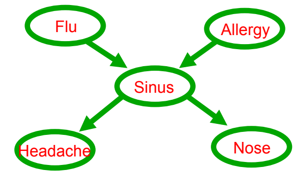
- For a general graph G



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

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

Information-theoretic interpretation of Maximum Likelihood Score



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

- Implications:
  - Intuitive: higher mutual info → higher score
  - Decomposes over families in BN (node and it's parents)
  - Same score for I-equivalent structures!
  - Information never hurts!

Chow-Liu tree learning algorithm 1

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

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

  - Compute mutual information:

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

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

Chow-Liu tree learning algorithm 2

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

Can we extend Chow-Liu?

- Tree augmented naïve Bayes (TAN) [Friedman et al. '97]
  - Naïve Bayes model overcounts, because correlation between features not considered
  - Same as Chow-Liu, but score edges with:

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