CSE 575: Statistical Machine Learning

Jingrui He

CIDSE, ASU

Graphical Models

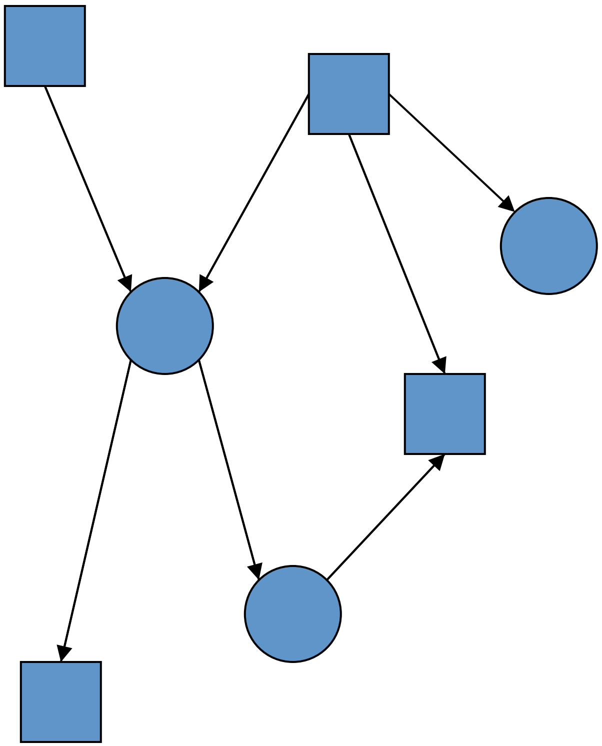
What is a graphical model ?

A graphical model is a way of representing probabilistic relationships between random variables.

Conditional (in)dependencies are represented by (missing) edges:

Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):



“Graphical models are a **marriage between probability theory and graph theory**.

They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering – **uncertainty and complexity** –

and in particular they are playing an increasingly important role in the design and analysis of **machine learning algorithms**.

Fundamental to the idea of a graphical model is the **notion of modularity** – a complex system is built by combining simpler parts.

The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

This view has many advantages -- in particular, **specialized techniques** that have been developed in one field can be **transferred between research communities** and exploited more widely.

Moreover, the graphical model formalism provides a **natural framework for the design of new systems**."

--- Michael Jordan, 1998.

What can we do with graphical models?

- Graphs are an intuitive way of representing and visualizing the relationships between many variables. (Examples: family trees, electric circuit diagrams, neural networks)

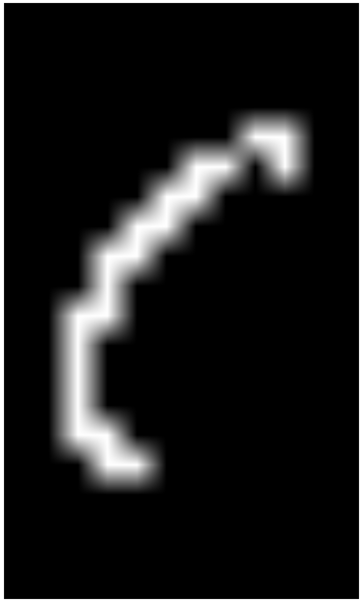
- Graphical models allow us to define general message-passing algorithms that implement probabilistic inference efficiently. Thus we can answer queries like "What is P(A|C = c)?" without enumerating all settings of all variables in the model.

- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can answer questions like: "Is A dependent of B given that we know the value of C ?" just by looking at the graph.

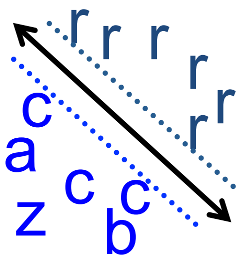
Applications of graphical models

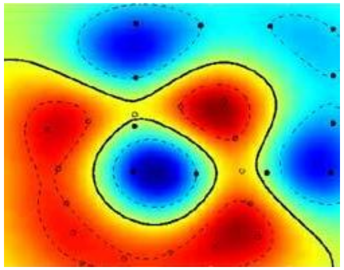
- Handwriting recognition
- Webpage classification
- Information extraction
- Speech recognition
- Computer vision
- Modeling of gene regulatory networks
- Gene finding and diagnosis of diseases
- Graphical models for protein structure

Handwriting recognition 1



Character recognition, e.g., kernel SVMs





Handwriting recognition 2

Webpage classification 1

Company home page

vs

Personal home page

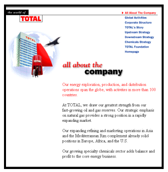
vs

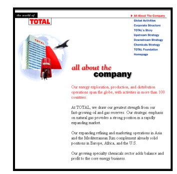
University home page

vs

…

Webpage classification 2





Probability Distributions

- Let $X_1,\ldots,X_p$ be discrete random variables
- Let P be a joint distribution over $X_1,\ldots,X_p$
- If the variables are binary, then we need $O(2^p)$ parameters to describe P
- Can we do better?
  - **Key idea**: use properties of independence

Independent Random Variables

- Two variables X and Y are **independent** if
  - $P(X = x|Y = y) = P(X = x)$ for all values x, y
  - That is, learning the values of Y does not change prediction of X

- If X and Y are independent then
  - $P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$

- In general, if $X_1,\ldots,X_p$ are independent, then
  - $P(X_1,\ldots,X_p)= P(X_1)\ldots P(X_p)$

Conditional Independence

- Unfortunately, most of random variables of interest are not independent of each other

- A more suitable notion is that of **conditional independence**

- Two variables X and Y are **conditionally independent** given Z if
  - $P(X = x|Y = y, Z=z) = P(X = x|Z=z)$ for all values x,y,z
  - That is, learning the values of Y does not change prediction of X once we know the value of Z
  - notation: $X \perp Y \mid Z$

Example: Naïve Bayes Model

- A common model in early diagnosis:
  - Symptoms are conditionally independent given the disease (or fault)
- Thus, if
  - $X_1,\ldots,X_p$ denote whether the symptoms are exhibited by the patient (headache, high-fever, etc.) and
  - H denotes the hypothesis about the patient's health
- then, $P(X_1,\ldots,X_p,H) = P(H)P(X_1|H)\ldots P(X_p|H)$,

- This **Naïve Bayes** model allows compact representation
  - It does make strong independence assumptions

Probabilistic Graphical Models I

- Probabilities play a central role in modern pattern recognition.
- The probabilistic inference and learning may be complex.
- It is advantageous to augment the analysis using diagrammatic representations of probability distributions, called probabilistic graphical models.

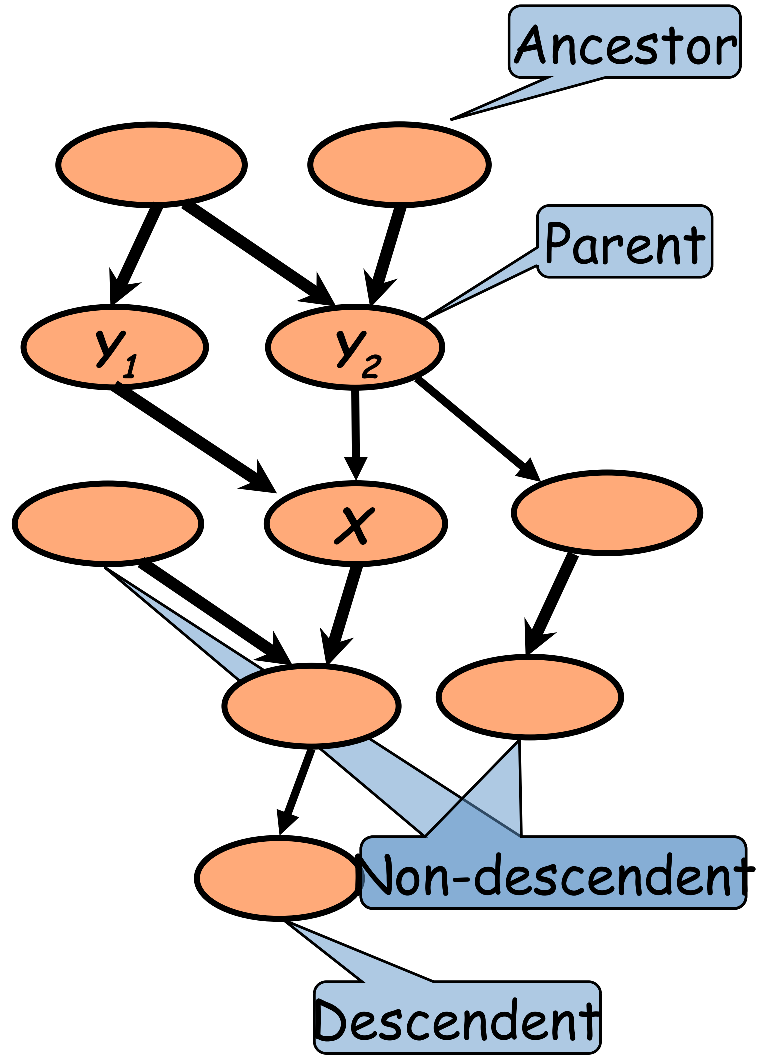
Probabilistic Graphical Models II

- Insights into the properties of the model, including **conditional independence properties**, can be obtained by inspection of the graph.

- Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

A Few Definitions

- Nodes (vertices) + links (arcs, edges)
  - Node: a random variable
  - Link: a probabilistic relationship
- Directed graphical models or Bayesian networks.
- Undirected graphical models or Markov random fields.



Different Types of BN

- ***Directed: Bayesian Networks***
  - ***E.g., Hidden Markov Model***
- Undirected: Markov Random Field
  - E.g., Restricted/Deep Boltzmann Machine
  - E.g., Conditional Random Fields
- Hybrid Graphical Models
  - E.g., Deep Belief Networks
  - E.g., Hierarchical-Deep Models

Bayesian Networks
Representation

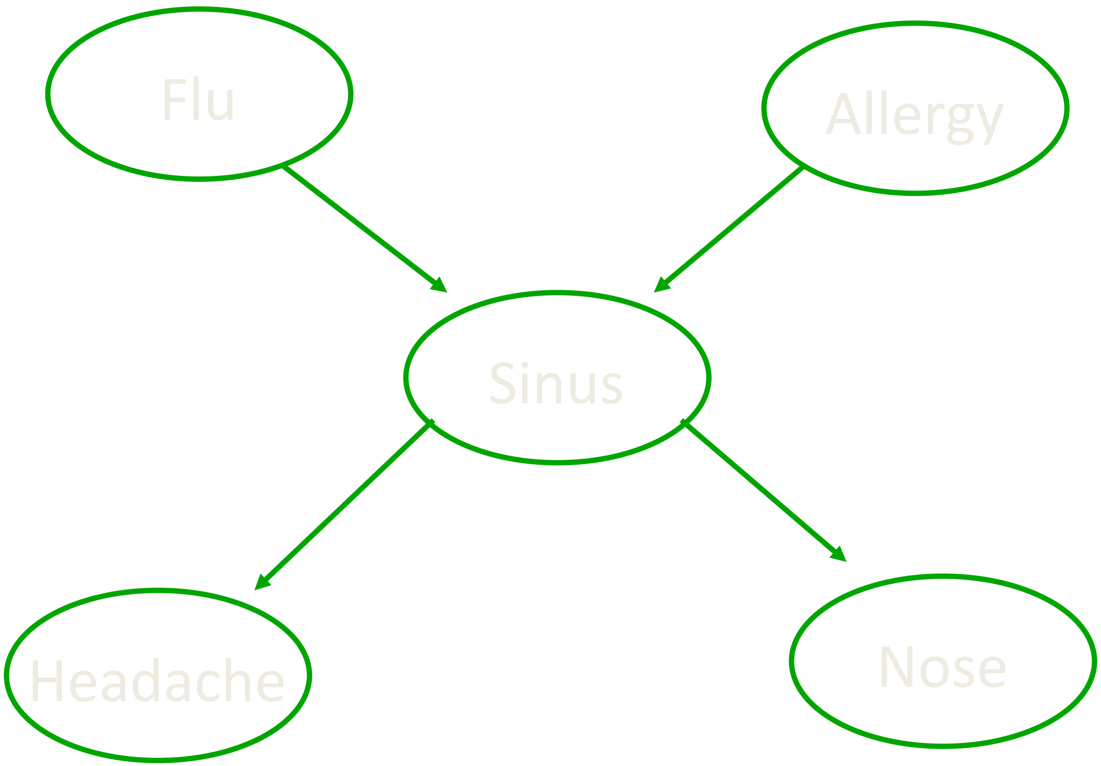
Bayesian networks

- 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

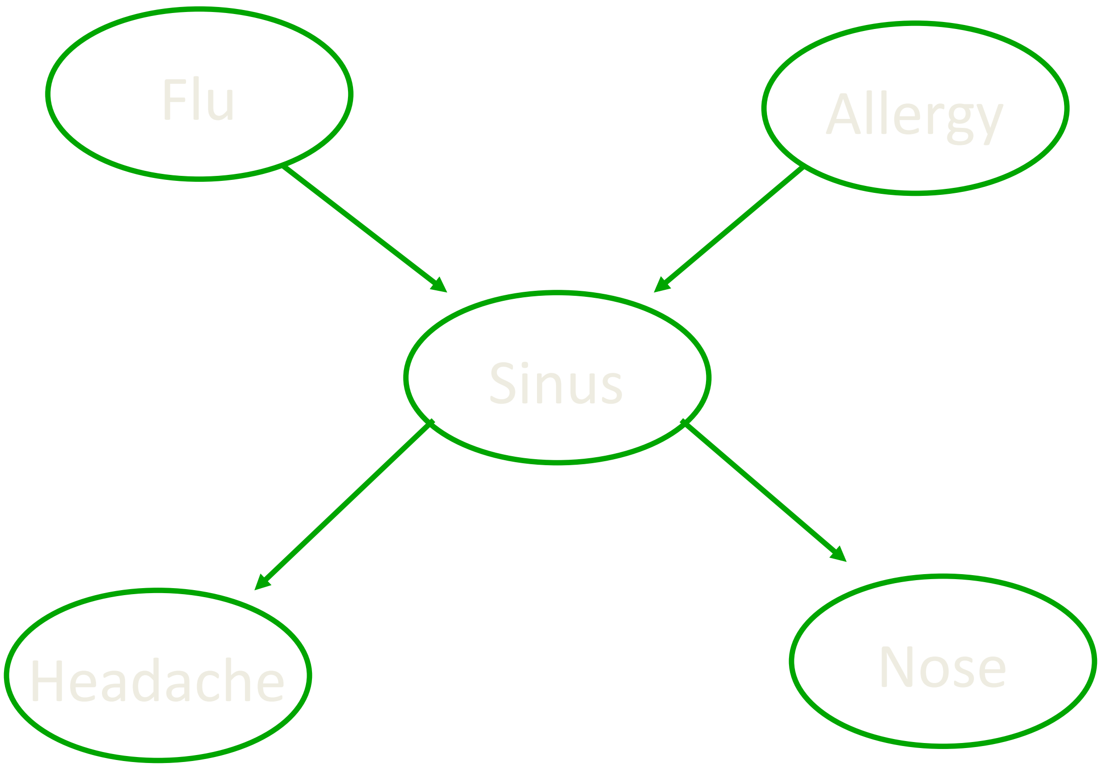
- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

Possible queries

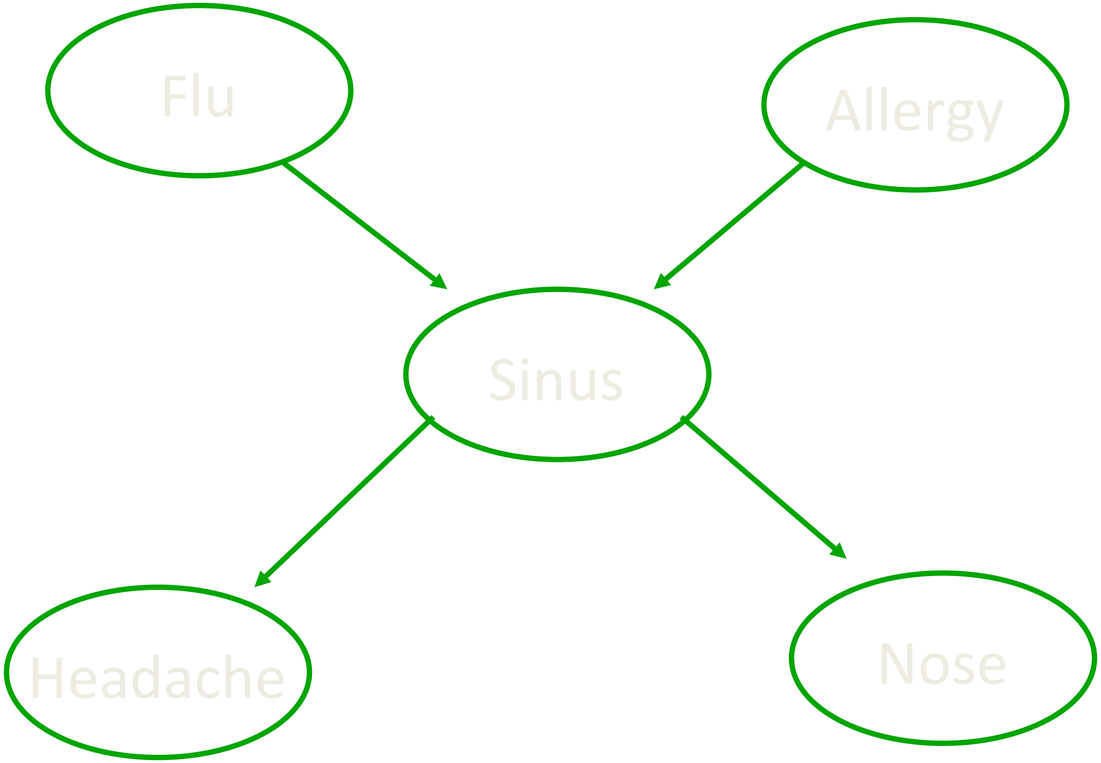


- Inference
- Most probable explanation
- Active data collection

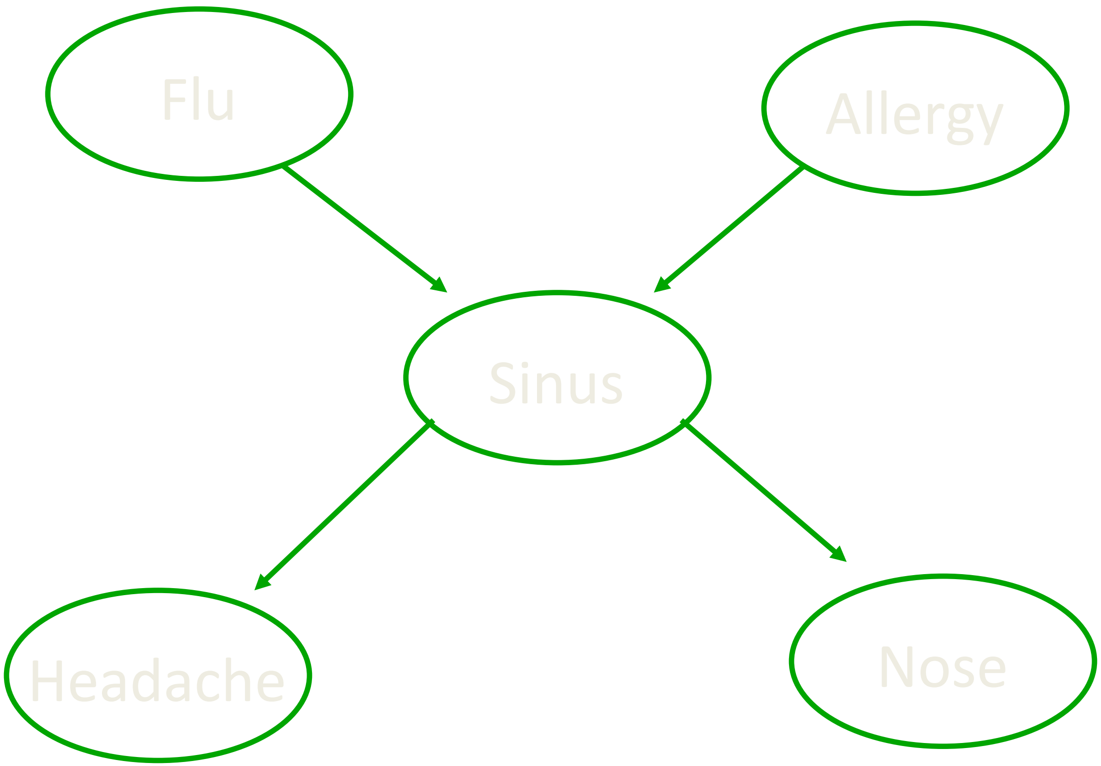
Factored joint distribution - Preview



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	

Allergy = t	
Allergy = f	

- More Generally:

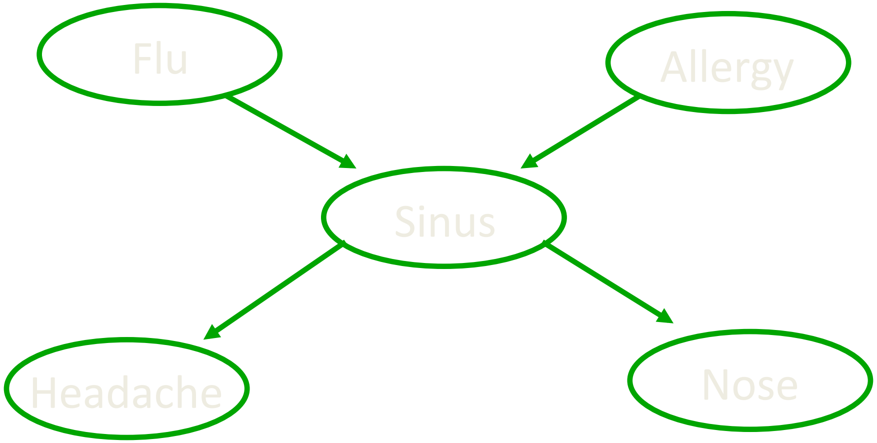
	Flu = t	Flu = f
Allergy = t		
Allergy = f		

Conditional independence

- Flu and Headache are not (marginally) independent

- Flu and Headache are independent given Sinus infection

The independence assumption

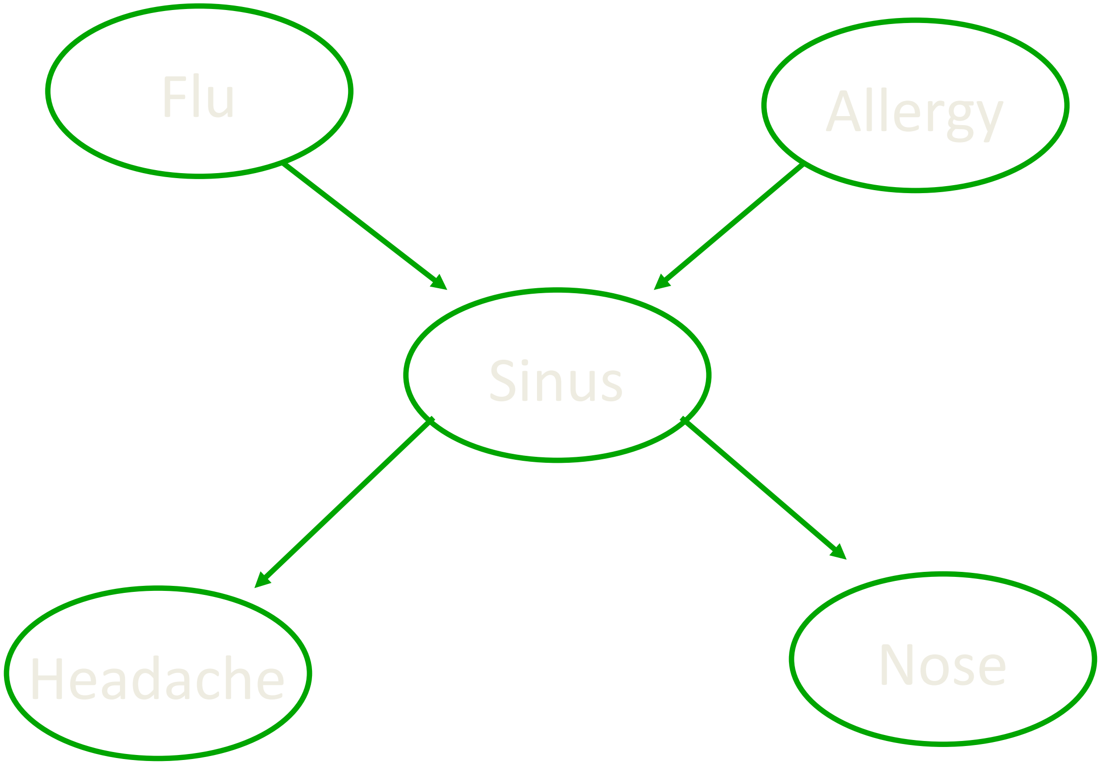


Local Markov Assumption:
A variable X is independent of its non-descendants given its parents and only its parents

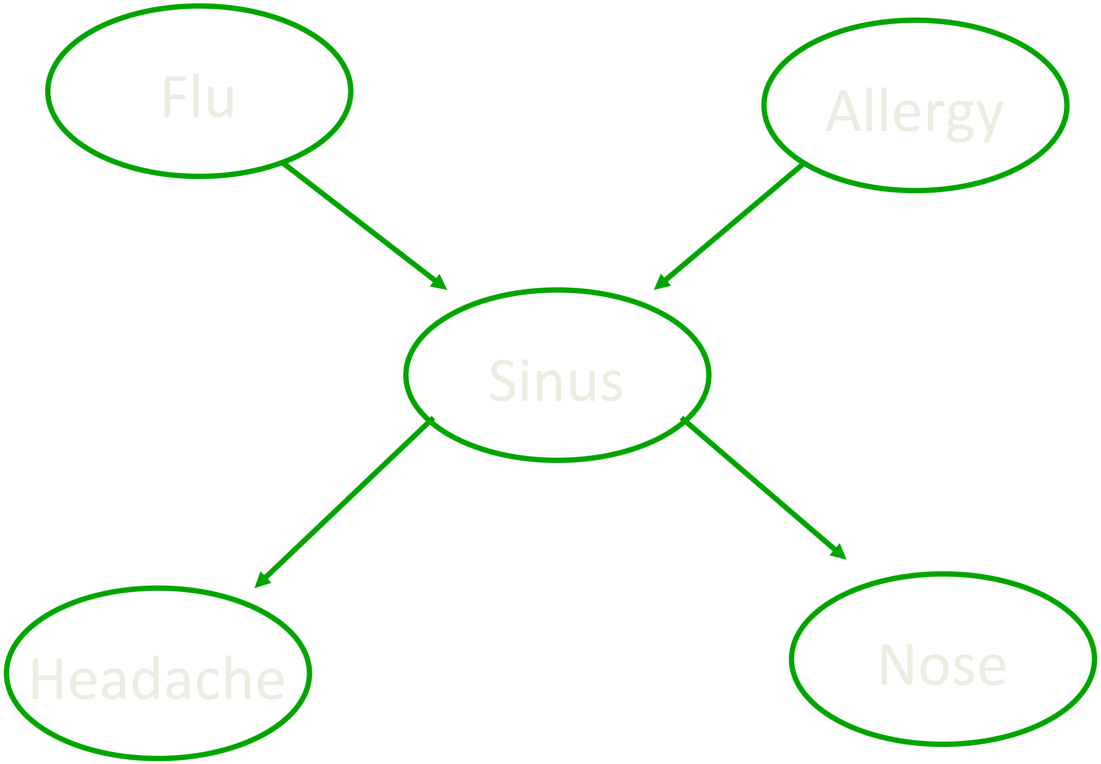
Naïve Bayes revisited

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

What about probabilities? Conditional probability tables (CPTs)

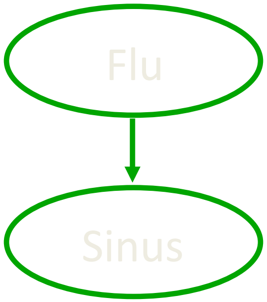


Joint distribution



Why can we decompose? Markov Assumption!

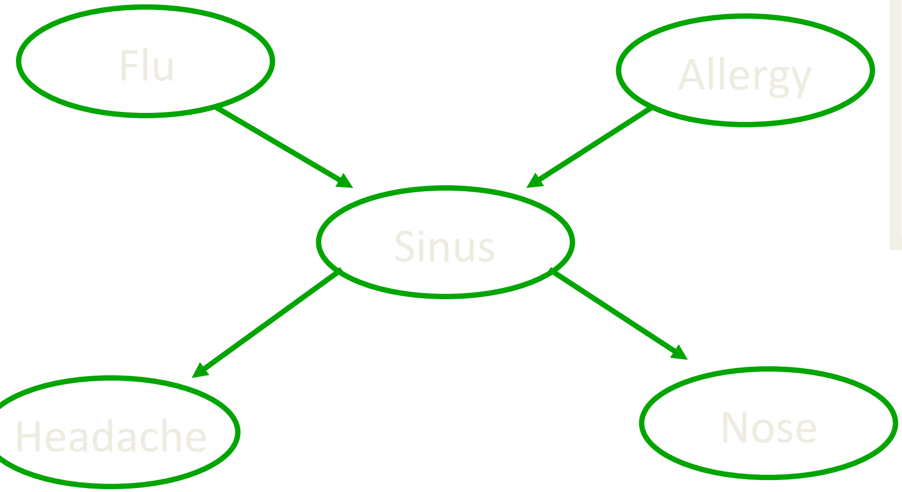
The chain rule of probabilities



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

- 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 and only its parents

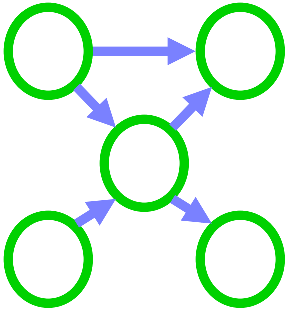
Two (trivial) special cases

Edgeless graph

Fully-connected graph

The Representation Theorem – Joint Distribution to BN

BN:



Encodes independence assumptions

If conditional independencies in BN are a 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 diagnosis
- 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 – $P(X_i| \mathbf{Pa}_{X_i})$

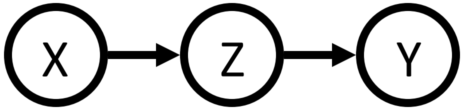
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp NonDescendants_{Xi} \mid \mathbf{Pa}_{Xi})$
- 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!!!

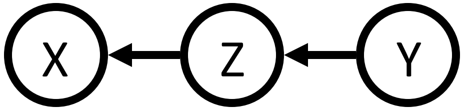
Understanding independencies in BNs – BNs with 3 nodes

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

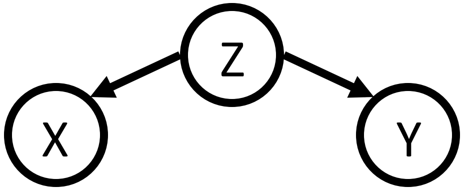
Indirect causal effect:



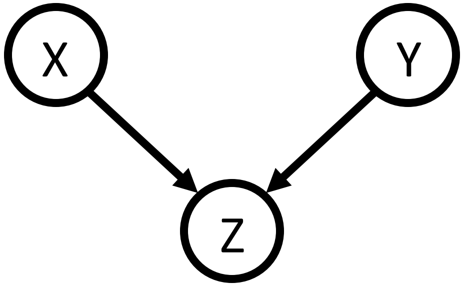
Indirect evidential effect:



Common cause:



Common effect:



Hidden Markov Models

Adventures of our BN hero

- Compact representation for probability distributions
- Fast inference
- Fast learning

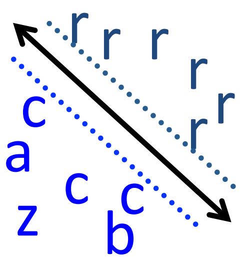
- But… Who are the most popular kids?

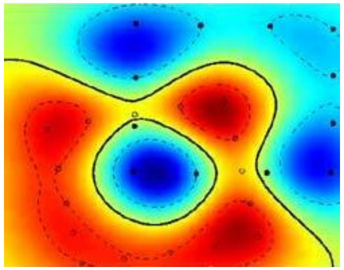
1. Naïve Bayes

2 and 3.
Hidden Markov models (HMMs)
Kalman Filters

Handwriting recognition

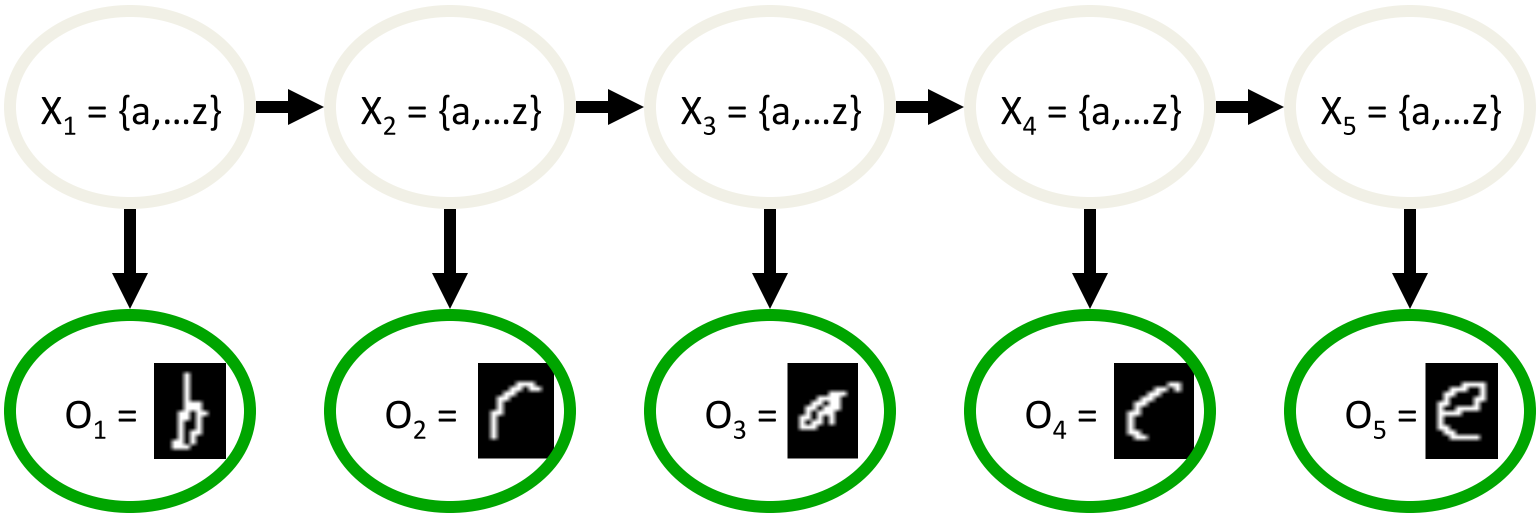
Character recognition, e.g., kernel SVMs



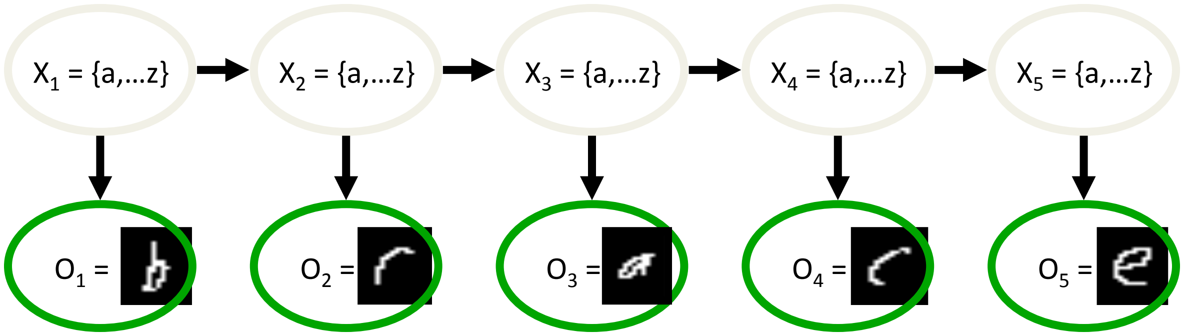


Example of a hidden Markov model (HMM)

Understanding the HMM Semantics



HMMs semantics: Details



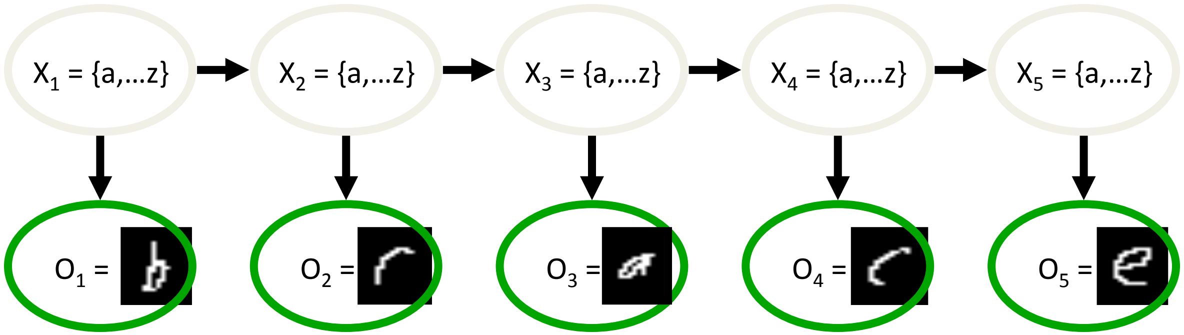
Just 3 distributions:

$$P(X_1)$$

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

$$P(O_i \mid X_i)$$

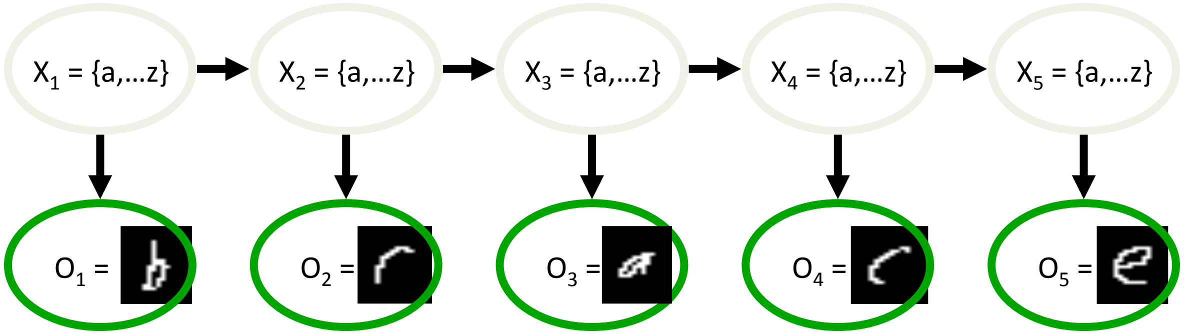
HMMs semantics: Joint distribution



$$P(X_1)$$
$$P(X_i \mid X_{i-1})$$
$$P(O_i \mid X_i)$$

$$P(X_1, \ldots, X_n \mid o_1, \ldots, o_n) = P(X_{1:n} \mid o_{1:n})$$
$$\propto P(X_1)P(o_1 \mid X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1})P(o_i \mid X_i)$$

Learning HMMs from fully observable data is easy



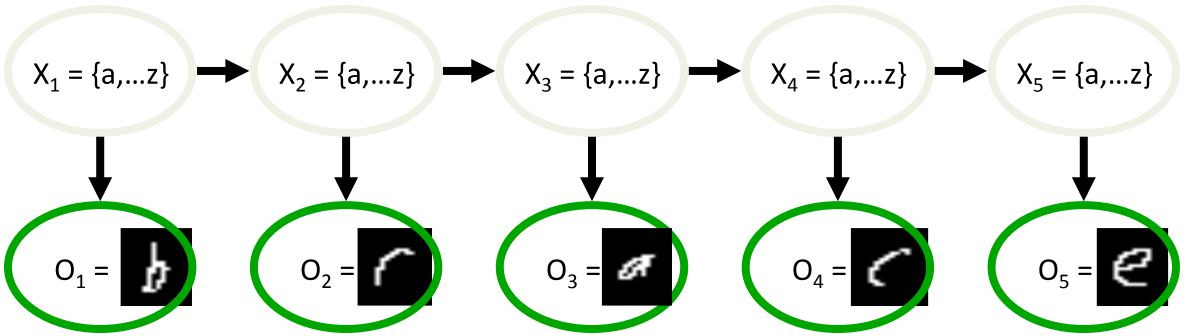
Learn 3 distributions:

$$P(X_1)$$

$$P(O_i \mid X_i)$$

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

Possible inference tasks in an HMM



Marginal probability of a hidden variable:

Viterbi decoding – most likely trajectory for hidden vars: