



"....so then we wondered how long  
it would take a million Shakespeares  
to type xlfwkljdryawehljqNuy."

## Graphical Models

Mark Dredze

Machine Learning  
CS 600.475

Based on intro by Kevin Murphy

## Probabilistic Models

- We have considered many probabilistic models
  - Logistic regression
  - Naïve Bayes
  - Linear Regression
  - Gaussian Mixture Models
- Most of these have been very simple
  - Assume a label (observed or unobserved)
  - Estimate probabilities from data

## Model Representations

- No formal language to talk about model
  - We've described the models and given intuition
- Example: Naïve Bayes
  - Assume that we first generate a label
  - We then generate features given the label
- How can we describe this model formally?

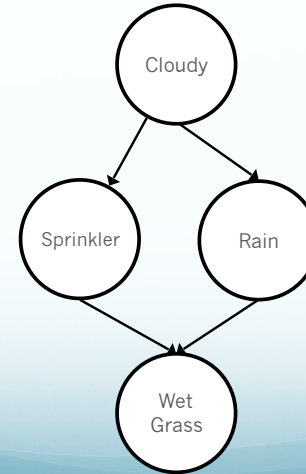
## Example Probabilistic System

- A collection of related binary random variables
  - The weather is cloudy
  - The sprinkler is turned on
  - It is raining
  - The grass is wet
- We can ask questions
  - If it is raining, what is the probability the grass is wet?
  - What is the probability that the grass is wet and its not cloudy?
  - Etc

## Example

- How do we answer these questions?
  - What is the structure of these variables?
  - What probabilities do I need to compute?
  - Are any of the variables independent of each other?
- We need some representation for these variables

## Graphical Models



## Outline

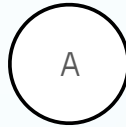
- Representation
  - **What is a graphical model?**
  - **What does it represent**
  - Conditional Independence
  - Types of probabilistic models
- Inference
  - How can we compute probabilities?
  - Message Passing
- Examples
  - Learning and inference

## Graphical Models

- Combination of probability theory and graph theory
  - Combines uncertainty (probability) and complexity (graphs)
  - Represent a complex system as a graph
    - Gives modularity
  - Standard algorithms for solving graph problems
- Your favorite algorithms are graphical models
  - Logistic regression, naïve Bayes, GMMs, etc.

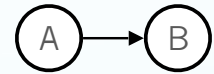
## Representation

- A probabilistic system is encoded as a graph
- Nodes
  - Random variables
    - Could be discrete (this lecture) or continuous
- Edges
  - Connections between two nodes
  - Indicates a direct relationship between two random variables
  - Note: the lack of an edge is very important
    - No direct relationship

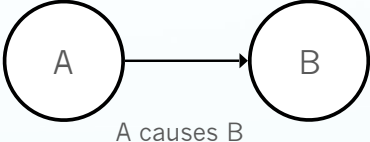


## Graph Types

- Edge type determines graph type
- Directed graphs
  - Edges have directions ( $A \rightarrow B$ )
  - Assume DAGs (no cycles)
  - Typically called Bayesian Networks
    - Popular in AI and stats
- Undirected graphs
  - Edges don't have directions ( $A - B$ )
  - Typically called Markov Random Fields (MRFs)
    - Popular in physics and vision

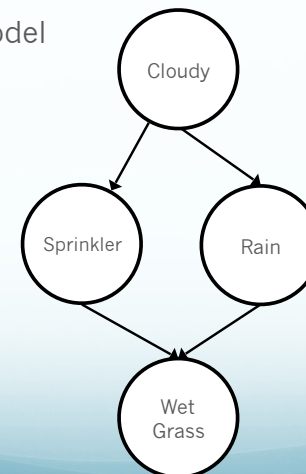


## Directed Graphs

- The direction of the edge indicates causation
- 
- A diagram of a directed graph with two circular nodes, 'A' and 'B'. A directed edge points from node 'A' to node 'B'. Below the edge, the text 'A causes B' is written.
- Causation can be very intuitive
    - We may know which random variable causes the other
    - Use this intuition to create a graph structure

## Example

Generative Model

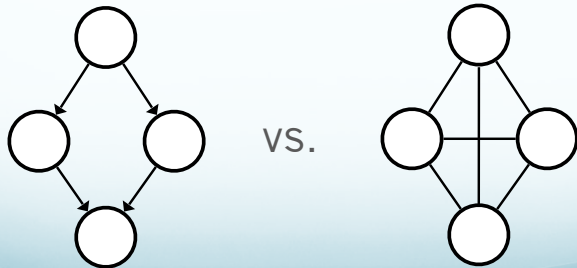


The Generative Story



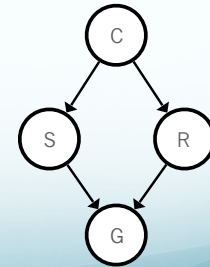
## Advantages?

- What have we gained by this representation?
  - We could just draw a graph where everything is connected



## Factorization

- Consider the joint probability of our example
  - $p(C,S,R,G)$ - this is complex
  - What can we do to simplify?
  - Notice that S and R are independent given C



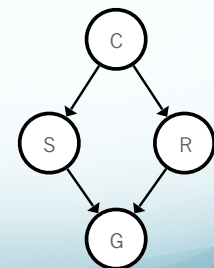
## Product Rule

- Can use the product rule to decompose joint probabilities
  - $p(a,b,c) = p(c|a,b) p(a,b)$
  - $p(a,b,c) = p(c|a,b) p(b|a) p(a)$
- This is true for any distribution
- Same for K variables
 
$$p(x_1 \dots x_K) = p(x_K | x_1 \dots x_{K-1}) \dots p(x_2 | x_1) p(x_1)$$

## Factorization

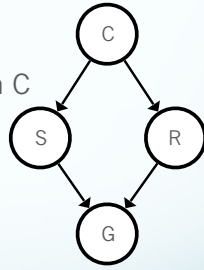
- For any graphical model we can write the joint distribution using conditional probabilities
  - We just need conditional probabilities for a node given its parents

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{parents}_k)$$



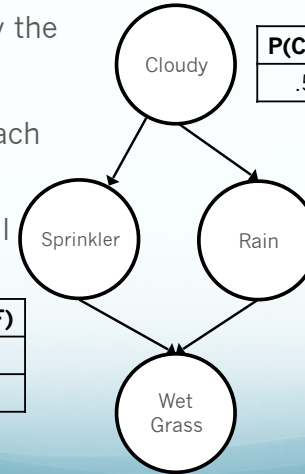
## Factorization

- Consider the joint probability of our example
  - $p(C,S,R,G)$ - this is complex
  - What can we do to simplify?
  - Notice that S and R are independent given C
- Factor the joint probability according to the graph
  - $p(C,S,R,G) = p(G|S,R) p(S|C) p(R|C) p(C)$
  - This is much simpler to compute
  - We are likely to have these conditional probabilities



## Conditional Probability Tables

- The CPTs specify the conditional probability distribution at each node
- CPTs reflect local information only



P(C=T)	P(C=F)
.5	.5

C	P(R=T)	P(R=F)
F	.2	.8
T	.8	.2

C	P(S=T)	P(S=F)
F	.5	.5
T	.1	.9

S	R	P(G=T)	P(G=F)
F	F	0	1
T	F	.9	.1
F	T	.9	.1
T	T	.99	.01

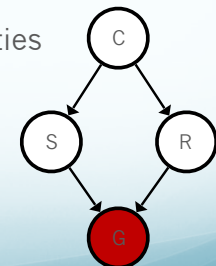
## Conditional Probability Tables

- Graph provides a problem structure that indicates relationships
- We use this structure to break down the problem into many local problems
- What is  $P(S=T|G=T)$ ?
  - Break down using the network and CPTs

$$p(S=T|G=T) = \frac{p(S=T, G=T)}{p(G=T)} = \frac{\sum_{c,r} p(C=c, S=T, R=r, G=T)}{\sum_{c,r,s} p(C=c, S=s, R=r, G=T)} = 0.430$$

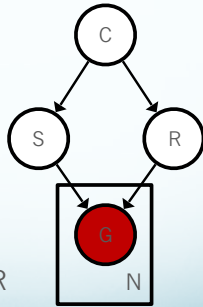
## Observed Variables

- Variables are either
  - Observed- we observe values in data
  - Hidden- we cannot see values in data
- Indicate observed variables by shading
- Compute the remaining probabilities given shaded value



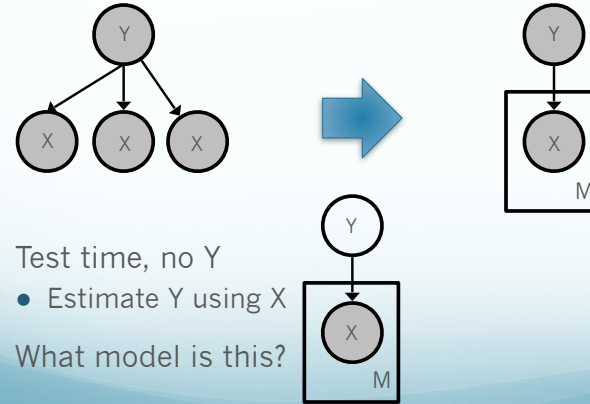
## Plate Notation

- Plates in graphical models
  - When many variables have same structure, we replace them with a plate
  - The plate indicates repetition
- There are N fields in which we can see if the grass is wet
- Each conditioned on the same S and R



## Example

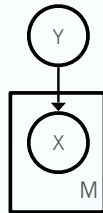
- A model where we have label Y and example X



- Test time, no Y
  - Estimate Y using X
- What model is this?

## Naïve Bayes

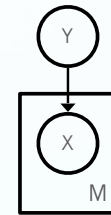
- Generative Story
  - Generate a label Y
  - Given Y, generate each feature X independently
- Learning
  - We observe X and Y, maximum likelihood solution
- Prediction
  - Compute most likely value for Y given X



## Factorization

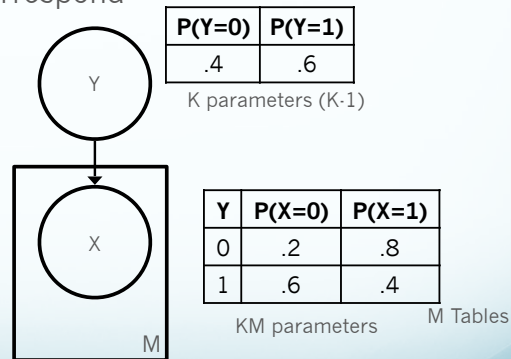
$$P(y, x) = P(x | y)P(y)$$

$$= \prod_{j=1}^M P(x_j | y)P(y)$$



## Conditional Probability Tables

- The parameters correspond to CPTs



## Learning

- We assumed both examples (X) and labels (Y) for learning naïve Bayes

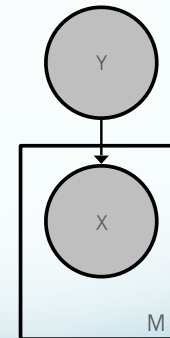
- Maximum likelihood solution
  - Each entry in table are based on counts

- What if we only have X?

- General purpose method for maximizing likelihood where we have missing variables

- $\max P(X) = \sum_{y \in Y} P(Y, X)$

- EM
- Unsupervised NB: clustering
- Some labels: semi-supervised NB



## Assumptions

- When we previously derived Naïve Bayes, we made an assumption
  - Features (X) conditionally independent given label (Y)
- How does independence fit in graphical models?

## Independence

- The best part of graphical models is what they do not show

- Consider the network



- A and B are independent
  - $P(A, B) = P(A) P(B)$
  - Variable independence allows us to build efficient models
    - Recall discussion on Naïve Bayes



# Conditional Independence

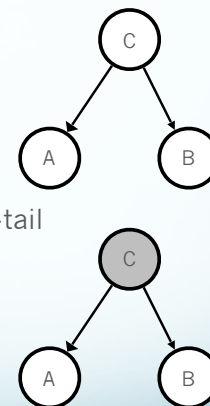
- Are A and B independent?



- A and B are conditionally independent given C
  - $P(A,B|C) = P(A|C) P(B|C)$
  - Once we know the value of C, no amount of information about B will change A
- How do we know if something is independent?
  - It's encoded in the paths of the graph!
  - No mathematical trickery needed

# Example 1

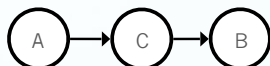
- Are A and B independent?
  - Clearly not. Both depend on C
- Are A and B conditionally independent?
  - Yes. Why?
  - The connection of A and B to C is tail-to-tail
    - Creates a dependence
  - When we condition on C, it blocks the path between A and B



# Example 2

- Are A and B independent?

- No. A cause C which causes B



- Are A and B conditionally independent?

- Yes. Why?

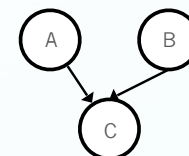


- The connection of A and B to C is head to tail
  - Creates a dependence
- When we condition on C, it blocks the path between A and B

# Example 3

- Are A and B independent?

- Yes. A and B are generated without common parents



- Are A and B conditionally independent given C?

- No. Why?
- The connection of A and B to C is head to head
  - Creates a dependence
- When C is unobserved, the path is **blocked**
- When C is observed, the path becomes **unblocked**

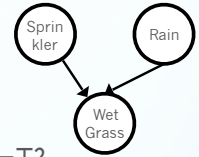


## Blocked vs. Unblocked?

- Terminology:  $y$  is a descendent of  $x$  if there is a path from  $x$  to  $y$  (following the arrows)
- Tail to tail or head to tail node only blocks a path when it is **observed**
- A head to head node blocks a path when it is **unobserved**
  - A head to head path will become unblocked if either node, or any of its descendants, is observed

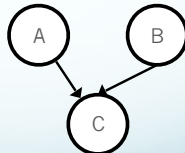
## Why?

- Recall the sprinkler/rain example
- The two causes (sprinkler/rain) compete to explain the grass
- Suppose  $G=T$ , what is the probability of  $S=T$ ?
  - $P(S=T|G=T) = .430$  (from before)
- Suppose we learn that  $R=T$ . What is  $S=T$  now?
  - $P(S=T|R=T, G=T) = 0.1945$



## Explaining Away

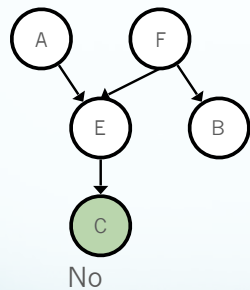
- This makes sense
  - The rain explained the grass, so sprinkler is now less likely
  - The rain explained away the state of the grass
  - Less need to use sprinkler to explain it
- This is why the observed head to head is unblocked
  - Once we know the value, we learn something about A and B



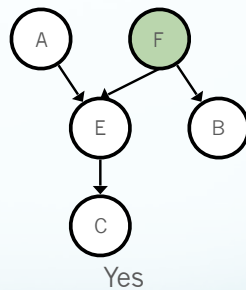
## D-Separation

- Two sets of nodes A and B are **d-separated** given observed set C if all paths between A and B are blocked
  - Blocked paths
    - The arrows on the path meet head to tail or tail to tail at a node in set C
      - OR
    - The arrows meet head to head at a node and neither the node, nor any of its descendants, is in set C
- If sets of nodes are d-separated they are conditionally independent

## D-Separation Examples

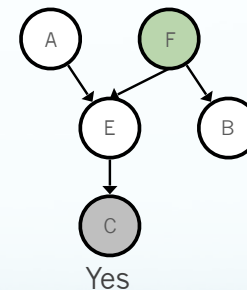


C is a descendent of head to head E



F is a tail to tail node

## D-Separation Example



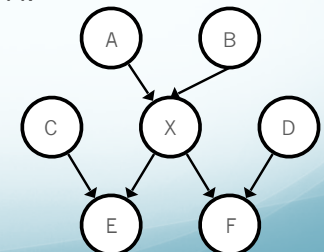
F is a tail to tail node and block path even though E is unblocked

## Isolating Nodes

- How do we isolate a variable in the graph?
  - We know how to make it conditionally independent
  - We want to experiment with a variable in isolation
  - We don't want to enumerate all possible values of the whole network

## Markov Blanket

- The Markov blanket of a node is the minimal set of nodes that isolates it from the graph
  - A node conditioned on its Markov blanket is independent from all other nodes in the graph
- What nodes are in the blanket for X?
  - Think about d-separation
  - All of them!
  - A Markov blanket depends on the parents, children, and co-parents



## Next Time

Inference in Graphical Models