Graphical Models

Mark Dredze

Machine Learning
CS 601.475

Based on intro by Kevin Murphy

Probabilistic Models

- We have considered many probabilistic models
  - Logistic regression
  - 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: Gaussian Mixture Models
  - Assume that we first select a cluster
  - We then generate an example (features) given the cluster
- How can we describe this model formally?

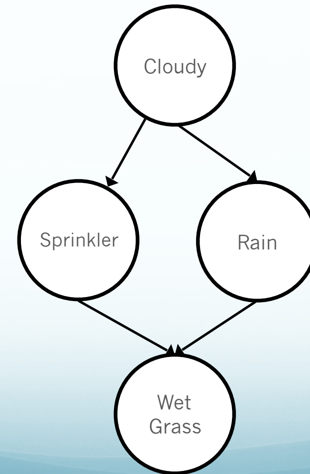
Example Probabilistic System

- 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

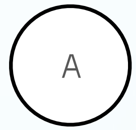
- Representation
  - **What is a graphical model?**
  - **What does it represent**
  - Conditional Independence
  - Types of probabilistic models
- Inference
  - How can we compute probabilities?
  - Message Passing
- Examples
  - Learning and inference

Graphical Models

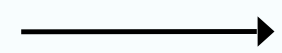
- Combination of probability theory and graph theory
  - Combines uncertainty (probability) and complexity (graphs)
  - Represent a complex system as a graph
    - Gives modularity
  - Standard algorithms for solving graph problems
- Your favorite algorithms are graphical models
  - Logistic regression,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

Representation

- A probabilistic system is encoded as a graph
- Nodes
  - Random variables
    - Could be discrete (this lecture) or continuous

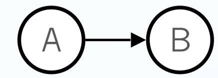


- Edges
  - Connections between two nodes
  - Indicates a direct relationship between two random variables
  - Note: the lack of an edge is very important
    - No direct relationship

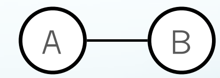


Graph Types

- Edge type determines graph type
- Directed graphs
  - Edges have directions (A -> B)
  - Assume DAGs (no cycles)
  - Typically called Bayesian Networks
    - Popular in AI and stats

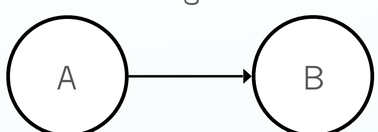


- Undirected graphs
  - Edges don't have directions (A – B)
  - Typically called Markov Random Fields (MRFs)
    - Popular in physics and vision



Directed Graphs

- The direction of the edge indicates causation

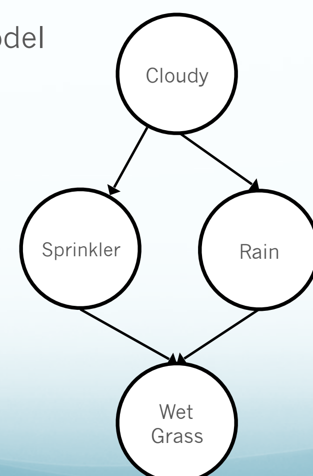


A causes B

- Causation can be very intuitive
  - We may know which random variable causes the other
  - Use this intuition to create a graph structure

Example

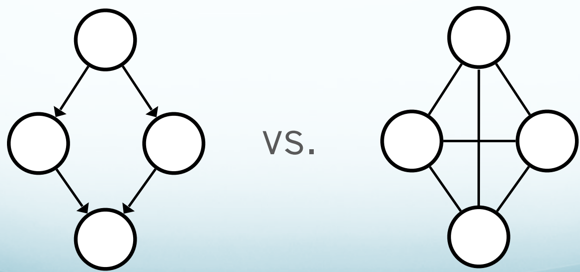
Generative Model



The Generative Story

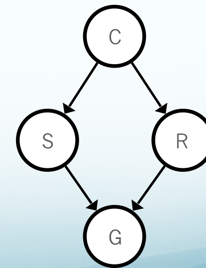
Advantages?

- 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

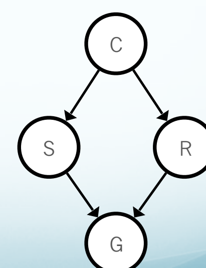
- Can use the product rule to decompose joint probabilities
  - p(a,b,c) = p(c|a,b) p(a,b)
  - p(a,b,c) = p(c|a,b) p(b|a) p(a)
- This is true for any distribution
- Same for K variables

$$p(x_1 \ldots x_K) = p(x_K \mid x_1 \ldots x_{K-1}) \ldots p(x_2 \mid x_1)\, p(x_1)$$

Factorization

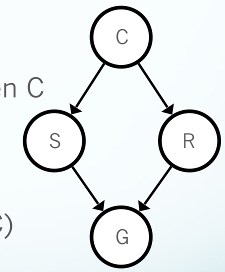
- For any graphical model we can write the joint distribution using conditional probabilities
  - We just need conditional probabilities for a node given its parents

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



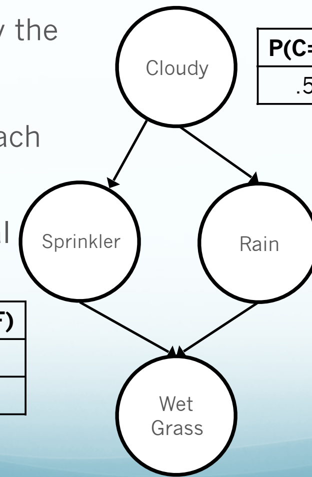
Factorization

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

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

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

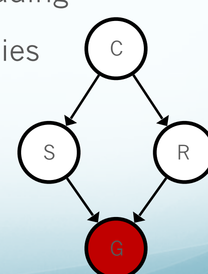
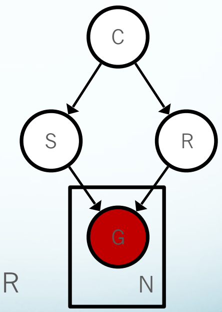


Plate Notation

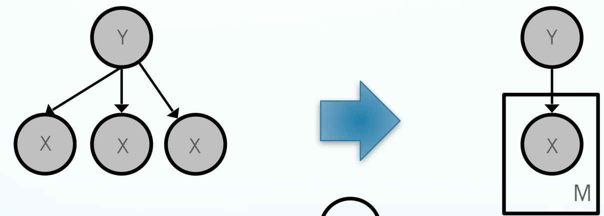
- Plates in graphical models
  - When many variables have same structure, we replace them with a plate
  - The plate indicates repetition



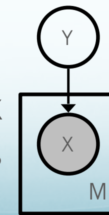
  - There are N fields in which we can see if the grass is wet
  - Each conditioned on the same S and R

Example

- A model where we have label Y and example X

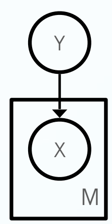


- Test time, no Y
  - Estimate Y using X
- What model is this?



Naïve Bayes

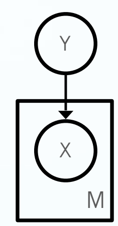
- Generative Story
  - Generate a label Y
  - Given Y, generate each feature X independently
- Learning
  - We observe X and Y, maximum likelihood solution
- Prediction
  - Compute most likely value for Y given X



Factorization

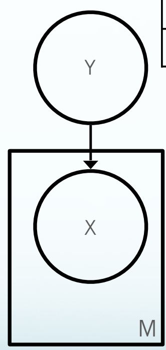
$$P(y,x) = P(x \mid y)P(y)$$

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



Conditional Probability Tables

- The parameters correspond to CPTs



P(Y=0)	P(Y=1)
.4	.6

K parameters (K-1)

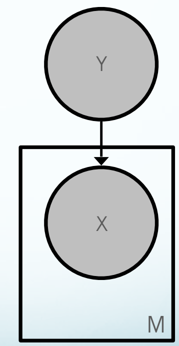
Y	P(X=0)	P(X=1)
0	.2	.8
1	.6	.4

KM parameters

M Tables

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



Conditional

- What is p(x|y)?
  - Probability of generating example x given that it has label y
- How hard is this?
  - Remember that x is a vector
  - Equivalent to $p(x_{i1}, x_{i2}, x_{i3} \ldots x_{iM} \mid y_i)$
  - Assuming binary features and binary label, how many parameters do we need?
    - 2 * (2^M-1) parameters!
      - (2^M-1) combinations for x
      - 2 labels

Conditional Independence

- **RV (random variable) X is conditionally independent of RV Y given RV Z if the probability of each is independent given Z**
- p(x,y|z) = p(x|z)p(y|z)
- Example
  - Probability that I need an umbrella and the ground is wet
  - Not independent! If its wet I probably need an umbrella because it is raining
  - I am told it is raining
  - Given this the probability that I need an umbrella is independent of the ground being wet
  - I gain no new information knowing that the ground is wet

Conditional Independence

- Assume each feature in x is independent given y
  - Once I know y each feature in x is independent
- Why is this helpful?

$$p(x_i \mid y_i) = \prod_{j=1}^{M} p(x_{ij} \mid y_i)$$

- This is a naïve assumption (it's very unlikely)

Conditional Independence

- How to estimate $p(x_{ij} \mid y_i)$?
  - Lots of data- every time feature x_{ij} occurs with y_i
- How many parameters do I need?
  - Before: 2 * (2^M-1)
  - Now: 2 * M
    - One parameter for each of M features
- Should be easier to learn so many fewer parameters

Naïve vs. Reality

- Positive: we now can parameterize our model
- Reality: naïve assumption very unlikely to be true
- Example:
  - Document classification: sports vs. finance
  - Each word in a document is a feature
  - Naïve assumption: once I know the topic is sports, every word is conditionally independent
    - Not true! Would be total nonsense.

Naïve vs. Reality

- Reality: works pretty well in practice
- Caution: features that are too dependent are difficult for model
  - Create features that are minimally dependent
  - Limits the expressiveness of features

Assumptions

- Naïve Bayes makes an assumption
  - Features (X) conditionally independent given label (Y)
- How does independence fit in graphical models?

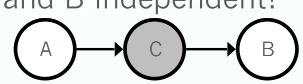
Independence

- 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

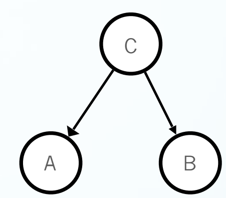
Conditional Independence

- Are A and B independent?

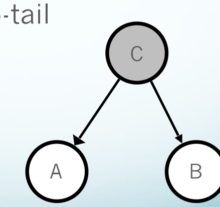


  - A and B are conditionally independent given C
    - P(A,B|C) = P(A|C) P(B|C)
    - Once we know the value of C, no amount of information about B will change A
- How do we know if something is independent?
  - It's encoded in the paths of the graph!
  - No mathematical trickery needed

Example 1

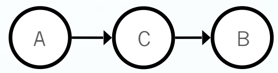


- Are A and B independent?
  - Clearly not. Both depend on C
- Are A and B conditionally independent?
  - Yes. Why?
  - The connection of A and B to C is tail-to-tail
    - Creates a dependence
  - When we condition on C, it blocks the path between A and B

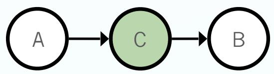


Example 2

- Are A and B independent?
  - No. A cause C which causes B



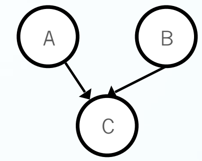
- Are A and B conditionally independent?
  - Yes. Why?



  - 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



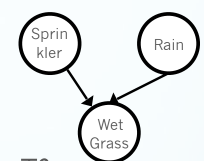
- Are A and B conditionally independent given C?
  - No. Why?
  - The connection of A and B to C is head to head
    - Creates a dependence
  - When C is unobserved, the path is **blocked**
  - When C is observed, the path becomes **unblocked**

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 descendents, is observed

Why?

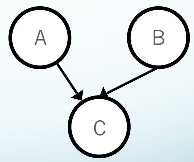
- Recall the sprinkler/rain example



- 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

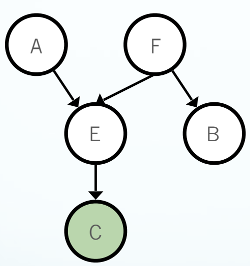
- This makes sense
  - The rain explained the grass, so sprinkler is now less likely
  - The rain explained away the state of the grass
  - Less need to use sprinkler to explain it
- This is why the observed head to head is unblocked
  - Once we know the value, we learn something about A and B



D-Separation

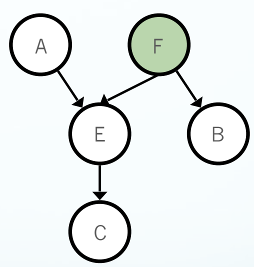
- Two sets of nodes A and B are **d-separated** given observed set C if all paths between A and B are blocked
  - Blocked paths
    - The arrows on the path meet head to tail or tail to tail at a node in set C
      - OR
    - The arrows meet head to head at a node and neither the node, nor any of its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

D-Separation Examples



No

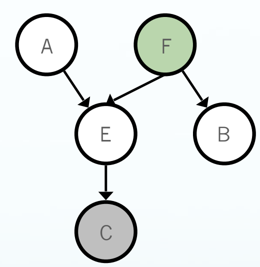
C is a descendent of head to head E



Yes

F is a tail to tail node

D-Separation Example



Yes

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