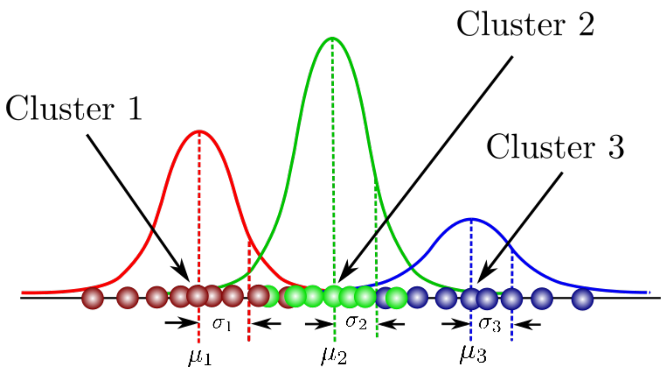
Graphical Models

Zach Wood-Doughty and Bryan Pardo

CS 349 Fall 2021

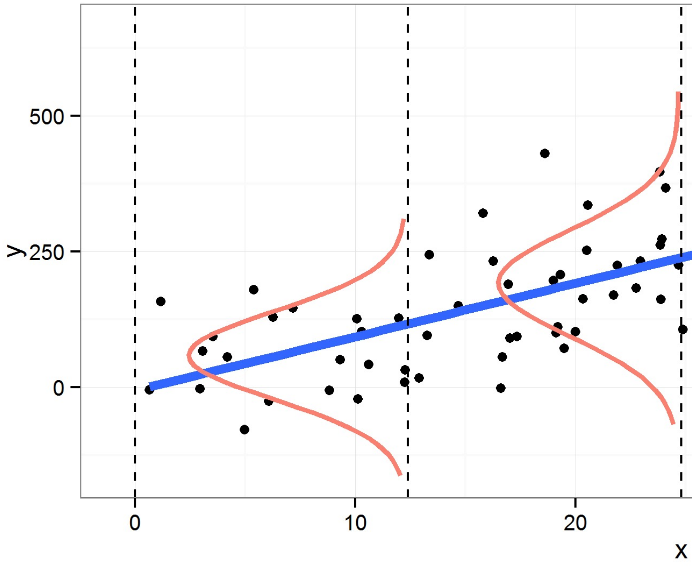
Some slides taken from Mark Dredze
And inspired by Kevin Murphy

Probabilistic Models



- Some models we've considered have a *probabilistic* interpretation
  - Linear Regression
  - Gaussian Mixture Models
- No formal language to talk about models
  - 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?



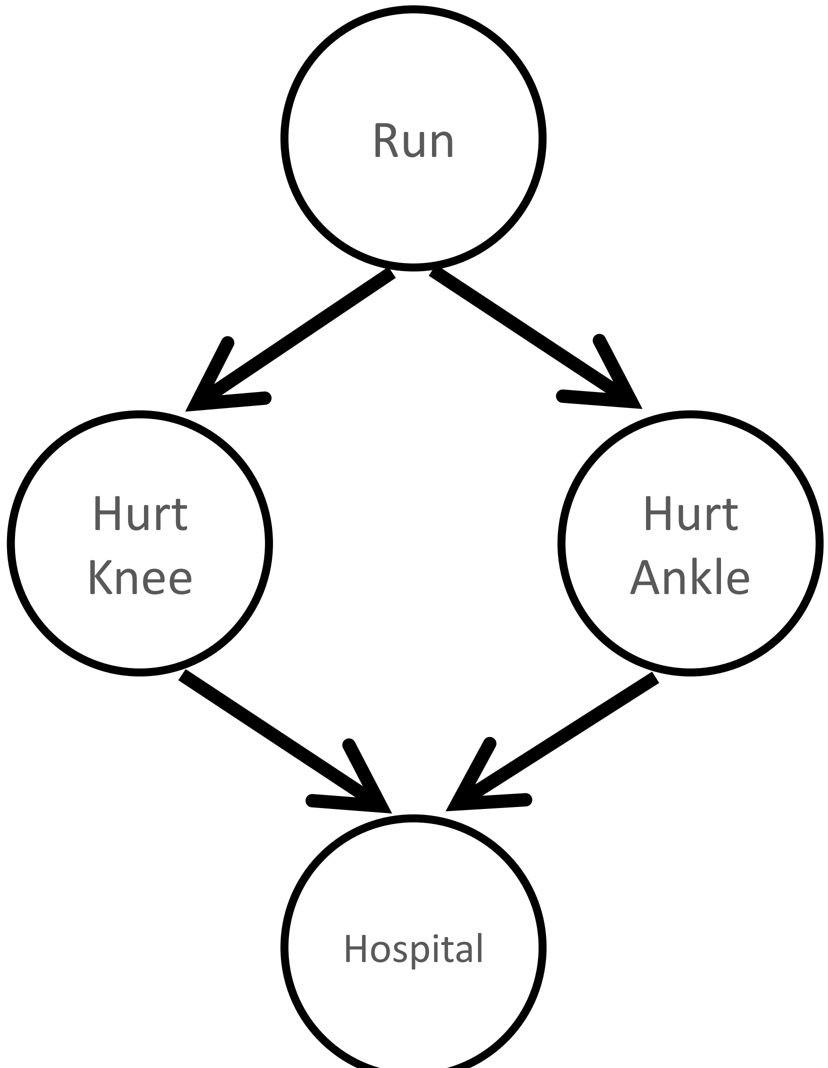
Example Probabilistic System

- A collection of related binary random variables
- Each day with some probability, a runner Avery:
  - Goes for a run
  - Sprains an ankle
  - Injuries their knee
  - Goes to the hospital
- Given a sprained ankle, what's the probability Avery goes to the hospital?
- What is the probability that Avery injuries their knee and goes to the hospital?
- 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?
- How can we represent the variables in a way that answers these questions?

Graphical Models

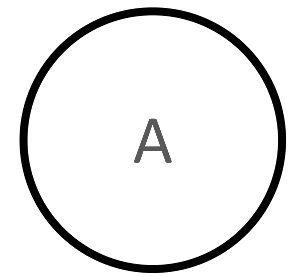


Graphical Models

- 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
- Many ML models can be framed as graphical models
    - Logistic regression, linear Regression, GMMs, etc.

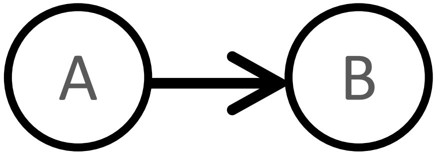
Representation

- 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

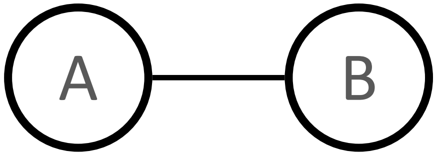


Graph Types

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

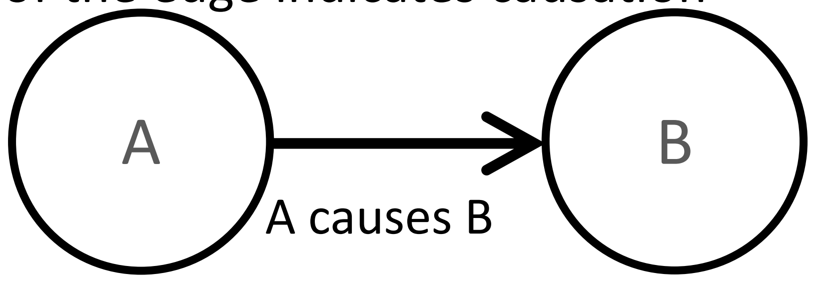


- Undirected graphs
    - Edges don’t have directions (A – B)
    - Typically called Markov Random Fields (MRFs)
        - Popular in physics and vision



Directed Graphs

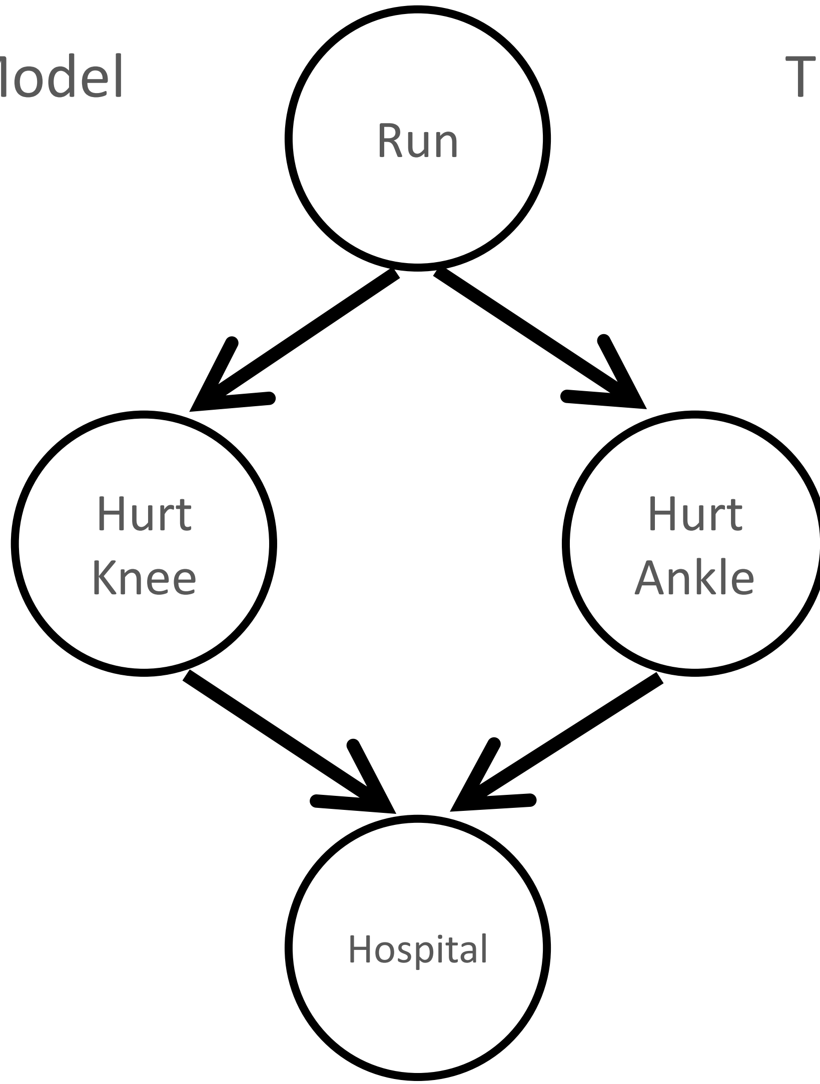
- The direction of the edge indicates causation



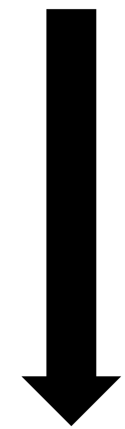
- Causation can be very intuitive
  - We may know which random variable causes the other
  - Use this intuition to create a graph structure

Example

Generative Model

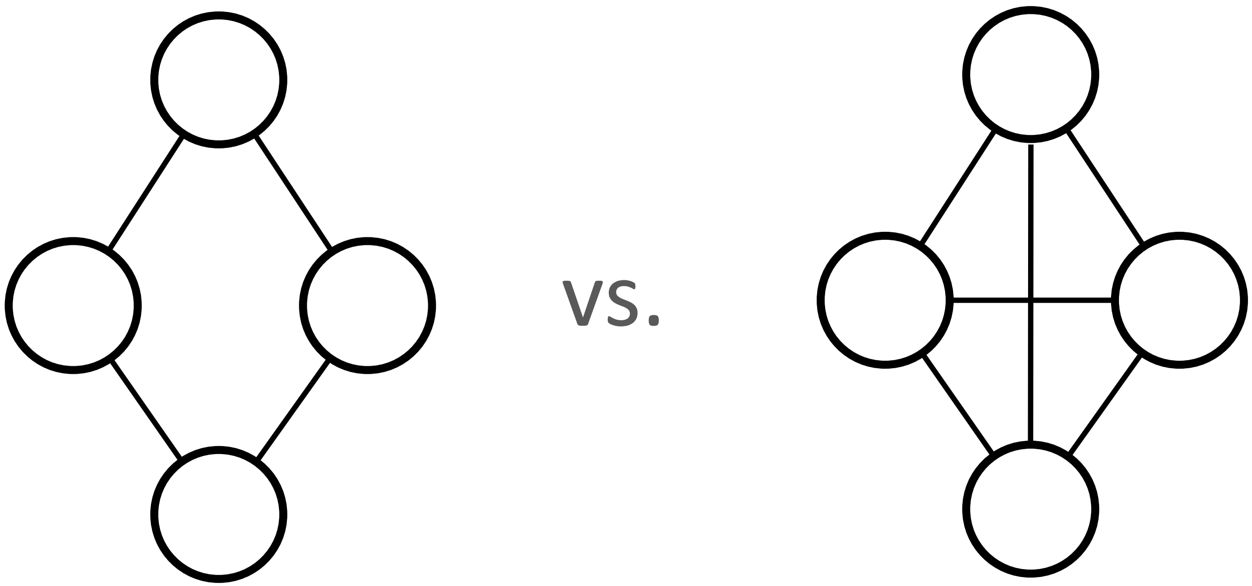


The Generative Story



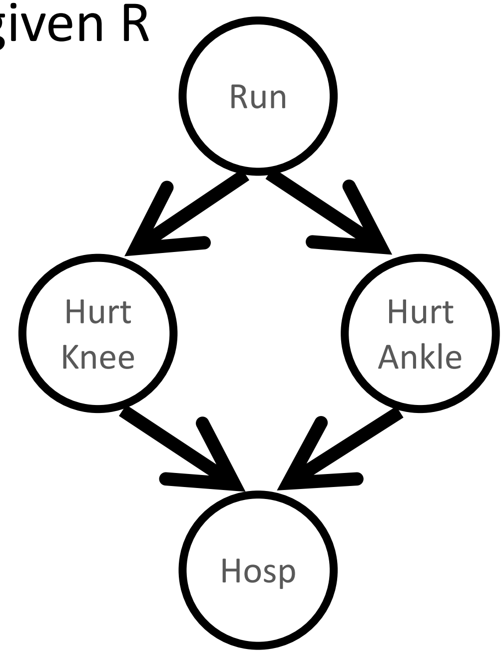
Advantages?

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



Factorization

- Consider the joint probability of our example
  - What is the size of the conditional probability table for the p(R, A, K, H) distribution?
  - What can we do to simplify?
  - Notice that A and K are independent given R



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

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

Recall: independence

- The probability I eat pie today is independent of the probability of a blizzard in Japan.
- This is DOMAIN knowledge, typically supplied by the problem designer
- Independence implies:

$$A \perp B \Rightarrow p(A \mid B) = p(A)$$

$$A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

How does independence help?

$$A \perp B \Rightarrow p(A \mid B) = p(A)$$

A	B	P(A, B)
F	F	0.56
T	F	0.24
F	T	0.14
T	T	0.06

$$\begin{aligned} p(A) &= \sum_B p(A, B) \\ &= p(A, B) + p(A, \neg B) \\ &= 0.24 + 0.06 = 0.3 \end{aligned}$$

$$\begin{aligned} p(A|B) &= \frac{p(A, B)}{p(B)} \\ &= \frac{p(A, B)}{\sum_A p(A, B)} \\ &= \frac{p(A, B)}{p(A, B) + p(\neg A, B)} \\ &= \frac{0.06}{0.06 + 0.14} \\ &= 0.06/0.2 = 0.3 \end{aligned}$$

$$A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

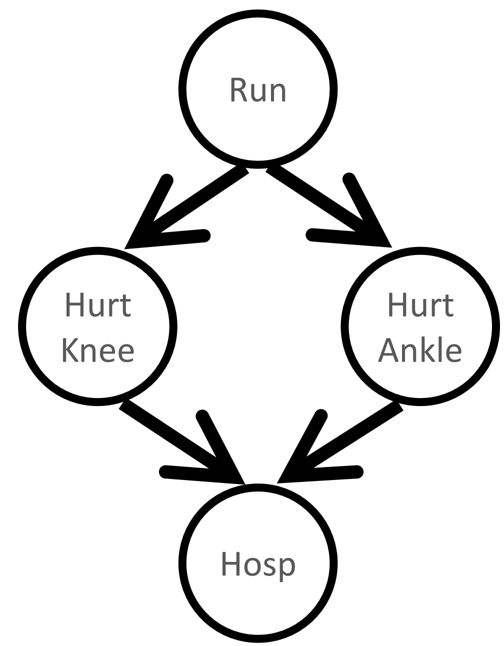
Conditional Independence

- **Random variable X is conditionally independent of Y given Z if the probability of each is independent given Z**
- $p(x,y|z) = p(x|z)p(y|z)$
- $p(x|z, y) = p(x \mid z)$
- Example
  - X: I need an umbrella and Y: the ground is wet
  - Not independent!
  - If ground is wet, it’s probably raining and I’ll need an umbrella
  - I am told it is raining; knowing 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
  - $P(x \mid z, y) = p(x, z)$

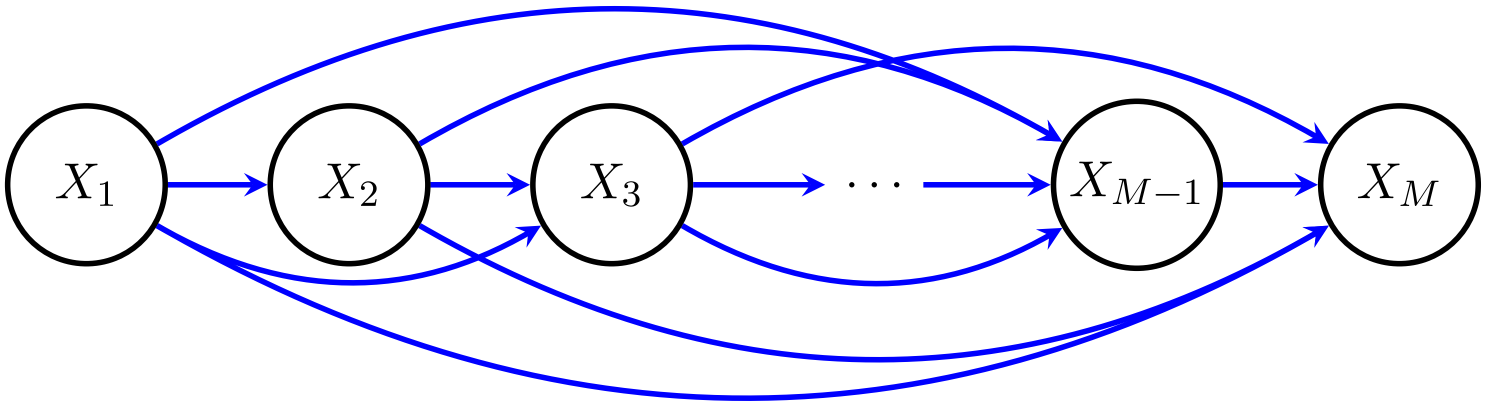
Factorization

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

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



Counting parameters in CPTs



X_1	X_2	…	X_M	P(X)
F	F	F	F	0.001
T	F	F	F	0.014
F	T	F	F	0.004
T	T	F	F	0.002
				…

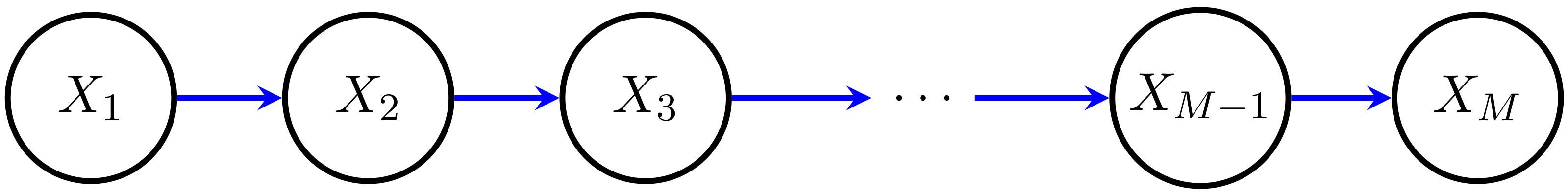
$P(X_1)$
.5

X_1	$P(X_2 \mid X_1)$
F	0.5
T	0.3

X_1	X_2	$P(X_3 \mid X_2, X_1)$
F	F	0.4
T	F	0.3
F	T	0.2
T	T	0.7

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

Counting parameters in CPTs



X_1	X_2	...	X_M	P(X)
F	F	F	F	0.001
T	F	F	F	0.014
F	T	F	F	0.004
T	T	F	F	0.002
				...

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

$P(X_1)$
.5

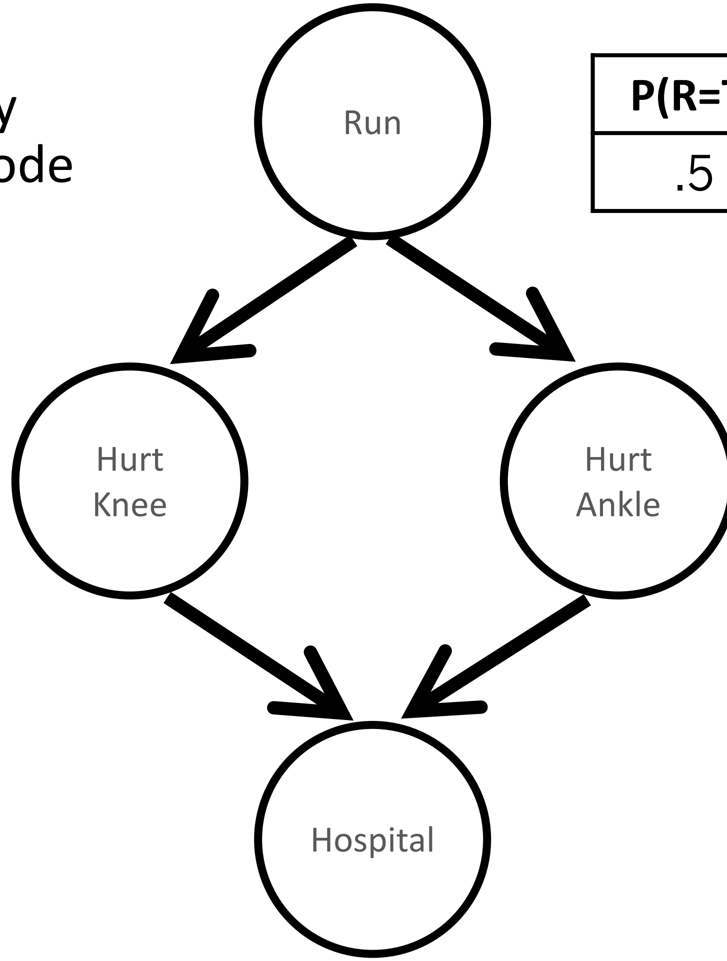
X_1	$P(X_2 \mid X_1)$
F	0.5
T	0.3

X_1	X_2	$P(X_3 \mid X_2, X_1)$
F	F	0.4
T	F	0.4
F	T	0.2
T	T	0.2

X_2	$P(X_3 \mid X_2)$
F	0.4
T	0.2

Conditional Probability Tables

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



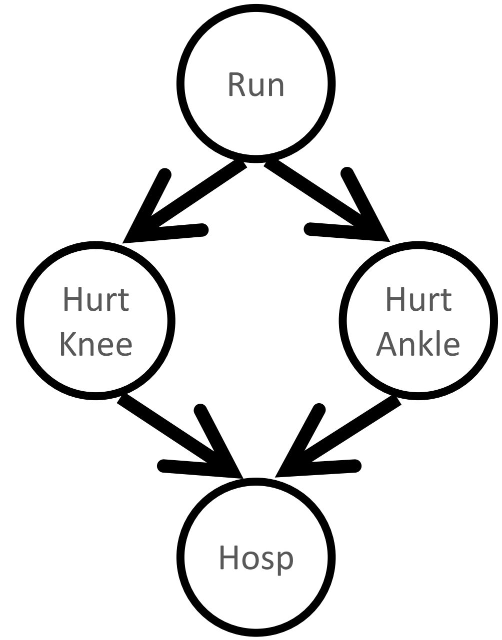
P(R=T)
.5

R	P(K=T)
F	0.01
T	0.1

A	K	P(H=T)
F	F	0
T	F	0.3
F	T	.9
T	T	.99

R	P(A=T)
F	0.05
T	0.2

Factorization



- Consider the joint probability of our example
  - The full p(R, A, K, H) is complex
  - What can we do to simplify?
  - Notice that A and K are independent given R
- Factor the joint probability according to the graph
  - $p(R, A, K, H) = p(H \mid A,K)\, p(A \mid R)\, p(K \mid R)\, p(R)$
  - This is much simpler to compute, with fewer conditional probabilities track.

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(A=T | H=T)?
  - Probability of ankle injury, given a trip to the hospital
  - Break down using the network and CPTs

$$p(A=T \mid H=T) = \frac{p(A=T, H=T)}{p(H=T)} = \frac{\sum_{r,k} p(R=r, K=k, A=T, H=T)}{\sum_{r,k,a} p(R=r, K=k, A=a, H=T)}$$

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

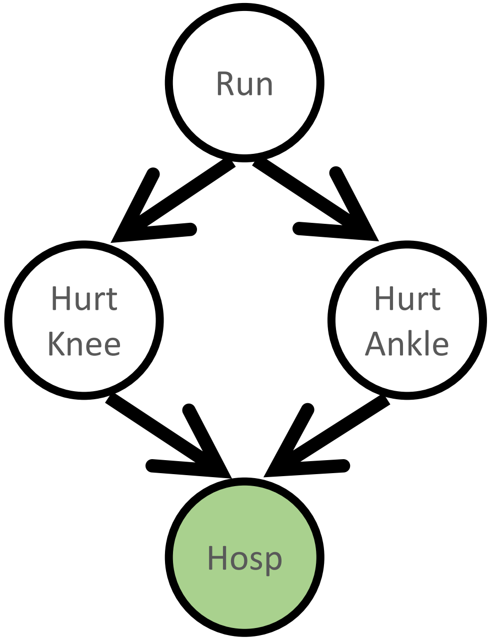
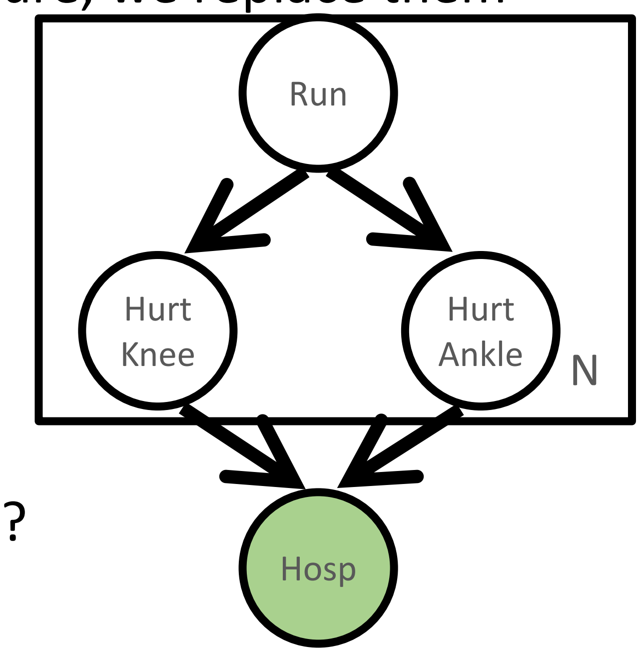


Plate Notation

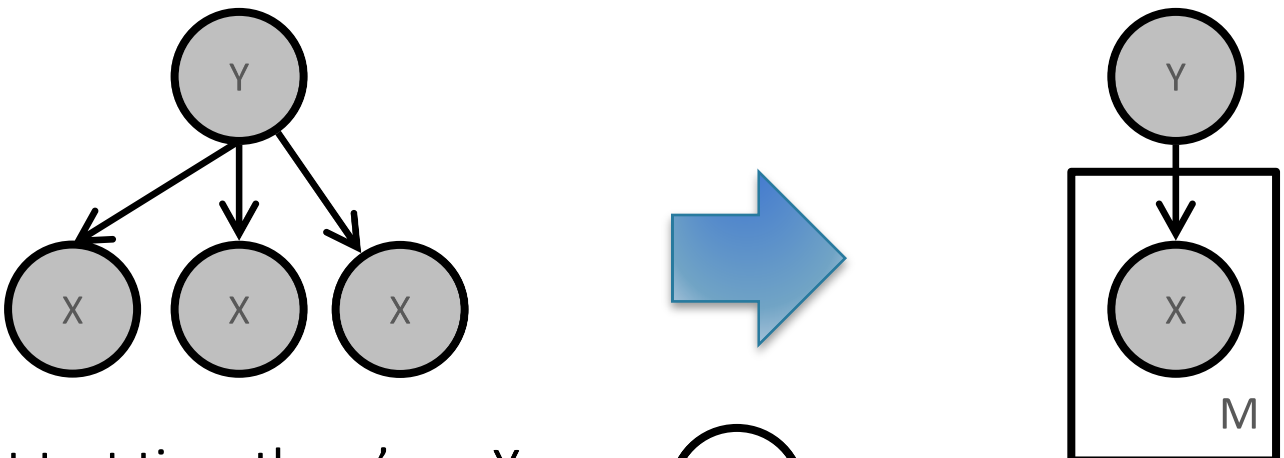
- Plates in graphical models
  - When many variables have same structure, we replace them with a plate
  - The plate indicates repetition



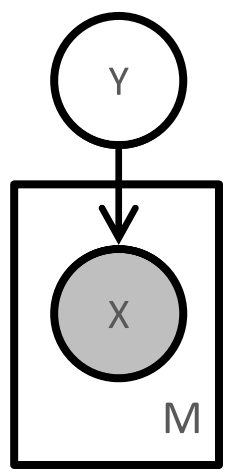
  - There are N days
  - Did Avery go to the hospital on any day?

Let’s consider a new model

- A model where we have label Y and example X

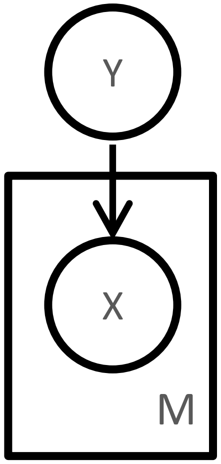


- At test time there’s no Y
    - Estimate Y using X
- What model is this?



Naïve Bayes

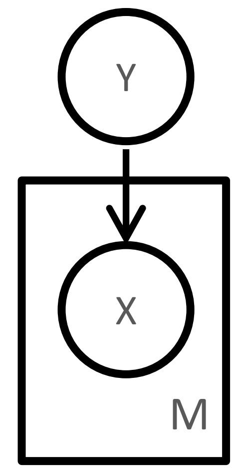
- Generative Story
  - Generate a label Y
  - Given Y, generate each feature X independently
- Learning
  - We observe X and Y, maximum likelihood solution
- Prediction
  - Compute most likely value for Y given X



Factorization

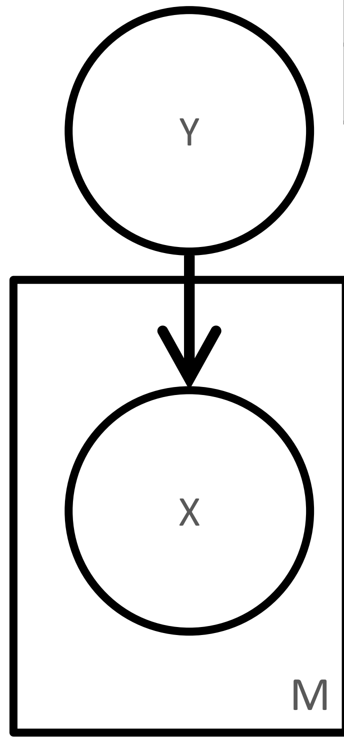
$$P(y,x) = P(x \mid y)P(y)$$

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



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

Argmax Derivation

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta \mid X, y)$$

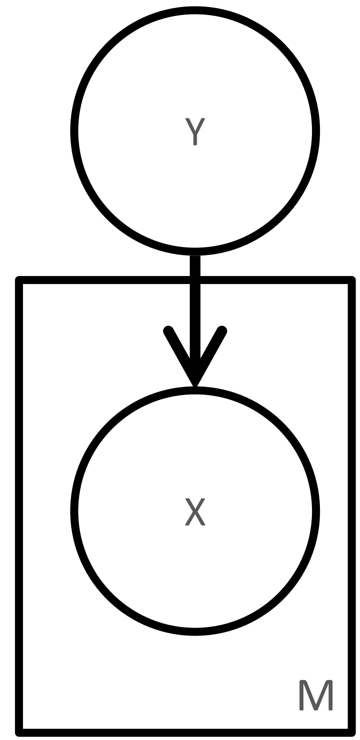
$$= \log p(y \mid \theta) + \log p(\theta) + \sum_{j=1}^{M} \log p(X_j \mid y, \theta)$$

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

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

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?
  - Can use EM! $\max P(X) = \sum_{y \in Y} P(Y, X)$
  - Unsupervised NB: clustering
  - Some labels: semi-supervised NB



Conditional Independence

- 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

- **Random variable X is conditionally independent of Y given Z if the probability of each is independent given Z**
- $p(x,y|z) = p(x|z)p(y|z)$
- $p(x|z, y) = p(x \mid z)$
- Example
  - X: I need an umbrella and Y: the ground is wet
  - Not independent!
  - If ground is wet, it’s probably raining and I’ll need an umbrella
  - I am told it is raining; knowing 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
  - $P(x \mid z, y) = p(x, z)$

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
- It's much easier to learn a smaller number of 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 grammatically nonsense.

Naïve Assumptions vs. Reality

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

Making more realistic assumptions

- Naïve Bayes makes assumptions
  - Features (X) conditionally independent given label (Y)
- What would be a more realistic assumption?
- How does independence fit in graphical models?

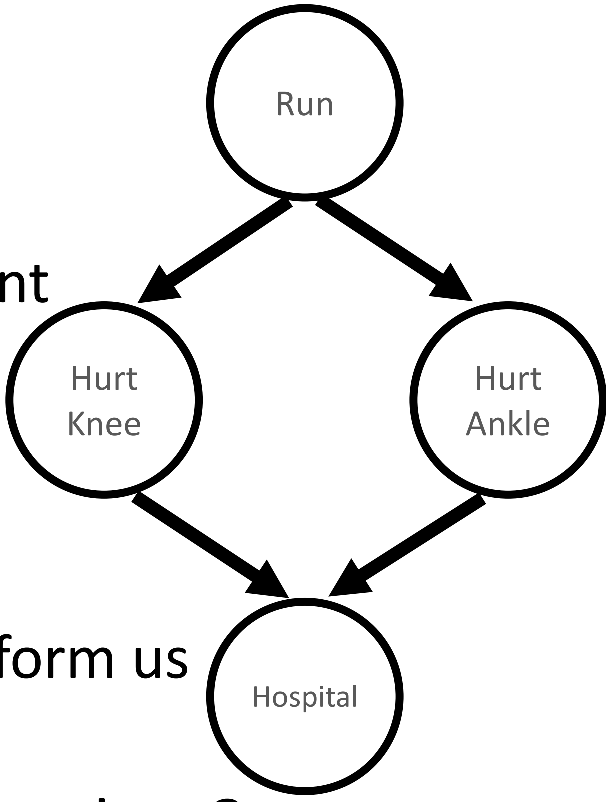
Independence

- 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

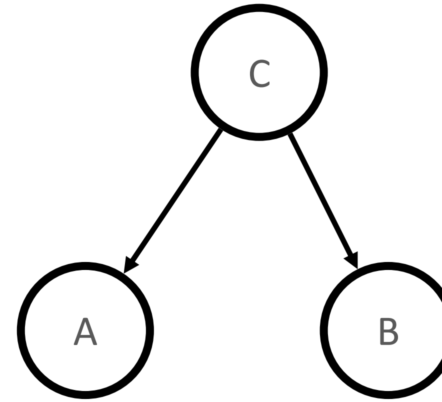
Conditional Independence

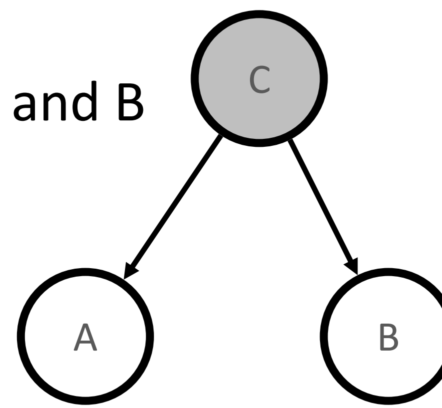
- Are Knee and Ankle independent?
  - No, but they are conditionally independent given Run
    - P(Knee, Ankle | Run)
      = p(Knee | Run) p(Ankle | Run)
    - Once we know whether Avery ran, no information about ankle injuries will inform us about knee injuries
- How do we know if something is independent?
  - We can read it from the paths of the graph!
  - No mathematical trickery needed



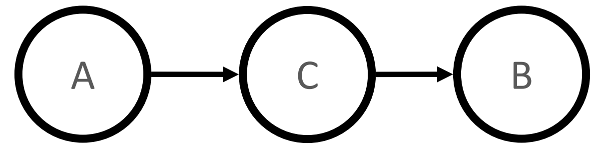
Example 1

- 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
  - Conditioning on C “blocks the path" between A and B

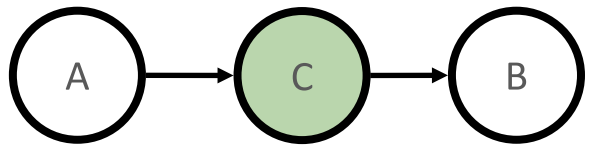




Example 2

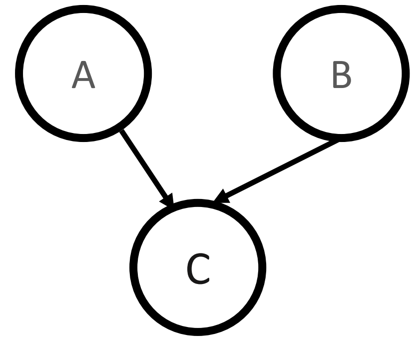


- Are A and B independent?
  - No. A cause C which causes B
- Are A and B conditionally independent?
  - Yes. Why?



  - 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



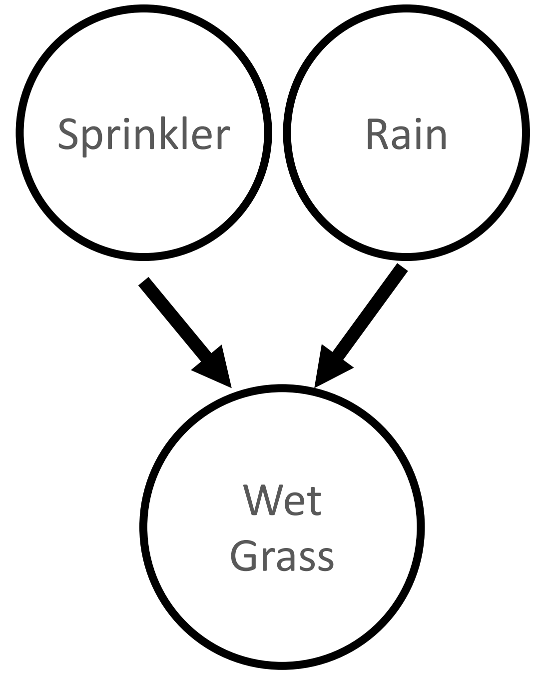
- Are A and B independent?
  - Yes. A and B are generated without common parents
- Are A and B conditionally independent given C?
  - No. Why?
  - The connection of A and B to C is "head-to-head"
    - Creates a dependence when C is observed
  - When C is unobserved, the path is **blocked**
  - When 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

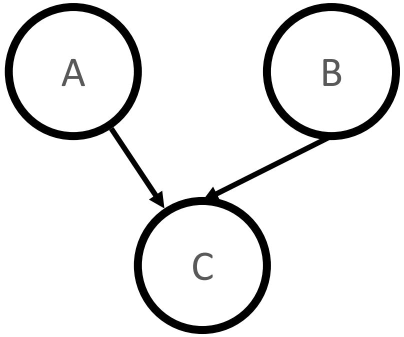
Head-to-head dependence

- Suppose you see the grass outside is wet
- The two causes (sprinkler/rain) compete to explain the grass



Explaining Away

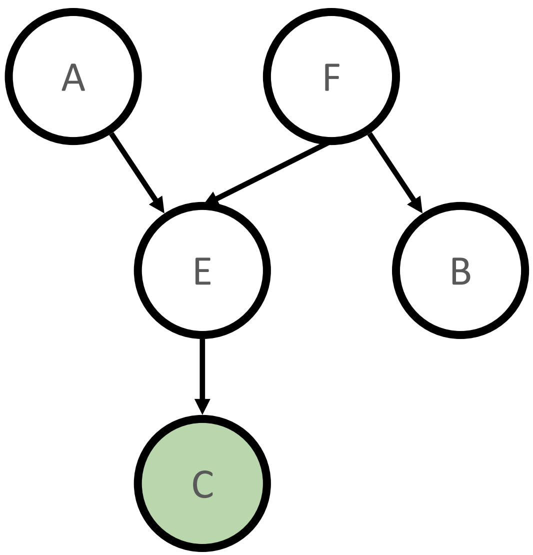
- This makes sense
  - The rain explained the grass, so sprinkler is now less likely
  - The rain explained away the state of the grass
  - Don’t "need" to use sprinkler to explain it
- Thus, the observed head-to-head is unblocked
  - Once we know the value of C,
    we learn something about A and B



D-Separation

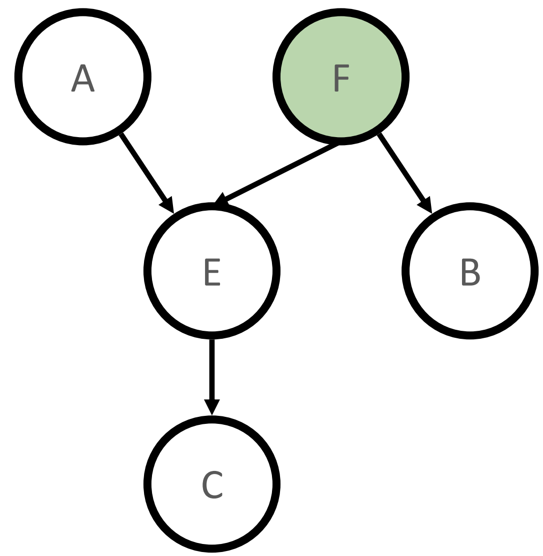
- Two nodes A and B are **d-separated** given observed node(s) C if all paths between A and B are blocked
  - Blocked paths: two arrows on the path meet head-to-tail or tail-to-tail at a node in set C
  - Or, the arrows meet head-to-head at a node which isn't in C
    - And none of its descendants are either
- If two (sets of) nodes are d-separated they are conditionally-independent!

Are A and B d-separated?



No

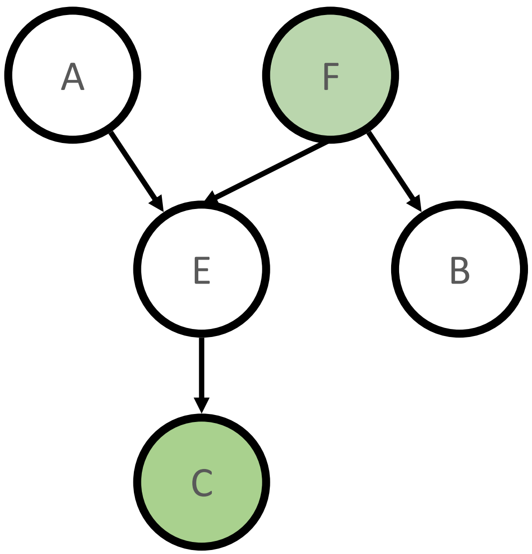
C is a descendent of head to head E



Yes

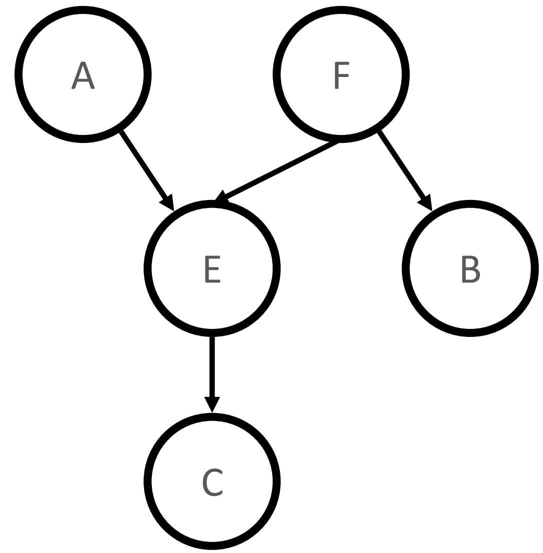
F is a tail to tail node

Are A and B d-separated?



Yes

F is a tail-to-tail node



Yes

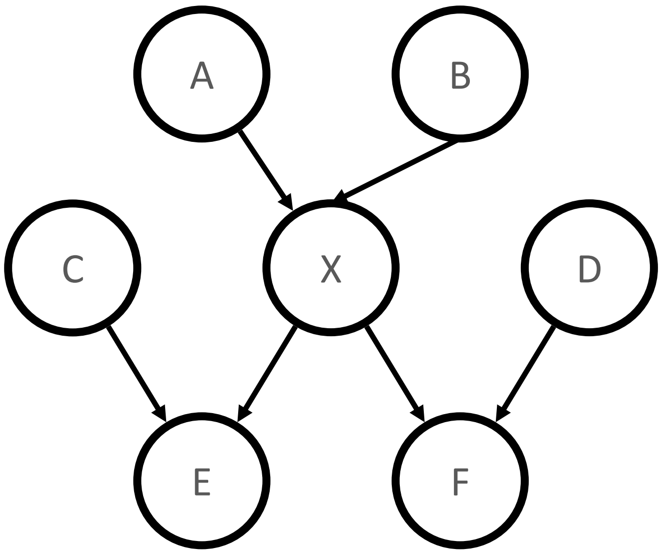
E is head-to-head

Isolating Nodes

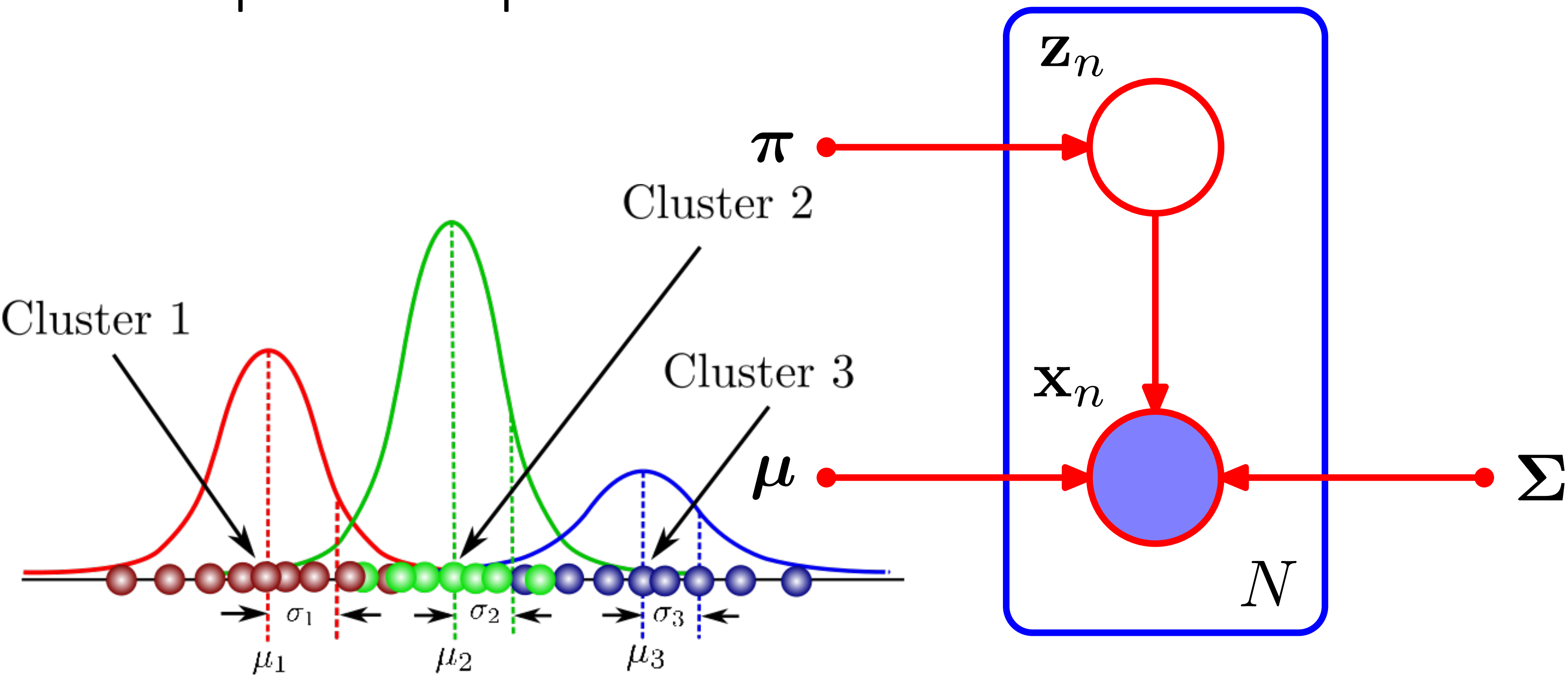
- How do we isolate a variable in the graph?
  - We know how to make it conditionally independent
  - We want to experiment with a variable in isolation
  - We don’t want to enumerate all possible values of the whole network

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

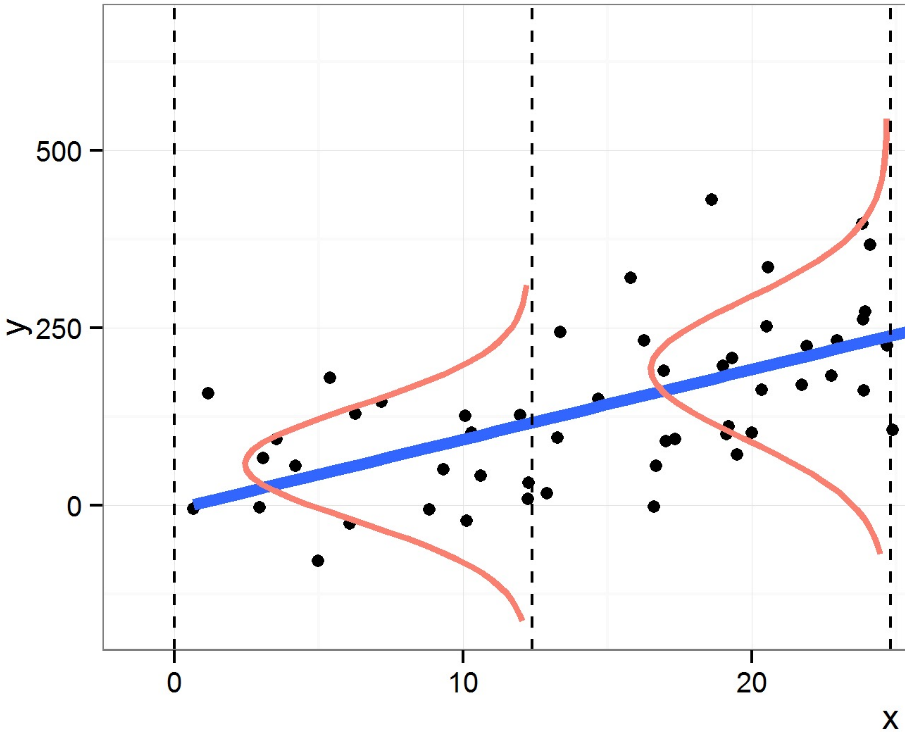


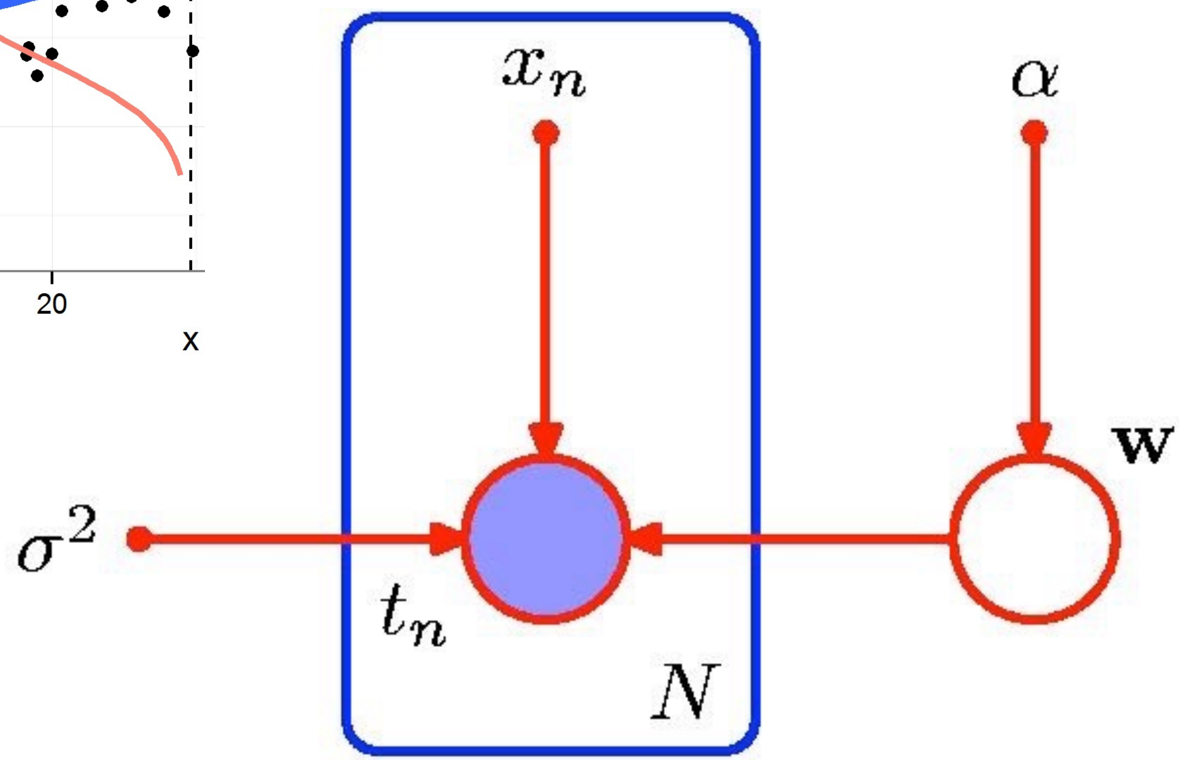
Graphical Representation



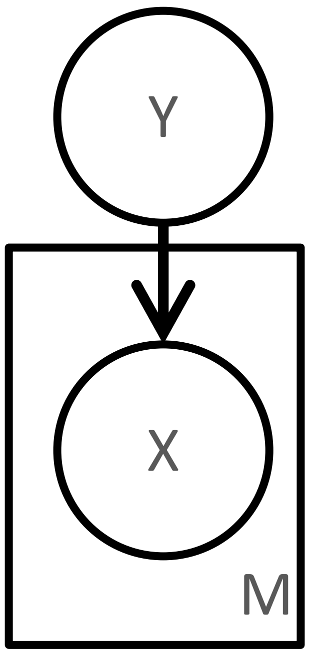
Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{\mathbf{x}_n\}$, with corresponding latent points $\{\mathbf{z}_n\}$, where $n = 1, \ldots, N$.

Graphical Representation





Return to Naïve Bayes



q(Y=1)
.6

Y	$q(X_i=1 \mid Y)$
0	.8
1	.4

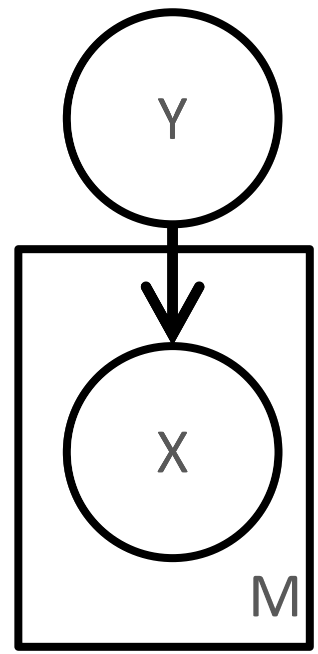
M Tables

2*M parameters

$$\arg\max_{y\in\{1\ldots k\}} p(y, x_1 \ldots x_d) = \arg\max_{y\in\{1\ldots k\}} \left(q(y) \prod_{j=1}^{d} q_j(x_j|y) \right)$$

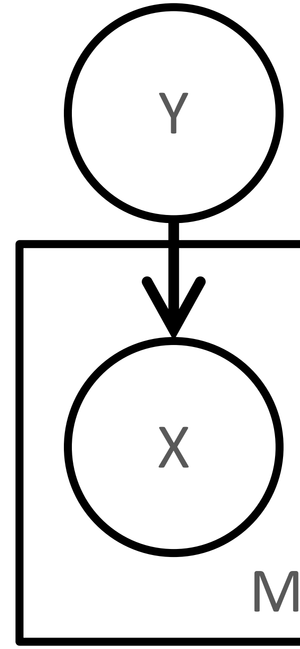
Maximum Likelihood Estimate for NB

- Suppose I have ten emails:
  one is spam (y=1) and nine are not (y=0)
- What's the MLE for p(y)?



p(spam)
0.1

Maximum Likelihood Estimate for NB



- Suppose I have only two emails:
  - Spam: “you win jackpot”
  - Not: “how are you”

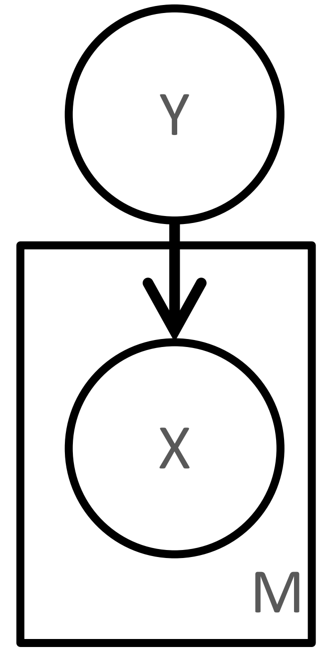
- What is p(“jackpot” | y)?

Y	p(“jackpot”=1 | Y)
Spam	0.33
Not	0.0

Y	p(“you”=1 | Y)
Spam	0.33
Not	0.33

Smoothing the MLE

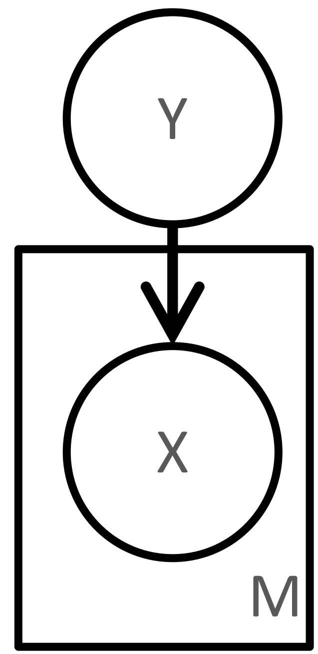
Y	p("jackpot"=1 | Y)
Spam	2/4
Not	1/4



$$\arg\max_{y \in \{1 \ldots k\}} p(y, x_1 \ldots x_d) = \arg\max_{y \in \{1 \ldots k\}} \left(q(y) \prod_{j=1}^{d} q_j(x_j | y) \right)$$

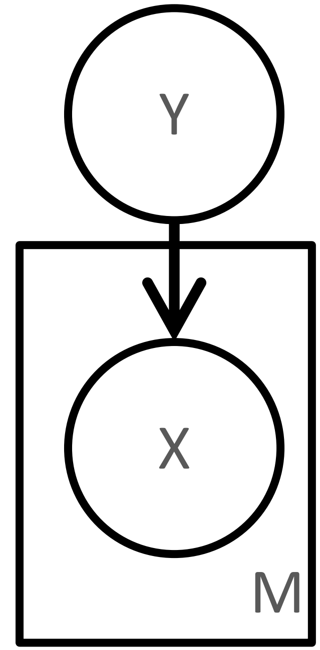
Maximum Likelihood Estimate for NB

- Suppose I have eleven emails:
  one is spam and nine are not
  one is unlabeled
- What's the MLE for q(y)?



q(Y=1)
?

Maximum Likelihood Estimate for NB



- Suppose I have only two emails:
  - Spam: “you win jackpot”
  - Not: “how are you”

- What is p(“jackpot” | y)?

Y	**p(“jackpot”=1 | Y)**
Spam	0.33
Not	0.0

Y	**p(“you”=1 | Y)**
Spam	0.33
Not	0.33

Expectation Maximization for Naïve Bayes

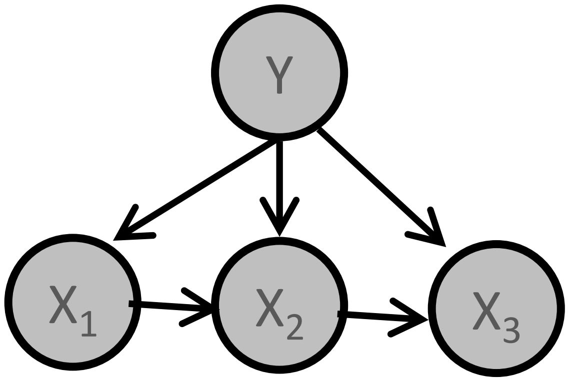
$$p(\underline{x}) = \sum_{y=1}^{k} p(\underline{x}, y) = \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j|y) \right)$$

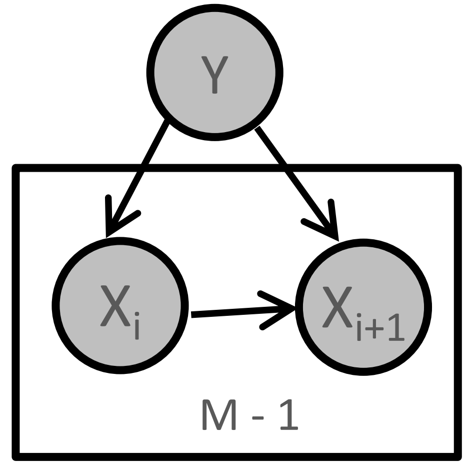
E-Step: $\delta(y|i) = p(y|\underline{x}^{(i)}; \underline{\theta}^{t-1}) = \dfrac{q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}{\sum_{y=1}^{k} q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}$

M-Step: $q^t(y) = \dfrac{1}{n} \sum_{i=1}^{n} \delta(y|i) \qquad q_j^t(x|y) = \dfrac{\sum_{i:x_j^{(i)}=x} \delta(y|i)}{\sum_i \delta(y|i)}$

http://www.cs.columbia.edu/~mcollins/em.pdf

Can we make weaker assumptions?





Y	**p("jackpot"=1 | Y)**
Spam	
Not	

Y	**X_i**	**p(X_{i+1}="win" | X_i, Y)**
1	"you"	
0	”you”	
1	"packers"	
0	"packers"	