CS 6347

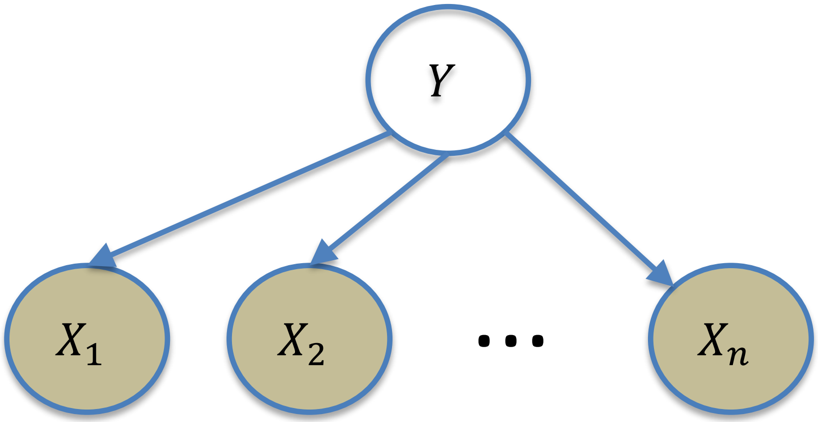
Lecture 4

Markov Random Fields

D-separation

- Let $I(p)$ be the set of all independence relationships in the joint distribution p and $I(G)$ be the set of all independence relationships implied by the graph G
- We say that G is an I-map for $I(p)$ if $I(G) \subseteq I(p)$
- Theorem: the joint probability distribution, p, factorizes with respect to the DAG $G = (V, E)$ iff G is an I-map for $I(p)$
- An I-map is perfect if $I(G) = I(p)$
  - Not always possible to perfectly represent all of the independence relations with a graph

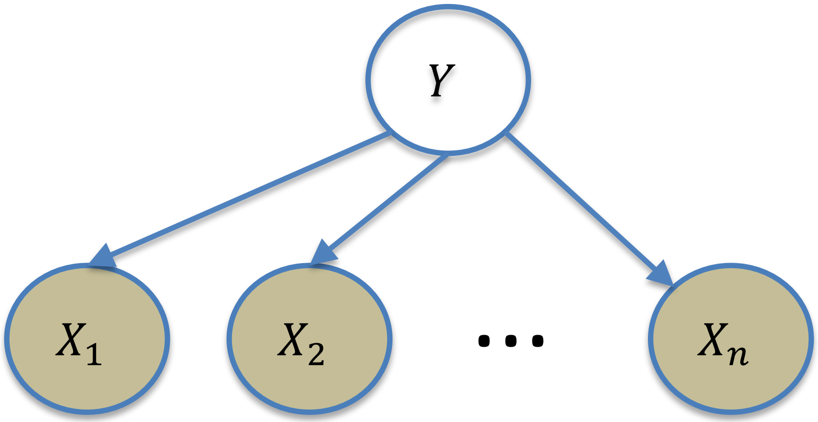
Naïve Bayes



$$p(y, x_1, \dots, x_n) = p(y)p(x_1|y) \dots p(x_n|y)$$

- In practice, we often have variables that we observe directly and those that can only be observed indirectly

Naïve Bayes



$$p(y, x_1, \dots, x_n) = p(y)p(x_1|y) \dots p(x_n|y)$$

- This model assumes that $X_1, \dots, X_n$ are independent given Y, sometimes called naïve Bayes

Example: Naïve Bayes

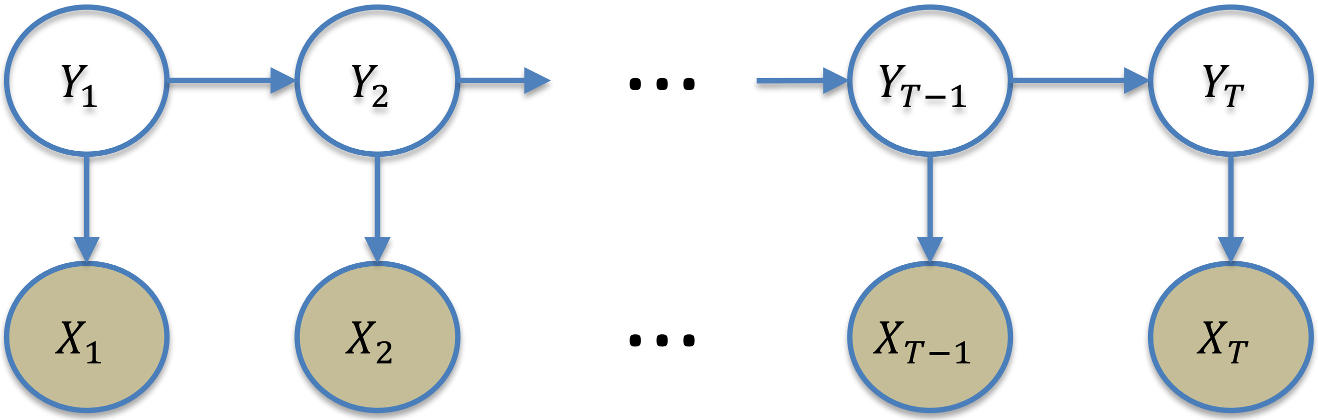
- Let Y be a binary random variable indicating whether or not an email is a piece of spam

- For each word in the dictionary, create a binary random variable X_i indicating whether or not word i appears in the email

- For simplicity, we will assume that $X_1, \dots, X_n$ are independent given Y

- How do we compute the probability that an email is spam?

Hidden Markov Models



$$p(x_1, \ldots, x_T, y_1, \ldots, y_T) = p(y_1)p(x_1|y_1)\prod_{t=2} p(y_t|y_{t-1})p(x_t|y_t)$$

- Used in coding, speech recognition, etc.
- Independence assertions?

Markov Random Fields (MRFs)

- A **Markov random field** is an undirected graphical model
  - Undirected graph $G = (V, E)$
  - One node for each random variable
  - Nonnegative potential function or "factor" associated with cliques, C, of the graph
  - Nonnegative potential functions represent interactions and need not correspond to conditional probabilities (may not even sum to one)

Markov Random Fields (MRFs)

- A **Markov random field** is an undirected graphical model
  - Corresponds to a **factorization** of the joint distribution

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

$$Z = \sum_{x'_1, \dots, x'_n} \prod_{c \in C} \psi_c(x'_c)$$

Markov Random Fields (MRFs)

- A **Markov random field** is an undirected graphical model
  - Corresponds to a **factorization** of the joint distribution

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

$$Z = \sum_{x'_1, \dots, x'_n} \prod_{c \in C} \psi_c(x'_c)$$

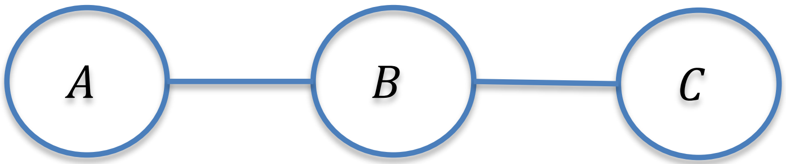
Normalizing constant, Z, often called the partition function

MRF Examples

- Given a graph $G = (V, E)$, express the following as probability distributions that factorize over G
  - Uniform distribution over independent sets
  - Uniform distribution over vertex covers

(done on the board)

Independence Assertions



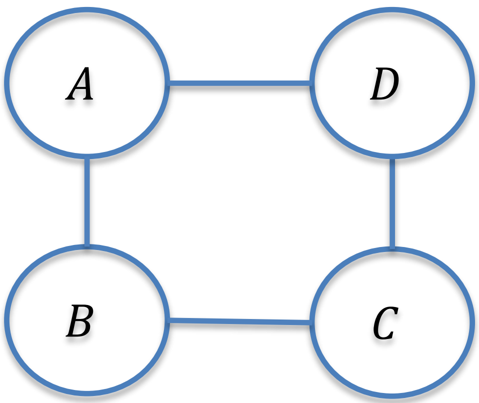
$$p(x_A, x_B, x_C) = \frac{1}{Z}\psi_{AB}(x_A, x_B)\psi_{BC}(x_B, x_C)$$

- How does separation imply independence?
- Showed that $A \perp C \mid B$ on board last lecture

Independence Assertions

- If $X \subseteq V$ is graph separated from $Y \subseteq V$ by $Z \subseteq V$, (i.e., all paths from X to Y go through Z) then $X \perp Y \mid Z$

- What independence assertions follow from this MRF?



Independence Assertions

- Each variable is independent of all of its non-neighbors given its neighbors
  - All paths leaving a single variable must pass through some neighbor
- If the joint probability distribution, p, factorizes with respect to the graph G, then G is an I-map for p
- If G is an I-map of a **strictly positive** distribution p, then p factorizes with respect to the graph G
  - Hamersley-Clifford Theorem

BNs vs. MRFs

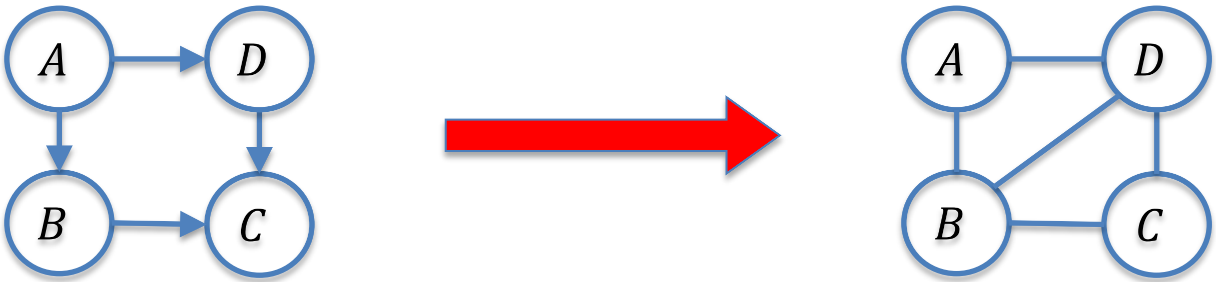
Property	Bayesian Networks	Markov Random Fields
Factorization	Conditional Distributions	Potential Functions
Distribution	Product of Conditional Distributions	Normalized Product of Potentials
Cycles	Directed Not Allowed	Allowed
Partition Function	1	Potentially NP-hard to Compute
Independence Test	d-Separation	Graph Separation

Moralization

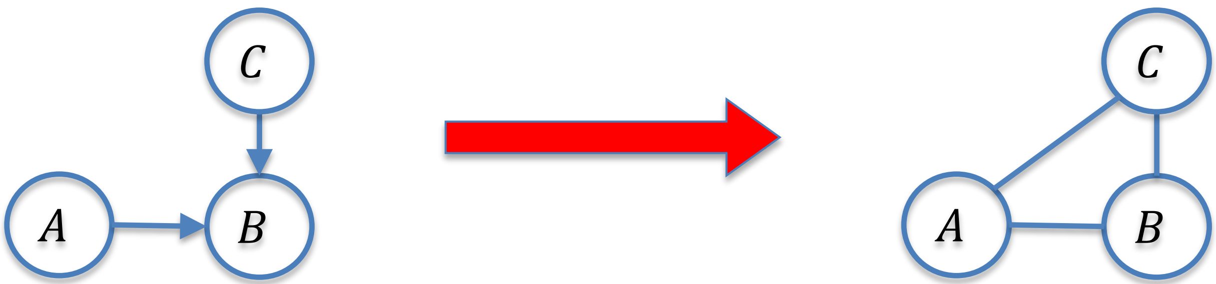
- Every Bayesian network can be converted into an MRF with some possible loss of independence information
  - Remove the direction of all arrows in the network
  - If A and B are parents of C in the Bayesian network, we add an edge between A and B in the MRF
- This procedure is called "**moralization**" because it "marries" the parents of every node

Moralization

- Every Bayesian network can be converted into an MRF with some possible loss of independence information
  - Remove the direction of all arrows in the network
  - If A and B are parents of C in the Bayesian network, we add an edge between A and B in the MRF
- This procedure is called "**moralization**" because it "marries" the parents of every node



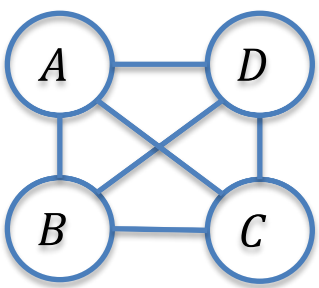
Moralization



- What independence information is lost?

Factorizations

- Many factorizations over the same graph may represent the same joint distribution
  - Some are better than others (e.g., they more compactly represent the distribution)
  - Simply looking at the graph is not enough to understand which specific factorization is being assumed

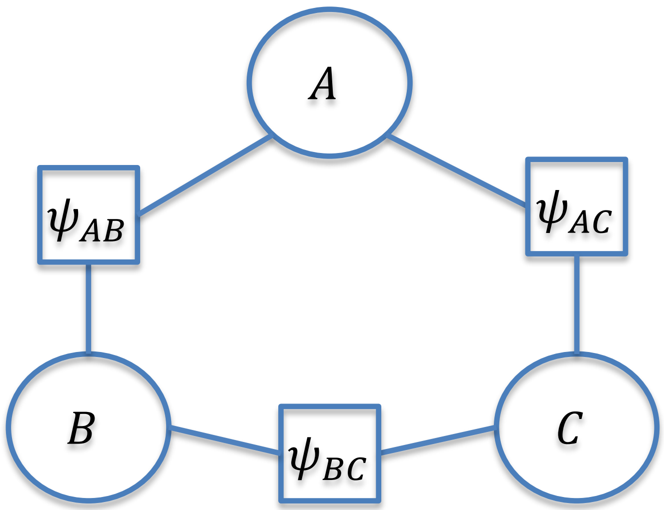


Factor Graphs

- Factor graphs are used to explicitly represent a given factorization over a given graph
  - Not a different model, but rather different way to visualize an MRF
  - Undirected bipartite graph with two types of nodes: variable nodes (circles) and factor nodes (squares)
  - Factor nodes are connected to the variable nodes on which they depend

Factor Graphs

$$p(x_A, x_B, x_C) = \frac{1}{Z}\psi_{AB}(x_A, x_B)\psi_{BC}(x_B, x_C)\,\psi_{AC}(x_A, x_C)$$



MRF Examples

- Given a graph $G = (V, E)$, express the following as probability distributions that factorize over G

  - Express the uniform distribution over matchings (i.e., subsets of edges such that no two edges in the set have a common endpoint) as a factor graph

(done on the board)

Conditional Random Fields (CRFs)

- Undirected graphical models that represent conditional probability distributions $p(Y \mid X)$
  - Potentials can depend on both X and Y

$$p(Y \mid X) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(x_c, y_c)$$

$$Z(x) = \sum_{y'} \prod_{c \in C} \psi_c(x_c, y'_c)$$

Log-Linear Models

- CRFs often assume that the potentials are log-linear functions

$$\psi_c(x_c, y_c) = \exp(w^T f_c(x_c, y_c))$$

f_c is referred to as a **feature vector** and w is some vector of feature weights

- The feature weights are typically learned from data

- CRFs don't require us to model the full joint distribution (which may not be possible anyhow)

Conditional Random Fields (CRFs)

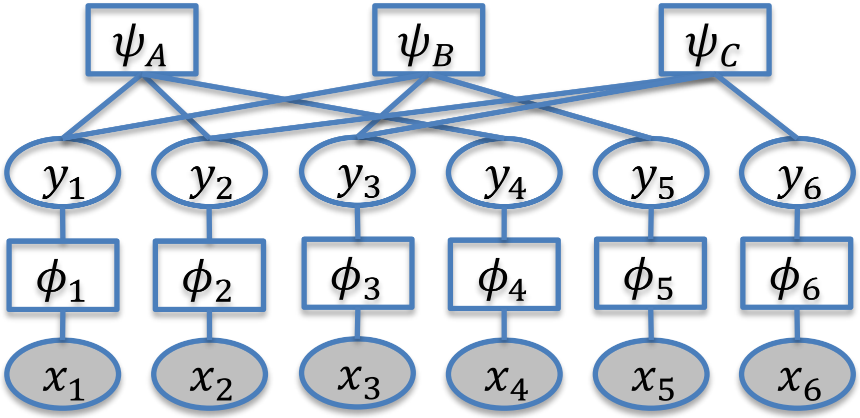
- Binary image segmentation
  - Label the pixels of an image as belonging to the foreground or background
  - +/- correspond to foreground/background
  - Interaction between neighboring pixels in the image depends on how similar the pixels are
    - Similar pixels should preference being in the same part of the image

Conditional Random Fields (CRFs)

- Binary image segmentation
  - This can be modeled as a CRF where the image information (e.g., pixel colors) is observed, but the segmentation is unobserved
  - Because the model is conditional, we don't need to describe the joint probability distribution of (natural) images and their foreground/background segmentations
  - CRFs will be particularly important when we want to learn graphical models from observed data

Low Density Parity Check Codes

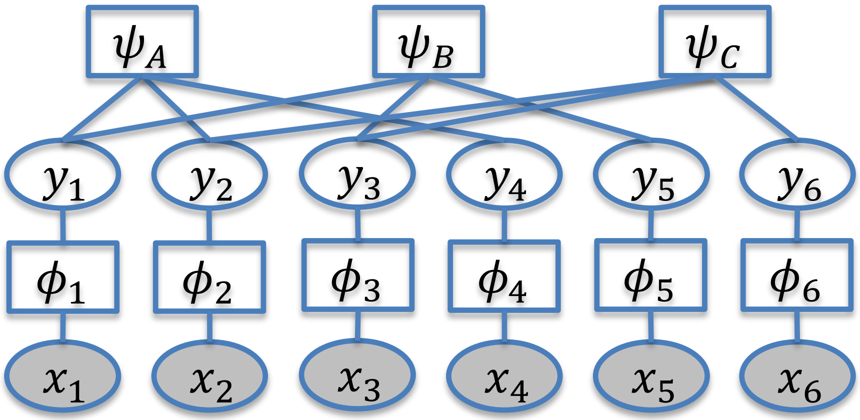
- Want to send a message across a noisy channel in which bits can be flipped with some probability – use error correcting codes



- ψ_A, ψ_B, ψ_C are all parity check constraints: they equal one if their input contains an even number of ones and zero otherwise

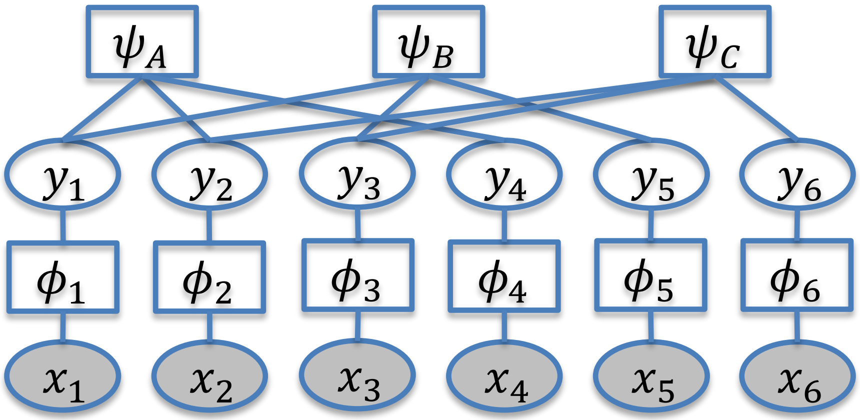
- $\phi_i(x_i, y_i) = p(y_i|x_i)$, the probability that the ith bit was flipped during transmission

Low Density Parity Check Codes



- The parity check constraints enforce that the y's can only be one of a few possible codewords: 000000, 001011, 010101, 011110, 100110, 101101, 110011, 111000
- Decoding the message that was sent is equivalent to computing the most likely codeword under the joint probability distribution

Low Density Parity Check Codes



- Most likely codeword is given by MAP inference

$$\arg\max_y p(y|x)$$

- Do we need to compute the partition function for MAP inference?