CS7140: Advanced Machine Learning **Spring 2020**

Lecture 1: Directed and Undirected Graphical Models

Lecturer: Rose Yu *Scribes: Divyarajsinhji Solanki, Wan He*

1.1 Graphical Models

Probabilistic graphical models provide framework for modeling relationships between random variables. Below figure gives brief introduction of the notations used to represent graphical models.

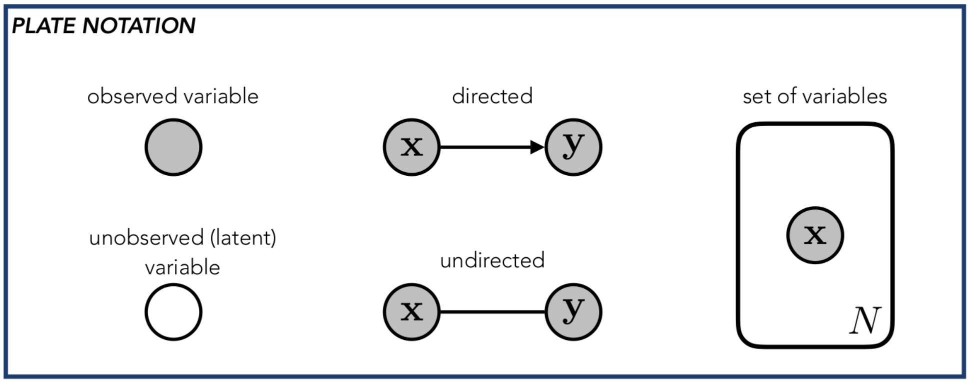


Figure 1.1: Plate notation

- Observed variable: The random variable whose values we can observe directly.
- Unobserved(latent) variable: The random variable whose values we cannot observe directly.
- Directed edges: There are used for causal or asymmetric interactions. In the above figure, X could be weather and Y could be person's mood, which models the effect of weather on person's mood.
- Undirected edges: These are used for a relationship between two variables but the exact direction of the information flow is not clear.
- Plate notation(Right most side of Figure 1.1): Plate notation is a method of representing variables that repeat i.i.d in a graphical model. Instead of drawing each repeated variable individually, a plate or rectangle is used to group variables into a subgraph that repeat together, and a number on the plate represents the number of repetitions of the subgraph in the plate.

1.2 Bayesian Network

Bayesian network representation is a directed acyclic graph (DAG), whose nodes are the random variables in our domain and whose edges correspond to direct influence of one node on another. If the graph is not acyclic then assigned probabilities are not proper. So it is essential for a Bayesian network to be acyclic.

1.2.1 Definition

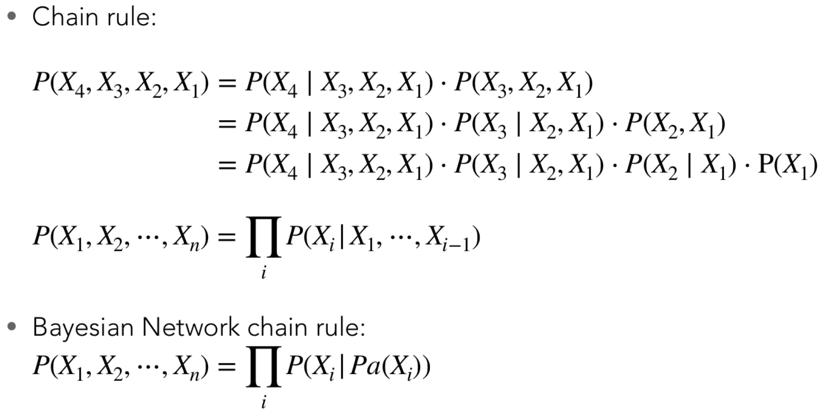
A DAG for Bayesian networks contains:

1. A node for each random variable X_i.

2. A conditional distribution per node $P(X_i \mid Pa(X_i))$.

Here $Pa(X_i)$ denotes the set of parents of the node X_i.

Below is a comparison of general chain rule and Bayesian Network Chain Rule. Note that no assumption of independence are made for general chain rules. As both formulas above represent the same distribution of variables, we can see that a Bayesian Network encodes conditional independence.



- Chain rule:

$$\begin{aligned} P(X_4, X_3, X_2, X_1) &= P(X_4 \mid X_3, X_2, X_1) \cdot P(X_3, X_2, X_1) \\ &= P(X_4 \mid X_3, X_2, X_1) \cdot P(X_3 \mid X_2, X_1) \cdot P(X_2, X_1) \\ &= P(X_4 \mid X_3, X_2, X_1) \cdot P(X_3 \mid X_2, X_1) \cdot P(X_2 \mid X_1) \cdot \mathrm{P}(X_1) \end{aligned}$$

$$P(X_1, X_2, \cdots, X_n) = \prod_i P(X_i | X_1, \cdots, X_{i-1})$$

- Bayesian Network chain rule:
$$P(X_1, X_2, \cdots, X_n) = \prod_i P(X_i | Pa(X_i))$$

Figure 1.2: Chain rule comparison

Bayesian network chain rule allows dependencies to be implicitly represented. That's why it is more compact than general chain rule. We will see that in the example below.

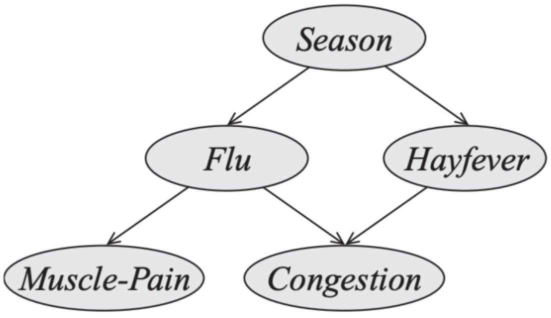


Figure 1.3: Sample Bayesian network

For above example, joint distribution of all the variables can be defined as:
Chain rule:

$$P(S, F, H, C, M) = P(S)P(F \mid S)P(H \mid F, S)P(C \mid F, H, S)P(M \mid F, S, H, C)$$

Bayesian Network chain rule:

$$P(S, F, H, C, M) = P(S)P(F \mid S)P(H \mid S)P(C \mid F, H)P(M \mid F)$$

From above two equations, Conditional Independence between variables can be inferred:

$$(H \perp F \mid S), \quad (C \perp S \mid F, H), \quad (M \perp C, H, S \mid F)$$

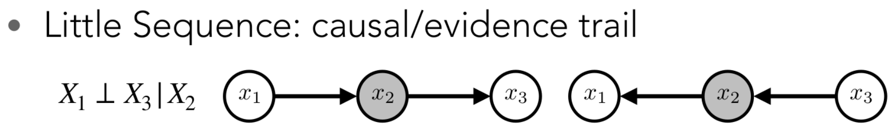
Here we can see that representation for BN chain rule seems more compact. As both formula represent similar distribution, by comparing both formulas we can write down conditional independence. For example, compare the corresponding distribution for H. General chain rule has the notation $P(H \mid F, S)$ and BN chain rule has $P(H \mid S)$. Now from rules of conditional independence we can infer that $(H \perp F \mid S)$. Therefore, a Bayesian Network encodes conditional independence.

1.2.2 Independence maps (I-map)

Let's first define a set of independencies associated with a distribution before coming to the definition of I-maps. If P is a distribution in $\mathcal{X}$, then we define $\mathcal{I}(P)$ as the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P. And let $\mathcal{I}(G)$ denote the set of all conditional independencies implied by the directed ayclic graph(DAG) G. Now, we can say that a DAG G is an I-map of a distribution P if $\mathcal{I}(G) \subseteq \mathcal{I}(P)$.

1.2.3 D(Dependence)-Separation

To decide independence, we use D-separation. Consider a toy example of three variables Bayesian Network structure.



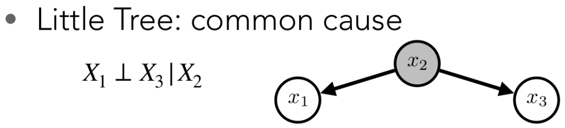
- Little Sequence: causal/evidence trail

$X_1 \perp X_3 | X_2$

Figure 1.4: Little Sequence

Figure 1.4 shows one type of relation we might find in DAG. It's called **"little sequence"** and shows direct causal effect of variables. Here, X_1 and X_3 are independent given observed variable X_2, $X_1 \perp X_3 \mid X_2$. It is also called causal/evidence trail based on direction of cause-effect. It is necessary for X_2 to be observed for mentioned independence to be true.

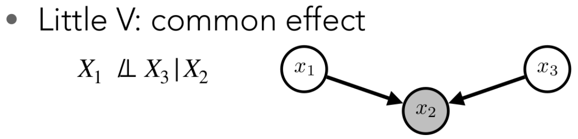
Figure 1.5 shows relationship between variables called **"Little Tree"**. It says that if two variable share a parent then they are independent of each other given that their parent is observed. In given example, nodes X_1 and X_3 shares the parent X_2. It is also called "common cause" because variable X_1 and X_3 are influenced by same variable X_2. And as before, it is necessary for X_2 to be observed variable.



- Little Tree: common cause

$X_1 \perp X_3 | X_2$

Figure 1.5: Little Tree



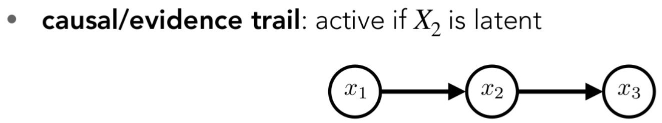
- Little V: common effect

$X_1 \not\perp X_3 | X_2$

Figure 1.6: Little V

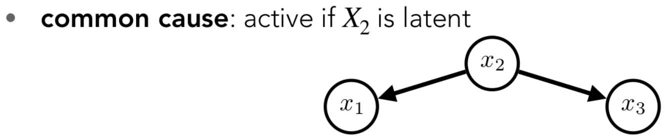
Figure 1.6 shows a relationship called **"Little V"**. This one is different from previous ones. It says that if two variables can influence a same third variable, then they are dependent given that third variable is observed. It might happen that those two variables were independent before we observed the third variable. In above example, X_1 and X_3 and dependent if X_2 is observed.

Active Trail: when influence can flow over to given path in the graph, it is said to be active trail. We can apply D-separation to find active trails. Figure 1.7 shows that when variable X_2 is not observed(latent) in the little sequence structure then the trail is active. That means that X_1 and X_3 are not independent.



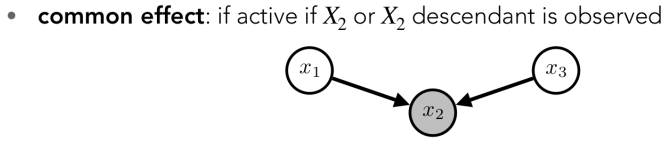
- **causal/evidence trail**: active if X_2 is latent

Figure 1.5: Causal/evidence trail



- **common cause**: active if X_2 is latent

Figure 1.6: Common cause



- **common effect**: if active if X_2 or X_2 descendant is observed

Figure 1.7: Common effect

Figure 1.8 shows that when variable X_2 is not observed (latent) in the little tree structure then the trail is active. That means that X_1 and X_3 are not independent. For the case of little V structure, the scenario is different than other ones. Figure 1.9 shows that. It says if a common child node of two nodes is observed then those two parent nodes are independent. Here in above figure 1.9, given X_2, X_1 and X_3 are independent.

1.2.4 I-Equivalence

Two Bayesian networks are I-equivalent if they encode precisely the same conditional independence assertions. In other words, two Bayesian networks are I-equivalent if:

1. they have same skeleton. Skeleton of a graph is a structure same as the original graph but all the edges are undirected.

2. they have same V-structure.

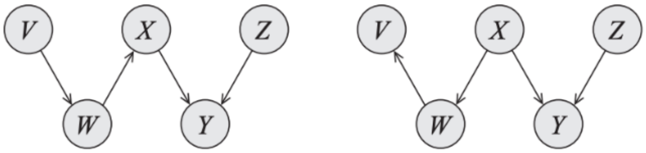


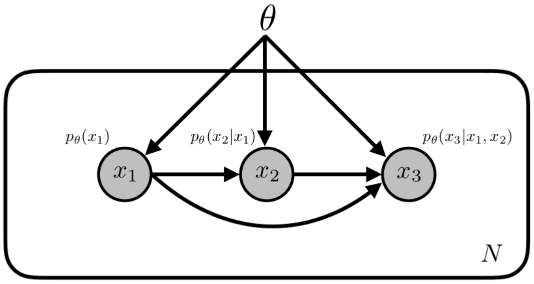
Figure 1.8: Example for I-equivalence

Figure 1.10 is an example of two networks which are I-equivalent. If we were to make all the edges in both graph undirected, we would end up with two same graphs. Also, V-structure is also same in both graphs, which is $X \rightarrow Y \leftarrow Z$.

1.2.5 Example Bayesian Networks

Here are some of the applications in which Bayesian networks are used to model the problem.

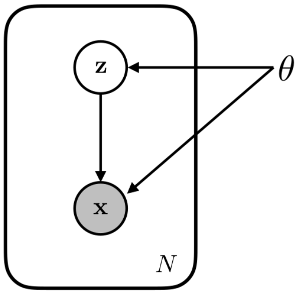
- **Autoregressive Model**: In this architecture next variable is conditioned upon the previous n variable as shown in figure 1.11. This kind of model can be used for time series forecasting.



$$p(x_1, x_2, x_3) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2)$$

Figure 1.9: Autoregressive Model

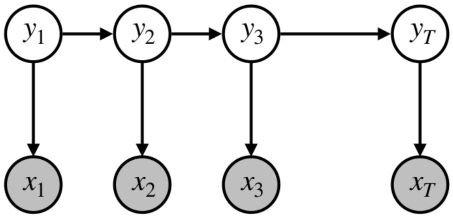
- **Latent Variable Model**: We want to infer latent variable Z from observed variable X. For example, observed variables can be words in a document and latent variables can be classes in which the word belongs e.g sports, education, art, etc.



$$p(x, z) = p(z)p(x \mid z)$$

Figure 1.10: Latent Variable Model

- **Hidden Markov Model**: It is used to model sequential relation between unobserved variables. One of the application of such model is speech recognition.



$$p(x_1, \cdots, x_T, y_1, \cdots, y_T) = p(y_1)p(x_1 \mid y_1)\prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t)$$

Figure 1.11: Hidden Markov Model

1.3 Markov Random Field

The representation power of the Bayesian network is limited in the sense that some independence information cannot be incorporated by it. Also, in the cases where interactions amongst the variables are symmetrical, incorporating directions into the graphs becomes redundant. An alternative graphical representation for the random variables, namely, the undirected graphical model or the Markov model is hence motivated to be introduced. Nodes in the Markov model represents random variables as in the Bayesian networks.

Definition 1.1 *A **Markov Random Field(MRF)** is an undirected graphical model for the joint probability distribution of a set of random variables satisfying the Markov properties.*

1.3.1 Local Markov Properties

The pairwise independence property requires any two random variables that are non-adjacent to each other to be conditionally independent given all other nodes in the graph. Formally,

Definition 1.2 (pairwise independence) *Given an undirected graph $\boldsymbol{G}$={V,E}, the random variable vector $\boldsymbol{X}$ with joint distribution P is Markov with respect to $\boldsymbol{G}$ if and only if for any pair of non-adjacent random variables X_u, X_v, the conditional independence $(X_u \perp X_u) \mid X_{V\setminus\{u,v\}}$ is satisfied.*

$$\mathcal{I}_P(G) = \{X_u \perp X_v \mid X_{V\setminus\{u,v\}} : \{u,v\} \notin E\}$$

Definition 1.3 (local independence) *Given an undirected graph $\boldsymbol{G} = \{V, E\}$, a random variable X_v with joint distribution P is Markov with respect to $\boldsymbol{G}$ if and only if it is independent of the rest of the nodes in the graph given its immediate neighbors. [KF09]*

$$\mathcal{I}_l(G) = \{X_v \perp X_{V\setminus\{v\cup N(v)\}} \mid X_{N(v)} : v \in V\}$$

1.3.2 Global Markov Property

Definition 1.4 (Separation) *Given an undirected graph $\boldsymbol{G} = \{V, E\}$, for any three disjoint subsets A, B, C $\subset$ V, B separates A and C if any path from A to C must pass through some vertex in B.*

Definition 1.5 (global independence) *Given an undirected graph $\boldsymbol{G} = \{V, E\}$, a random variable vector* **X** *with joint distribution P is globally Markov with respect to $\boldsymbol{G}$ if for any three disjoint subsets X_A, X_B, X_C $\subset X$, (i.e. A, B, C $\subset$ V) where X_B separates X_A and X_C, we have X_A is independent of X_C given X_B.*

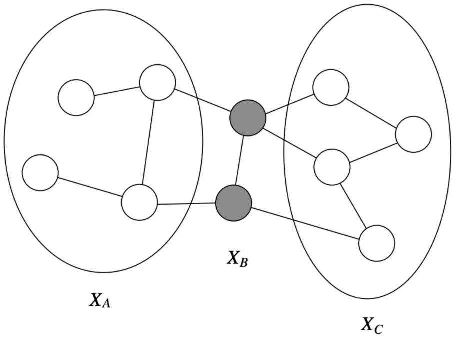


Figure 1.12: Markov Network Independence

This conditional independence $X_A \perp X_C \mid X_B$ is given by graph separation If there is no path from a $\in$ A to c $\in$ C after removing all variables in B.

For general distributions, the pairwise independence assumption is weaker than the local independence assumption and the local independence assumption is weaker than the global independence assumption. However, for strictly positive distributions, these three independence assumptions are equivalent.

Proposition 1.6 *Given an undirected Markov graph* **G** *and a distribution P, if P satisfies the local independence assumption then P satisfies the pairwise independence assumption.*

$$P \models \mathcal{I}_l(G) \Rightarrow P \models \mathcal{I}_p(G)$$

Proposition 1.7 *Given an undirected Markov graph* **G** *and a distribution P, if P satisfies the global independence assumption then P satisfies the local independence assumption.*

$$P \models \mathcal{I}(G) \Rightarrow P \models \mathcal{I}_l(G)$$

Proposition 1.8 *Given an undirected Markov graph* **G** *and a strictly positive distribution P, if P satisfies the pairwise independence assumption then P satisfies the global independence assumption.*

$$P > 0 \textit{ and } P \models \mathcal{I}_p(G) \Rightarrow P \models \mathcal{I}(G)$$

Corollary 1.9 *Given an undirected Markov graph* **G** *and a strictly positive distribution P, if P satisfies one of the following independence assumptions, then P satisfies all of them, i.e. the following independence assumptions are equivalent .*

1. $P \models \mathcal{I}_l(G)$
2. $P \models \mathcal{I}_p(G)$
3. $P \models \mathcal{I}(G)$

1.3.3 Parameterization

Definition 1.10 (Boltzmann-Gibbs distribution) *A distribution P_Φ is a Gibbs distribution parameterized by a set of factors $\Phi = \{\phi_1(D_1), ..., \phi_K(D_K)\}$ if it is defined as the following:*

$$P_\Phi(X_1, ..., X_n) = \frac{1}{Z}\tilde{P_\Phi}(X_1, ..., X_n),$$

where

$$\tilde{P_\Phi}(X_1, ..., X_n) = \phi_1(\mathbf{D_1}) \times \phi_{\mathbf{2}}(\mathbf{D_2}) \times ... \times \phi_{\mathbf{m}}(\mathbf{D_m})$$

and

$$Z = \sum_{X_1,...,X_n} \tilde{P_\Phi}(X_1, ..., X_n)$$

is a normalizing constant known as the partition function. [KF09] Note that the factors does not correspond to conditional probabilities or probabilities themselves. The joint distribution is a product of all the individual factors and conversely, each factor merely acts as a contribution to the joint distribution.

Theorem 1.11 (HammersleyâClifford theorem) *A strictly positive probability distribution satisfies one of the Markov properties with respect to an undirected graph G if and only if it is a Gibbs distribution, that is, its density can be factorized over the cliques of the graph.*

Definition 1.12 (Markov Blanket) *A set* **U** *is a Markov blanket of X in a distribution P if $X \notin U$ and if U is a minimal set of nodes such that*

$$(X \perp \mathcal{X} - \{X\} - \mathbf{U} | \mathbf{U}) \in \mathcal{I}(\mathbf{P})$$

In undirected graphical models, the Markov blanket of a variable is precisely its neighbors in the graph. Figure 1.13 shows an example where the shaded nodes are the Markov blanket of node X.

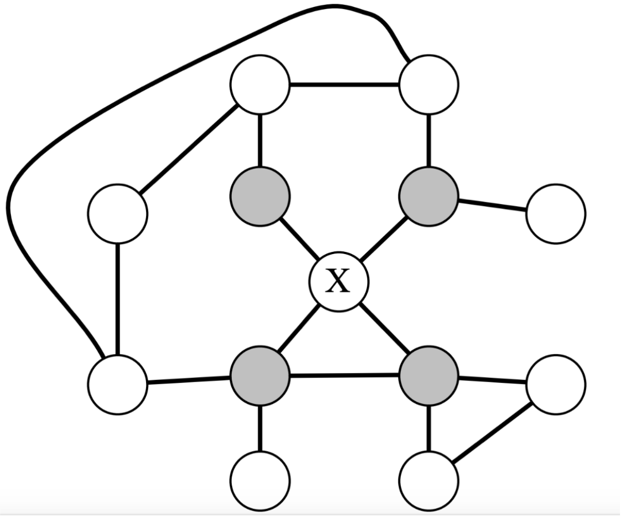


Figure 1.13: The Markov blanket of a MRF of node X is the shaded nodes

1.3.4 Example Markov Random Fields

Ising model was first introduced in statistical physics for systems of interacting atoms and it is one of the earliest types of Markov network models. Each node in the Ising model is a binary random variable $X_i \in \{+1, -1\}$ associated to an atom and its value is defined by the direction of the respective atomâs spin. [KF09]

The energy function is expressed over sub-cliques or edges of the graph:

$$p(X_1, ..., X_T) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$$

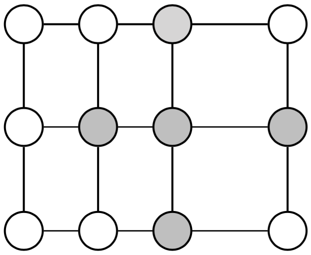
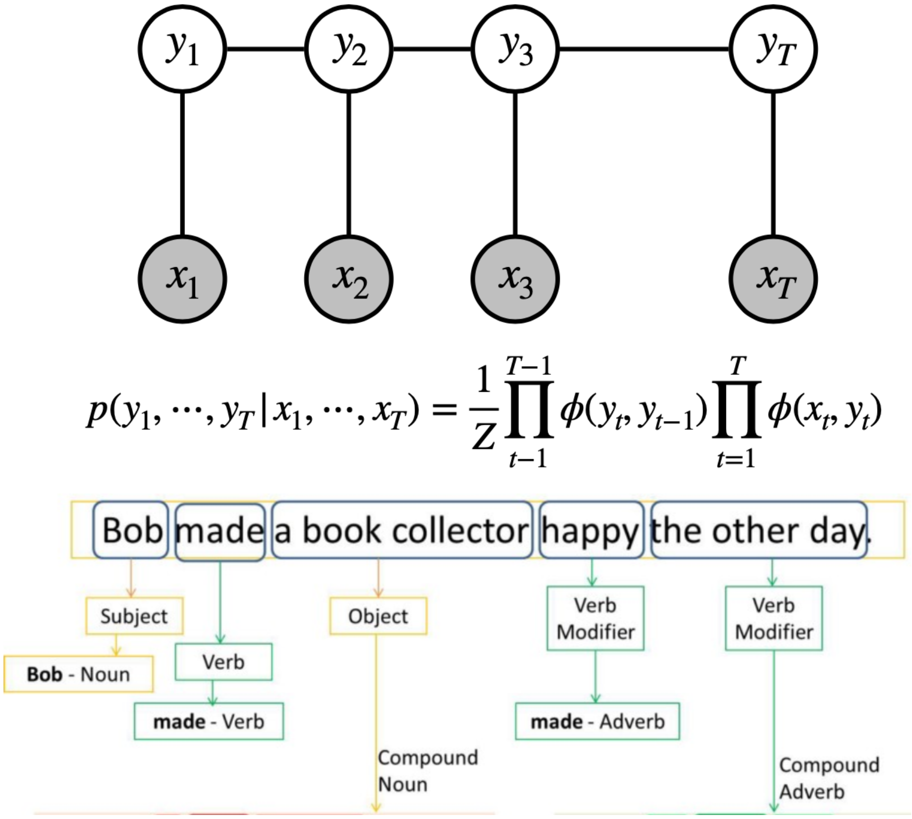


Figure 1.14: Ising Model

1.3.4.1 Conditional Random Fields

We have seen the application of using Markov networks to compactly represent a joint distribution over $\mathcal{X}$. The same Markov network framework could also be used to represent a conditional distribution. A conditional random field is a discriminative undirected graph model whose nodes correspond to the observed and latent variables $\mathbf{X} \cup \mathbf{Y}$ that represents the conditional distribution of $\mathbf{Y}|\mathbf{X}$. It is a form of Markov random filed that defines a posterior for variables Y's given observed variables X's as illustrated by the below diagram. The factorization gives a posterior distribution $P(Y|X)$ for the target variable Y given observed data X instead of a joint distribution P(Y,X).



$$p(y_1, \cdots, y_T | x_1, \cdots, x_T) = \frac{1}{Z} \prod_{t-1}^{T-1} \phi(y_t, y_{t-1}) \prod_{t=1}^{T} \phi(x_t, y_t)$$

Figure 1.15: Conditional Random Field

The conditional distribution factorization and the unconditional Gibbs distribution is only different by the normalization function Z(X). This difference is denoted graphically by having the feature variables grayed out. Note that the edges in the CRF does not explicitly encode the structure of any distribution over X and this flexibility allows us to incorporate into the model observed variables whose dependencies may be quite complex or even poorly understood. [KF09]

In the case of a simple latent variable model, the CRF is

$$p(Z|X) = \phi_{X,Z}(X, Z)$$

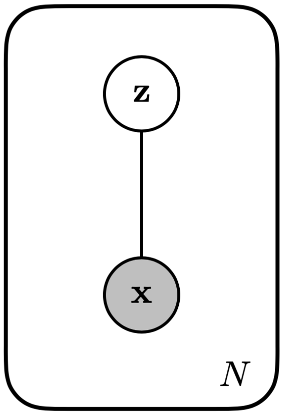


Figure 1.16: Latent Variable Model

References

[CW87] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Proceedings of the 19th ACM Symposium on Theory of Computing*, 1987, pp. 1–6.

[KF09] D. Koller and N. Friedman, "Probabilistic graphical models: principles and techniques," *MIT press*, 2009, pp. 102–156.