Probabilistic Graphical Models
Part II: Undirected Graphical Models

Selim Aksoy

Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

CS 551, Fall 2019

Introduction

- We looked at directed graphical models whose structure and parametrization provide a natural representation for many real-world problems.
- Undirected graphical models are useful where one cannot naturally ascribe a directionality to the interaction between the variables.

Introduction

- An example model that satisfies:
  - $(A \perp C | \{B, D\})$
  - $(B \perp D | \{A, C\})$
  - No other independencies
- These independencies cannot be naturally captured in a Bayesian network.

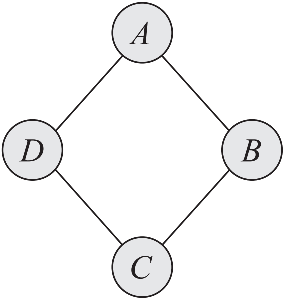

Figure 1: An example undirected graphical model.

An Example

- Four students are working together in pairs on a homework.
- Alice and Charles cannot stand each other, and Bob and Debbie had a relationship that ended badly.
- Only the following pairs meet: Alice and Bob; Bob and Charles; Charles and Debbie; and Debbie and Alice.
- The professor accidentally misspoke in the class, giving rise to a possible misconception.
- In study pairs, each student transmits her/his understanding of the problem.

An Example

- Four binary random variables are defined, representing whether the student has a misconception or not.
- Assume that for each $X \in \{A, B, C, D\}$, x^1 denotes the case where the student has the misconception, and x^0 denotes the case where she/he does not.
- Alice and Charles never speak to each other directly, so A and C are conditionally independent given B and D.
- Similarly, B and D are conditionally independent given A and C.

An Example

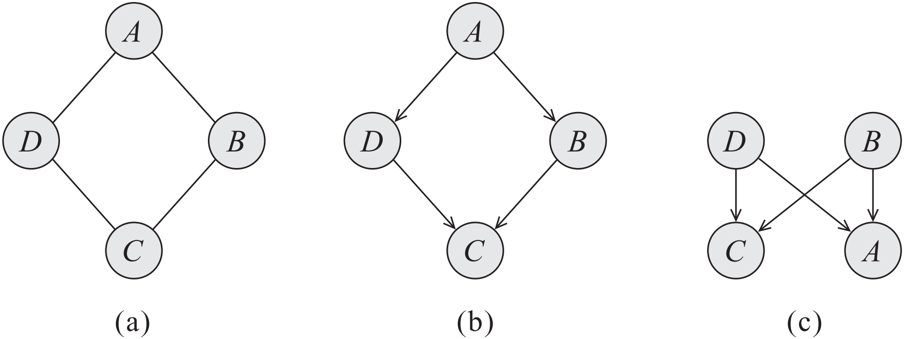

Figure 2: Example models for the misconception example. (a) An undirected graph modeling study pairs over four students. (b) An unsuccessful attempt to model the problem using a Bayesian network. (c) Another unsuccessful attempt.

Parametrization

- How to parametrize this undirected graph?
- We want to capture the affinities between related variables.
- Conditional probability distributions cannot be used because they are not symmetric.
- Marginals cannot be used because a product of marginals does not define a consistent joint.
- A general purpose function: *factor* (also called *potential*).

Parametrization

- Let $\mathbf{D}$ is a set of random variables.
  - A factor ϕ is a function from Val$(\mathbf{D})$ to $\mathbb{R}$.
  - A factor is nonnegative if all its entries are nonnegative.
  - The set of variables $\mathbf{D}$ is called the scope of the factor.
- In the example in Figure 2, an example factor is

$$\phi_1(A, B) : \text{Val}(A, B) \mapsto \mathbb{R}^+.$$

Parametrization

Table 1: Factors for the misconception example.

$\phi_1(A, B)$			$\phi_2(B, C)$			$\phi_3(C, D)$			$\phi_4(D, A)$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Parametrization

- The value associated with a particular assignment a, b denotes the affinity between these two variables: the higher the value $\phi_1(a, b)$, the more compatible these two values are.
- For ϕ_1, if A and B disagree, there is less weight.
- For ϕ_3 , if C and D disagree, there is more weight.
- A factor is not normalized, i.e., the entries are not necessarily in $[0, 1]$.

Parametrization

- The Markov network defines the local interactions between directly related variables.
- To define a global model, we need to combine these interactions.
- We combine the local models by multiplying them as

$$P(a, b, c, d) = \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(d, a).$$

Parametrization

- However, there is no guarantee that the result of this process is a normalized joint distribution.
- Thus, it is normalized as

$$P(a, b, c, d) = \frac{1}{Z} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a)$$

where

$$Z = \sum_{a,b,c,d} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a).$$

- Z is known as the *partition function*.

Parametrization

Table 2: Joint distribution for the misconception example.

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300,000	0.04
a^0	b^0	c^0	d^1	300,000	0.04
a^0	b^0	c^1	d^0	300,000	0.04
a^0	b^0	c^1	d^1	30	4.110^{-6}
a^0	b^1	c^0	d^0	500	6.910^{-5}
a^0	b^1	c^0	d^1	500	6.910^{-5}
a^0	b^1	c^1	d^0	5,000,000	0.69
a^0	b^1	c^1	d^1	500	6.910^{-5}
a^1	b^0	c^0	d^0	100	1.410^{-5}
a^1	b^0	c^0	d^1	1,000,000	0.14
a^1	b^0	c^1	d^0	100	1.410^{-5}
a^1	b^0	c^1	d^1	100	1.410^{-5}
a^1	b^1	c^0	d^0	10	1.410^{-6}
a^1	b^1	c^0	d^1	100,000	0.014
a^1	b^1	c^1	d^0	100,000	0.014
a^1	b^1	c^1	d^1	100,000	0.014

Parametrization

- Note that the factors are not marginals.
- In the misconception model, the marginal over A, B is

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

but the factor is

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

- A factor is only one contribution to the overall joint distribution.
- The distribution as a whole has to take into consideration the contributions from all of the factors involved.

Parametrization

- There is a tight connection between the factorization of the distribution and its independence properties.
- In undirected models, conditional independence is given by graph separation.
- $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ if there is no path from any node in $\mathbf{X}$ to any node in $\mathbf{Y}$ after removing all variables in $\mathbf{Z}$.
- In other words, all paths between nodes in $\mathbf{X}$ and $\mathbf{Y}$ pass through at least one of the nodes in $\mathbf{Z}$.
- Thus, the *Markov blanket* of a variable is its neighbors in the graph (i.e., a node is independent of the rest of the nodes in the graph given its immediate neighbors).

Parametrization

- Markov blanket tells that nodes that are not neighbors are conditionally independent given the remainder of the nodes.
- Factorization should be chosen in such a way that nodes that are not neighbors are not in the same factor.
- In other words, whatever factorization we pick, we know that only connected nodes can be arguments of a single local function.

Parametrization

- *Clique:* A subset of nodes such that there is an edge between all pairs of nodes (i.e., a fully connected subset).
- *Maximal clique:* A clique that cannot include any further node from the graph without ceasing to be a clique.
- We factorize an undirected graphical model as

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{X}_c),$$

  where $\mathcal{C}$ is the set of all maximal cliques in the graph, $\mathbf{X}_c$ is the set of nodes in clique c, $\phi_c(\mathbf{X}_c)$ is the potential function defined over $\mathbf{X}_c$, and Z is the partition function

$$Z = \sum_{X_1, \ldots, X_n} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{X}_c).$$

Hammersley-Clifford Theorem

- The Hammersley-Clifford theorem tells us that the family of distributions defined by the conditional independence semantics on the graph and the family defined by products of potential functions on cliques are the same.
- Tricky point: the potential functions are arbitrary real valued, but strictly positive.
- Notice the crucial difference between graphs, which tell us independencies that are true no matter what local functions we choose, and numerical functions which could introduce some extra independencies, once we know them.

Boltzmann Distributions

- Constraining clique potentials to be positive can be inconvenient.
- We often represent the clique potentials with their logs as

$$\phi_c(\mathbf{X}_c) = \exp\{-H_c(\mathbf{X}_c)\}$$

  by using arbitrary real valued energy functions $H_c(\mathbf{X}_c)$.
- This gives the joint a nice additive structure

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \exp\{-H_c(\mathbf{X}_c)\} = \frac{1}{Z} \exp\left\{-\sum_{c \in \mathcal{C}} H_c(\mathbf{X}_c)\right\}.$$

- This way of defining a probability distribution based on energies is the Boltzmann distribution from statistical physics.

Reduced Markov Networks

- If we observe some values, $\mathbf{U} = \mathbf{u}$, in the factor value table, we can eliminate the entries which are inconsistent with $\mathbf{U} = \mathbf{u}$.
- Let $\mathcal{H}$ be a Markov network over $\mathbf{X}$ and $\mathbf{U} = \mathbf{u}$ a context. The reduced Markov network $\mathcal{H}[\mathbf{u}]$ is a Markov network over the nodes $\mathbf{W} = \mathbf{X} - \mathbf{U}$, where we have an edge X—Y if there is an edge X—Y in $\mathcal{H}$.

Reduced Markov Networks

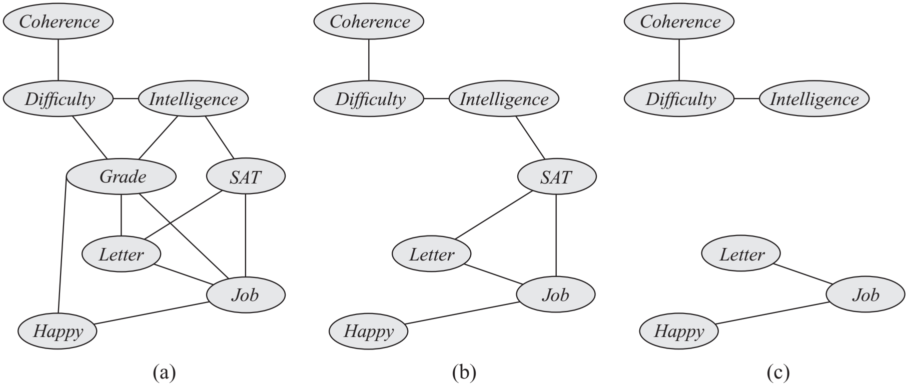

Figure 3: A reduced Markov network example. (a) Original set of factors. (b) Reduced to the context $G = g$. (c) Reduced to the context $G = g, S = s$.

Reduced Markov Networks

- Conditioning on a context $\mathbf{U}$ in Markov networks eliminates edges from the graph.
- In a Bayesian network, conditioning on evidence can create new dependencies.

Learning Undirected Models

- Like in Bayesian networks, once the joint distribution is generated, any kind of question can be answered using conditional probabilities and marginalization.
- However, a key distinction between Markov networks and Bayesian networks is normalization.
- Markov networks use a global normalization constant called the partition function.
- Bayesian networks involve local normalization within each conditional probability distribution.

Learning Undirected Models

- The global factor couples all of the parameters across the network, preventing us from decomposing the problem and estimating local groups of parameters separately.
- The global parameter coupling has significant computational ramifications.
- Even the simple maximum likelihood parameter estimation with complete data cannot be solved in closed form.

Learning Undirected Models

- We generally have to resort to iterative methods such as gradient ascent.
- The good news is that the likelihood objective is concave, so the methods are guaranteed to converge to the global optimum.
- The bad news is that each of the steps in the iterative algorithm requires that we run inference on the network, making even simple parameter estimation a fairly expensive process.