More about Undirected Graphical Models

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Outline

Factorization and Markov Properties on Undirected Graphs
- Relations between the Properties
- Verifying the Relations

Modeling with Undirected Graphs and Using Them in Practice
- Graph and Other Model Elements
- Gibbs Sampling
- An MRF/CRF Application Example

Notation

For an undirected graph $G = (V, E)$,

- $\mathcal{C}$: the set of cliques of G
- neighbors of v: $\mathcal{N}_v$
- For $A \subseteq V$, boundary of A: $\text{bd}(A) = \cup_{v \in A} \mathcal{N}_v \setminus A$, closure of A: $\text{cl}(A) = A \cup \text{bd}(A)$
- For $A, B, S \subseteq V$, $A \perp B \,|\, S$: S separates A from B in G
- $X = \{X_v, v \in V\}$: random variables associated with V
- X_A for $A \in V$: $\{X_v, v \in A\}$
- $X_A \perp X_B \,|\, X_S$: X_A and X_B are independent conditional on X_S

Illustration:

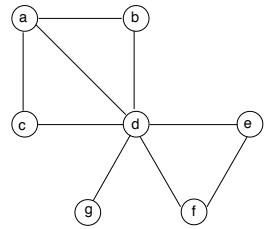
$$\mathcal{N}_a = \{b, c, d\}$$

$$\text{bd}(\{a, b\}) = \{c, d\}, \quad \text{cl}(\{a, b\}) = \{a, b, c, d\}$$

$$\mathcal{C} = \{\{a, b, d\}, \{a, c, d\}, \{d, e, f\}, \{d, g\}\}$$

$$\{a, b, c\} \perp \{e, f\} \,|\, \{d\}$$

G:



Markov Properties on Undirected Graphs

Fact: (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

(F) P *factorizes* according to G: $p(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C)$.

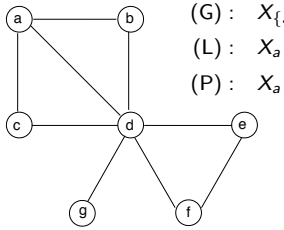
(G) For any disjoint subsets A, B, S of V, $A \perp B \,|\, S \quad \Rightarrow \quad X_A \perp X_B \,|\, X_S$.

(L) For all $v \in V$, $X_v \perp X_{V \setminus \mathrm{cl}(v)} \,|\, X_{\mathrm{bd}(v)}$.

(P) For all pairs of non-adjacent vertices (v, v') in G, $X_v \perp X_{v'} \,|\, V \setminus \{v, v'\}$.

Illustration:

G:



(F) : $p(x) = \phi_1(x_a, x_b, x_d)\phi_2(x_a, x_c, x_d)\phi_3(x_d, x_e, x_f)\phi_4(x_d, x_g)$
(This is the most general form of p that satisfies (F).)

(G) : $X_{\{a,b\}} \perp X_{\{e,g\}} \,|\, X_d$, $X_{\{c,g\}} \perp X_{\{b,f\}} \,|\, X_{\{a,d\}}$, etc.

(L) : $X_a \perp X_{\{e,f,g\}} \,|\, X_{\{b,c,d\}}$, $X_g \perp X_{\{a,b,c,e,f\}} \,|\, X_d$, etc.

(P) : $X_a \perp X_e \,|\, X_{\{b,c,d,e,f\}}$, $X_c \perp X_b \,|\, X_{\{a,d,e,f,g\}}$, etc.

Usefulness of the Graphical Representation

The fact "(F) $\Rightarrow$ (G)" is extremely useful:

- Read off conditional independence relations from the graph
- Create structures among variables to streamline computation
- Provide the machinery we need to study directed graphical models (Bayesian networks)

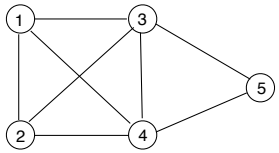
A few practices before we continue:

According to which graph, the following distribution factorizes?

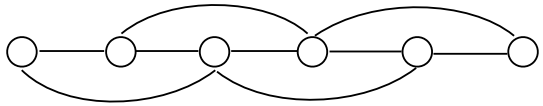
- $p(x_1, x_2, x_3, x_4, x_5, x_6) = \phi_1(x_1, x_2, x_3, x_4)\phi_2(x_3, x_4, x_5)$
- $p(x_1, x_2, x_3, x_4, x_5, x_6) = \phi_1(x_1, x_2, x_3)\phi_2(x_1, x_4)\phi_3(x_2, x_3, x_4)\phi_4(x_3, x_4, x_5)$
- The distribution of a 2nd-order Markov chain

Answers to Previous Questions

Graph G for the first two questions:



Graph G for a 2nd-order Markov chain:



Multivariate Gaussian Random Variables

The conclusion (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P) extends to rather general random variables. In particular, they hold for continuous random variables with positive and continuous densities, in which case (P) $\Rightarrow$ (F) also holds.

Consider a non-degenerate multivariate Gaussian random variable

$$X = (X_1, \ldots, X_n) \sim N(\mu, \Sigma).$$

Its density function is

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big\}.$$

So $f(x) \propto \exp\big\{ -\psi(x)\big\}$, where

$$\psi(x) = \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) = a + b'x + \frac{1}{2}\sum_{i,j} \Delta_{ij} x_i x_j, \quad \text{with} \quad \Delta = \Sigma^{-1},$$

and a, b being some constants. This shows:

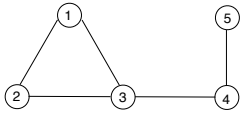
- The inverse covariance matrix Σ^{-1} reveals the graph G according to which $f(x)$ factorizes, and from which we can read off the conditional independence relations among $X_1, \ldots, X_n$. (The structure of Σ shows marginal independence relations.)

Multivariate Gaussian Random Variables

Example: $n = 5$,

$$\Sigma = \begin{pmatrix} 5 & -1 & -3 & 2 & -1 \\ -1 & 5 & -3 & 2 & -1 \\ -3 & -3 & 9 & -6 & 3 \\ 2 & 2 & -6 & 8 & -4 \\ -1 & -1 & 3 & -4 & 5 \end{pmatrix}, \quad \Sigma^{-1} = \frac{1}{3} \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}.$$

Graph G:



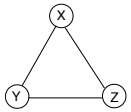
Remark:

- Recall that for estimating the parameter of a model, we can choose a parametrization suitable for the problem at hand. In graphical Gaussian modeling, some elements of Σ^{-1} are constrained to be zeros, so (μ, Σ^{-1}) is a more convenient parametrization than (μ, Σ).

Does G capture all independence relations?

Generally, G cannot.

- A simple counterexample:
  X, Y, Z are pairwise independent but not mutually independent.
  Then $P(X, Y, Z)$ factorizes only according to the fully connected graph



  So G does not capture the marginal independence.
- Directed graphical models/Bayesian networks (to be introduced in the next lecture) are more expressive from this perspective.

Outline

Verify (G) $\Rightarrow$ (L) $\Rightarrow$ (P)

(G) $\Rightarrow$ (L): evident, because $\{v\} \perp V \setminus \mathrm{cl}(\{v\}) \,|\, \mathrm{bd}(\{v\})$.

(L) $\Rightarrow$ (P), i.e., $X_v \perp X_{V \setminus \mathrm{cl}(\{v\})} \,|\, X_{\mathcal{N}_v} \quad \Rightarrow \quad X_v \perp X_{v'} \,|\, X_{V \setminus \{v, v'\}}$ for non-adjancent pairs (v, v'):

We use the fact

$$X \perp (Y, W) \,|\, Z \quad \Rightarrow \quad X \perp Y \,|\, (Z, W).$$

To see this,

- Intuitive argument: if given Z, knowing further the values of (Y, W) will not change our uncertainty about X, then given both Z and W, knowing further the value of Y will not change our uncertainty about X.
- Formal argument: as shown in the first exercise,

$$p(x, y, w, z) = a(x, z)b(y, w, z) \quad \Leftrightarrow \quad X \perp (Y, W) \,|\, Z$$
$$\Rightarrow \quad p(x, y, w, z) = c(x, w, z)d(y, w, z) \quad \Leftrightarrow \quad X \perp Y \,|\, (Z, W)$$

  where a, b, c, d are some functions.

"(L) $\Rightarrow$ (P)" then follows by taking

$$X = X_v, \quad Y = X_{v'}, \quad Z = X_{\mathcal{N}_v}, \quad W = X_{V \setminus \left(\{v'\} \cup \mathrm{cl}\left(\{v\}\right)\right)},$$

and noticing for a non-adjacent pair (v, v'),

$$V \setminus \Big(\{v'\} \cup \mathrm{cl}(\{v\})\Big) \cup \mathcal{N}_v = V \setminus \{v, v'\}, \quad \text{so} \quad (Z, W) = X_{V \setminus \{v, v'\}}.$$

Verify (F) ⇒ (G)

(F) ⇒ (G), i.e., $p(x) = \prod_{C\in\mathcal{C}} \phi_C(x_C) \quad \Rightarrow \quad X_A \perp X_B \,|\, X_S$, for all disjoint subsets $A, B, S \subseteq V$ such that $A \perp B \,|\, S$.

For any such subset A, B, S, we show that the marginal PMF of X_A, X_B, X_S satisfies

$$p(x_A, x_B, x_S) = a(x_A, x_S)b(x_B, x_S), \quad \text{for some functions } a, b,$$

which then implies $X_A \perp X_B \,|\, X_S$ (by the first exercise).

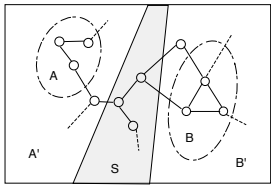
Let $A', B' \subseteq V$ be such that A', B', S are disjoint, and

$$A \subseteq A', \quad B \subseteq B', \quad A' \cup B' \cup S = V, \quad A' \perp B' \,|\, S.$$

We first show (F) implies $X_{A'} \perp X_{B'} \,|\, X_S$.

We examine which variables co-occur with $X_{A'}$ as the arguments of some function ϕ_C. Denote

$$\mathcal{C}_1 = \{C \in \mathcal{C} \mid A' \cap C \neq \emptyset\}.$$



Because $C \in \mathcal{C}$ is a complete subset,

$$i \in C, \text{ and } A' \cap C \neq \emptyset \quad \Rightarrow \quad i \in \mathrm{cl}(A') = A' \cup \mathrm{bd}(A'),$$

and because S separates A' from B',

$$\mathrm{bd}(A') \subseteq S, \quad \text{so} \quad C \subseteq A' \cup S, \quad \forall C \in \mathcal{C}_1.$$

Similarly,

$$C \subseteq B' \cup S, \quad \forall C \in \mathcal{C} \setminus \mathcal{C}_1.$$

Verify (F) ⇒ (G)

Therefore

$$p(x_{A'}, x_{B'}, x_S) = \prod_{C\in\mathcal{C}} \phi_C(x_C) = \Big(\prod_{C\in\mathcal{C}_1} \phi_C(x_C)\Big) \cdot \Big(\prod_{C\in\mathcal{C}\setminus\mathcal{C}_1} \phi_C(x_C)\Big)$$
$$= \psi_1(x_{A'}, x_S)\psi_2(x_{B'}, x_S)$$

for some functions ψ_1, ψ_2, (which implies $X_{A'} \perp X_{B'} \,|\, X_S$).

We then marginalize over $X_{A'\setminus A}$ and $X_{B'\setminus B}$ to obtain

$$p(x_A, x_B, x_S) = \sum_{x_{A'\setminus A}} \sum_{x_{B'\setminus B}} \psi_1(x_{A'}, x_S)\psi_2(x_{B'}, x_S)$$
$$= \Big(\sum_{x_{A'\setminus A}} \psi_1(X_{A'}, X_S)\Big)\Big(\sum_{x_{B'\setminus B}} \psi_2(x_{B'}, x_S)\Big)$$
$$= a(x_A, x_S)b(x_B, x_S)$$

for some functions a, b. This establishes "(F) ⇒ (G)."

Outline

Reading and Building the Graph

The graph shows conditional independence relations.

Besides, it also visualizes local dependence structures:

- The edges are naturally thought to represent direct interaction/association among the variables. Directions in the interaction are lost in the representation, so cause/effect relations cannot be modeled explicitly.
- Interactions are not restricted to be among pairs of variables. Indeed, local interaction is between a variable and its neighbors, with contributions from individual complete subsets (corresponding to terms appeared in the factorized p). This suggests that we think of a group of edges/nodes as a whole unit, when we interpret the graph in terms of interactions.

When building the graph of a model:

- We may start by specifying a factorized representation of P and obtain the associated graph. This is natural in the case where there is a certain global "energy" function we want to minimize.
- Or, more generally, we may start by considering properties (L) or (P) for our problem, obtain the associated graph, and hope (G), (L), (P) and (F) are all equivalent for the problem.

The graph is useful also for model checking: When property (G) is implied, the graph reveals other conditional independence assumptions implicit in the model. If some of them appear to be inappropriate, we can revise the model.

Elements of an Undirected Graphical Model

Typical elements:

- Graph G
- Form of distribution P: with $\mathcal{C}$ being the set of complete subsets of G,

$$p(x) \propto \prod_{C \in \mathcal{C}} \phi_C(x_c), \quad \text{or} \quad p(x) \propto \exp\Big\{ - \sum_{C \in \mathcal{C}} \phi_C(x_c) \Big\}.$$

- Function forms of ϕ_C and parameters in them

Maximum likelihood estimation is in general not easy, because, for example, when

$$p(x;\theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \phi_C(x_c;\theta),$$

$$L(\theta;x) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \phi_C(x_c;\theta), \quad \text{and} \quad \ell(\theta;x) = \sum_{C \in \mathcal{C}} \ln \phi_C(x_c;\theta) - \ln Z(\theta),$$

and the normalizing constant $Z(\theta)$ is a complicated, unknown function of θ, which makes the maximization of $\ell(\theta)$ difficult.

$Z(\theta)$: also called the *partition function*.

Outline

Gibbs Sampling

Goal: draw samples from an unknown distribution $P(X)$

E.g., $p(x)$ may be known only up to a normalizing constant,
or $p(x)$ may be known only through its local characteristics.

Use of the samples:

- understand the global behavior of the system
- find minimal energy configurations (when $p(x) \propto e^{-\psi(x)/T}$)
- approximate expected values

Gibbs sampling (Geman and Geman, 1984):

- We decompose x into d components:

$$x = (x_1, \ldots, x_d),$$

update each component while fixing the others,
and generate a sequence of $x^t = (x_1^t, \ldots x_d^t)$ with the goal that

$$P(X^t) \to P(X), \text{ and the frequency of } x \text{ in } \{x^t\} \to P(X = x), \text{ as } t \to \infty.$$

Gibbs Samplers

Gibbs sampling algorithm:

- Start with any initial $(x_1^0, \ldots, x_d^0)$.
- At iteration $t+1$, select a coordinate i,

$$\text{draw } \; X_i^{t+1} \sim P\big(X_i \,|\, X_{-i} = x_{-i}^t\big), \qquad X_{-i}^{t+1} = x_{-i}^t.$$

Two basic samplers:

- Random-scan Gibbs sampler:
  select i randomly according to a given distribution.
- Systematic-scan Gibbs sampler:
  select i according to a given order.

Notes:

- $\{X^t\}$ is a Markov chain on the space of x. Here the properties of long-term behavior of Markov chains are key to achieve the goal that

  $$P(X^t) \to P(X), \text{ and the frequency of } x \text{ in } \{x^t\} \to P(X = x), \text{ as } t \to \infty.$$

  (We talk about why in the future.)
- Intuitively, the components of the system interact with each other, and after sufficiently long time, the system is expected to be at "equilibrium."

Gibbs Sampling

Gibbs sampling is particularly appealing for MRF, because

- The full conditional distributions used in sampling reduces to the local characteristics

$$P(X_i \,|\, X_{-i} = x_{-i}^t) = P(X_i \,|\, X_{\mathcal{N}_i} = x_{\mathcal{N}_i}^t),$$

  and local characteristics are much simpler, of much lower dimension, and easy to normalize even if the normalizing constant is unknown.
- If the graph corresponds to a network and each node has a processor, sampling can be carried out in parallel and asynchronously throughout the network. Each processor only needs to do simple local computation. Thus large-scale problems can be handled.

Convergence can be slow when local interactions are strong, but not pushing towards the same direction.

Things can also go wrong, for example, when

- From some state x' not all states x with $p(x) > 0$ are reachable.
- The state space is infinite and the chain somehow drifts to infinity.
- We hypothesized some wrong local characteristics for which there exists no compatible joint distribution. (Existence and uniqueness results such as Besag and Hammersley-Clifford theorems are useful here to prevent pathologies.)

Outline

HMM for Sequence Labeling

Recall the HMM for parts-of-speech tagging example in Lec. 2:

pron	*v*	*adv*	*final punct.*
I	drove	home	.

- Possible tags:
  I: *n, pron*
  drove: *v, n*
  home: *n, adj, adv, v*

Two models for words $W = \{W_i\}$ and associated tags $T = \{T_i\}$:

$$p(w,t) = \prod_i p(w_i \,|\, t_i) p(t_i \,|\, t_{i-1});$$

$$p(w,t) = \prod_i p(w_i \,|\, t_i) p(t_i \,|\, t_{i-1}, t_{i-2}).$$

Question: What are the undirected graphs according to which the above $P(W,T)$ factorize?

Conditional Random Fields (CRF)

Suppose that $X = \{X_n\}$ and $Y = \{Y_n\}$ correspond to the latent and observable random variables, respectively, in a particular application.

In an MRF model: we specify the graph and the functions ϕ_C for (x, y).

Main features of CRF models:

(i) $P(X \,|\, Y)$ factorizes according to an undirected graph G', and

$$p(x \,|\, y) \propto \exp\{-\psi(x,y)\}, \quad \psi(x,y) = \sum_{C \in \mathcal{C}} \phi(x_c, y).$$

(ii) $P(Y)$ is not modeled.
We specify neither the probabilities nor the dependence among Y_is.

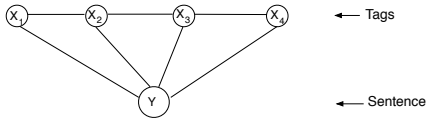
(iii) Maximum likelihood estimation: maximize $\prod_j P(X = x^j \,|\, Y = y^j; \theta)$ over Θ, based on complete data $\{(x^j, y^j)\}$.

Notes:

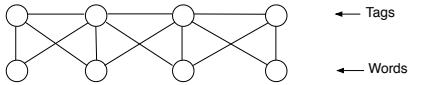
- Equivalently to (i), $P(X,Y)$ factorizes according to the graph which is G' added with the node Y and edges linking Y to the rest of the nodes.
- The main difference between CRF and a typical MRF model of (X,Y) is in (ii), namely, CRF models only the conditional distributions.

Three CRF Models for Sequence Labeling

(i)

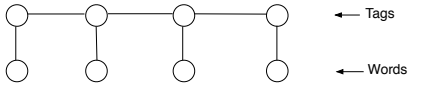


(ii)



Pairwise edges among the word variables are not drawn in the left.

(iii)



Questions:

(1) Suppose $p(x \mid y) \propto e^{-\psi(y,x)}$. What are the most general forms of ψ for the three models?

(2) Is model (iii) the same as an HMM? Do we obtain the same $P(X \mid Y)$, if we train the HMM also by maximizing $\prod_j P(X = x^j \mid Y = y^j; \theta)$ based on complete data $\{(x^j, y^j)\}$, like when training a CRF?

Further Readings

About MRF:

1. A. C. Davison. *Statistical Models*, Cambridge Univ. Press, 2003. Chap. 6.2.

For the next class, it would be good to take a look of

2. Robert G. Cowell et al. *Probabilistic Networks and Expert Systems*, Springer, 2007. Chap. 2.8, 2.9.