Inference and Representation

Rachel Hodos

New York University

Lab 4, September 23, 2015

Outline

Clarification on I-equivalent Bayesian Networks

- Theorem:

  If DAG's G and G' have the same V-structures and the same skeleton **then** $I(G) = I(G')$.

Clarification on I-equivalent Bayesian Networks

- Theorem:

  If DAG's G and G' have the same V-structures and the same skeleton **then** $I(G) = I(G')$.

- But the converse is not always true!

Clarification on I-equivalent Bayesian Networks

- Theorem:

  If DAG's G and G' have the same V-structures and the same skeleton **then** $I(G) = I(G')$.

- But the converse is not always true!
- Counterexample: two different, fully connected triplets

Clarification on I-equivalent Bayesian Networks

- Theorem:

  If DAG's G and G' have the same V-structures and the same skeleton **then** $I(G) = I(G')$.

- But the converse is not always true!
- Counterexample: two different, fully connected triplets
- Definition: *immorality* = v-structure where parents are not connected

Clarification on I-equivalent Bayesian Networks

- Theorem:

  If DAG's G and G' have the same V-structures and the same skeleton **then** $I(G) = I(G')$.

- But the converse is not always true!
- Counterexample: two different, fully connected triplets
- Definition: *immorality* = v-structure where parents are not connected
- Revised statement that is true in both directions:

  DAG's G and G' have the same **immoralities** and the same skeleton **iff** $I(G) = I(G')$.

Q: Why don't we have to worry about V-structures when we factorize a distribution?

- In order to go from a simple chain rule factorization:

$$P(X_1, \ldots, X_n) = \prod P(X_i | X_1, \ldots, X_{i-1})$$

to the canonical BN factorization,

$$P(X_1, \ldots, X_n) = \prod P(X_i | Pa(X_i))$$

we only use the following type of conditional independence:

$$X_i \perp X_{non-desc} | Pa(X_i).$$

- The conditional independence follows from d-separation.
- So, we never condition on children, and hence don't have to worry about V-structures.

Proof: Assume topological ordering... (Theorem 3.1 Koller & Friedman)

Let $\mathcal{G}$ be a BN structure over a set of random variables $\mathcal{X}$, and let P be a joint distribution over the same space. If $\mathcal{G}$ is an I-map for P, then P factorizes according to $\mathcal{G}$.

PROOF Assume, without loss of generality, that $X_1, \ldots, X_n$ is a *topological ordering* of the variables in $\mathcal{X}$ relative to $\mathcal{G}$ (see definition 2.19). As in our example, we first use the chain rule for probabilities:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}).$$

Now, consider one of the factors $P(X_i \mid X_1, \ldots, X_{i-1})$. As $\mathcal{G}$ is an I-map for P, we have that $(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}^{\mathcal{G}}) \in \mathcal{I}(P)$. By assumption, all of X_i's parents are in the set $X_1, \ldots, X_{i-1}$. Furthermore, none of X_i's descendants can possibly be in the set. Hence,

$$\{X_1, \ldots, X_{i-1}\} = \text{Pa}_{X_i} \cup Z$$

where $Z \subseteq \text{NonDescendants}_{X_i}$. From the local independencies for X_i and from the decomposition property (equation (2.8)) it follows that $(X_i \perp Z \mid \text{Pa}_{X_i})$. Hence, we have that

$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid \text{Pa}_{X_i}).$$

Applying this transformation to all of the factors in the chain rule decomposition, the result follows. ■

Q: Is there an algorithm to construct all possible graphs for a given set of independence statements?

No, but there is an algorithm to construct *a* minimal I-map given some I(p) *and* some variable ordering:

Algorithm 3.2 Procedure to build a minimal I-map given an ordering

Procedure Build-Minimal-I-Map (
  $X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
  $\mathcal{I}$ // Set of independencies
)
1 Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
2 **for** $i = 1, \ldots, n$
3 $U \leftarrow \{X_1, \ldots, X_{i-1}\}$ // U is the current candidate for parents of X_i
4 **for** $U' \subseteq \{X_1, \ldots, X_{i-1}\}$
5 **if** $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in \mathcal{I}$ **then**
6 $U \leftarrow U'$
7 // At this stage U is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
8 // Now set U to be the parents of X_i
9 **for** $X_j \in U$
10 Add $X_j \rightarrow X_i$ to $\mathcal{G}$
11 return $\mathcal{G}$

Markov Random Fields (undirected graphical models)

- Rather than CPDs, we specify (non-negative) **potential functions** over sets of variables associated with cliques C of the graph,

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(\mathbf{x}_c)$$

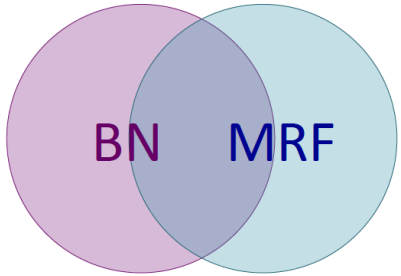
- Z is the **partition function** and normalizes the distribution:

$$Z = \sum_{\hat{x}_1, \ldots, \hat{x}_n} \prod_{c \in C} \phi_c(\hat{\mathbf{x}}_c)$$

- Like CPD's, $\phi_c(\mathbf{x}_c)$ can be represented as a table, but it is *not normalized*
- Called *undirected graphical models, Markov random fields* (MRFs), or *Markov networks*
- Independence given simply by graph separation

Comparing BNs to MRFs

- There are some $I(p)$'s that can be represented by MRFs but not BNs, and vice versa. (Examples are v-structure, and four friends' hair color from yesterday).



- Advantage of MRFs: marginalization and inference are *local* operations
- Disadvantage: hard to compute the partition function (sum over all possible states), often resort to approximations
- Disadvantage: no longer a natural way to *sample* data

Remarks on MRFs

- Cliques are *not* the same thing as CPD's or marginals
- However, setting a clique potential to 0 for a particular state will result in probability being equal to 0
- Edges are undirected but cliques potentials *do not* have to be symmetric
- Maximal cliques provide sufficient parametrization, so why not only use maximal cliques?

Remarks on MRFs

- Cliques are *not* the same thing as CPD's or marginals
- However, setting a clique potential to 0 for a particular state will result in probability being equal to 0
- Edges are undirected but cliques potentials *do not* have to be symmetric
- Maximal cliques provide sufficient parametrization, so why not only use maximal cliques?
- One reason: may want to use sub-cliques to decrease number of parameters

Motivation for conditional random fields

- Suppose **Y** is a set of variables that we want to estimate (e.g. class labels)
- Suppose **X** is a set of variables that are always observed, i.e., we have empirical distribution $P(\mathbf{X})$.

Motivation for conditional random fields

- Suppose $\mathbf{Y}$ is a set of variables that we want to estimate (e.g. class labels)
- Suppose $\mathbf{X}$ is a set of variables that are always observed, i.e., we have empirical distribution $P(\mathbf{X})$.
- We could model the full joint distribution $P(\mathbf{X}, \mathbf{Y})$ as $P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})$. But can be difficult to model $P(\mathbf{Y})$, e.g. what is the distribution of labels of natural images?

Motivation for conditional random fields

- Suppose $\mathbf{Y}$ is a set of variables that we want to estimate (e.g. class labels)
- Suppose $\mathbf{X}$ is a set of variables that are always observed, i.e., we have empirical distribution $P(\mathbf{X})$.
- We could model the full joint distribution $P(\mathbf{X}, \mathbf{Y})$ as $P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})$. But can be difficult to model $P(\mathbf{Y})$, e.g. what is the distribution of labels of natural images?
- But, the joint distribution can equivalently be factored as $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{Y}|\mathbf{X})P(\mathbf{X})$. Now we only need $P(\mathbf{Y}|\mathbf{X})$.

Conditional random fields (CRFs)

- **Conditional random fields** are undirected graphical models of conditional distributions $p(\mathbf{Y} \mid \mathbf{X})$
- We typically show the graphical model using just the **Y** variables
- Potentials are a function of **X** and **Y**
- Can still use all the tools we've learned so far to model this joint distribution over **Y**

Formal definition

- A CRF is a Markov network on variables $\mathbf{X} \cup \mathbf{Y}$, which specifies the conditional distribution

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \phi_c(\mathbf{x}_c, \mathbf{y}_c)$$

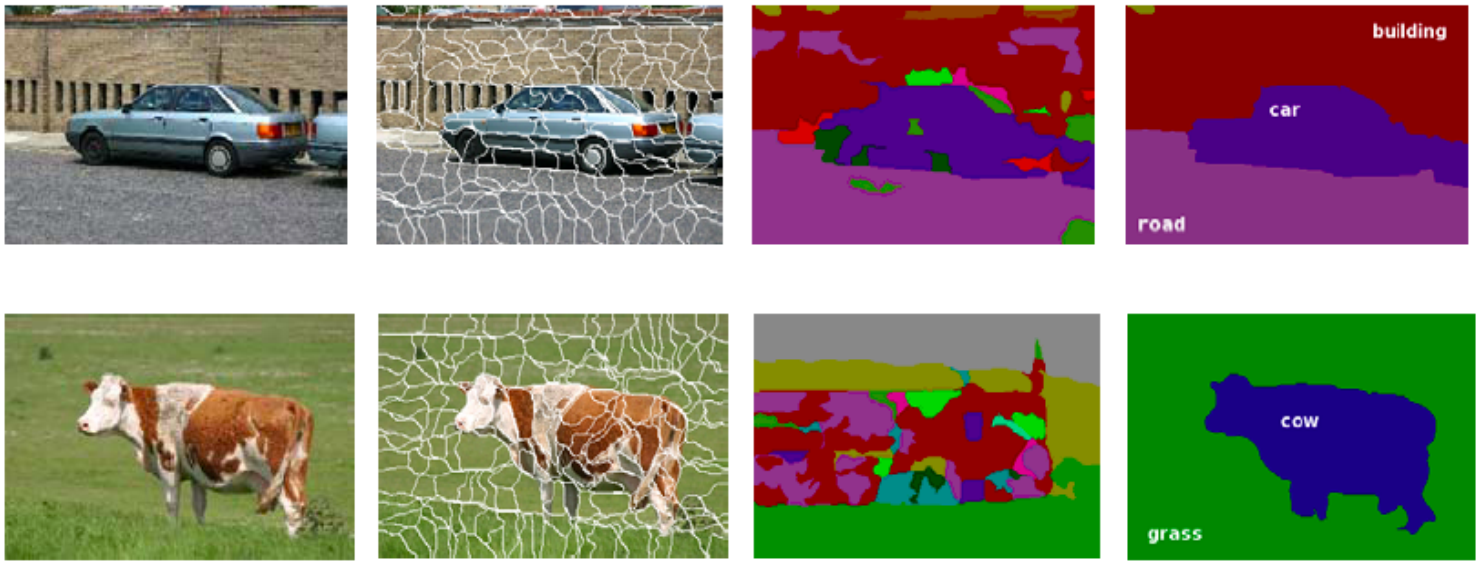
  with partition function

$$Z(\mathbf{x}) = \sum_{\hat{\mathbf{y}}} \prod_{c \in C} \phi_c(\mathbf{x}_c, \hat{\mathbf{y}}_c).$$

- As before, two variables in the graph are connected with an undirected edge if they appear together in the scope of some factor
- The only difference with a standard Markov network is the normalization term – before marginalized over $\mathbf{X}$ and $\mathbf{Y}$, now only over $\mathbf{Y}$

CRFs in computer vision

- Example applications: segmentation, stereo, de-noising
- Grids are particularly popular, e.g., pixels in an image with 4-connectivity



- How would you define the clique potentials for a given image **X** in order to perform image segmentation?

Parameterization of CRFs

- Factors may depend on a large number of variables
- We typically parameterize each factor as a log-linear function,

$$\phi_c(\mathbf{x}_c, \mathbf{y}_c) = \exp\{\mathbf{w} \cdot \mathbf{f}_c(\mathbf{x}_c, \mathbf{y}_c)\}$$

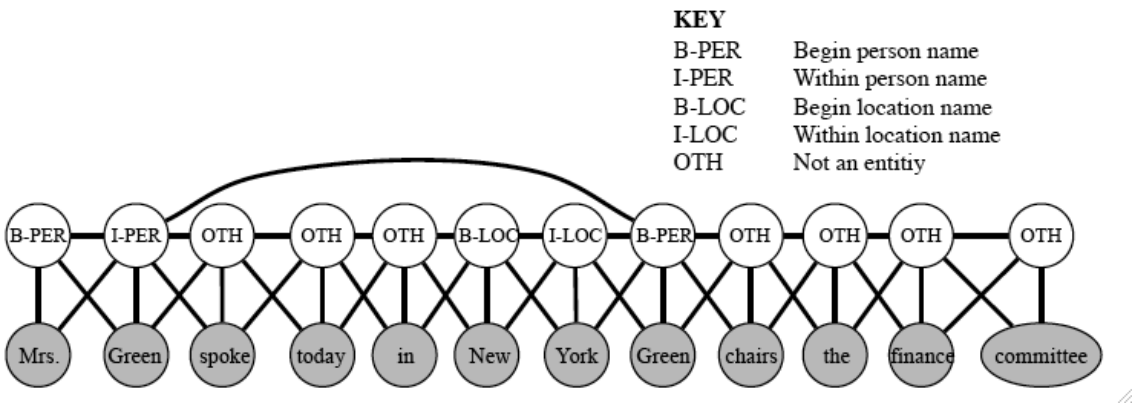
- $\mathbf{f}_c(\mathbf{x}_c, \mathbf{y}_c)$ is a feature vector
- $\mathbf{w}$ is a weight vector which is typically learned – we will discuss this extensively in later lectures

NLP example: named-entity recognition

- Given a sentence, determine the people and organizations involved and the relevant locations:
  "Mrs. Green spoke today in New York. Green chairs the finance committee."
- Entities sometimes span multiple words. Entity of a word not obvious without considering its *context*
- CRF has one variable X_i for each word, which encodes the possible labels of that word
- The labels are, for example, "B-person, I-person, B-location, I-location, B-organization, I-organization"
  - Having beginning (B) and within (I) allows the model to segment adjacent entities

NLP example: named-entity recognition

The graphical model looks like (called a *skip-chain CRF*):



There are three types of potentials:

- $\phi^1(Y_t, Y_{t+1})$ represents dependencies between neighboring target variables [analogous to transition distribution in a HMM]
- $\phi^2(Y_t, Y_{t'})$ for all pairs t, t' such that $x_t = x_{t'}$, because if a word appears twice, it is likely to be the same entity
- $\phi^3(Y_t, X_1, \cdots, X_T)$ for dependencies between an entity and the word sequence [e.g., may have features taking into consideration capitalization]