

CS 3750 Machine Learning
Lecture 6

**Markov Random Fields IV:
Learning**

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

CS 3750 Advanced Machine Learning

Markov random fields

- **Probabilistic models with symmetric dependences.**

- Typically models spatially varying quantities

$$P(x) \propto \prod_{c \in cl(x)} \phi_c(x_c)$$

$\phi_c(x_c)$ - A potential function (defined over factors)

- If $\phi_c(x_c)$ is strictly positive we can rewrite the definition as:

$$P(x) = \frac{1}{Z} \exp \left(- \sum_{c \in cl(x)} E_c(x_c) \right) \quad \text{- Energy function}$$

- Gibbs (Boltzman) distribution

$$Z = \sum_{x \in \{x\}} \exp \left(- \sum_{c \in cl(x)} E_c(x_c) \right) \quad \text{- A partition function}$$

CS 3750 Advanced Machine Learning

Types of Markov random fields

- **MRFs with discrete random variables**

- Clique potentials can be defined by mapping all clique-variable instances to \mathbb{R}
- Example: Assume two binary variables A,B with values $\{a1,a2,a3\}$ and $\{b1,b2\}$ are in the same clique c. Then:

$$\phi_c(A, B) \cong$$

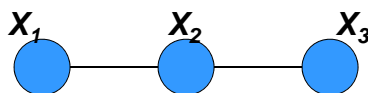
| | | |
|----|----|-----|
| a1 | b1 | 0.5 |
| a1 | b2 | 0.2 |
| a2 | b1 | 0.1 |
| a2 | b2 | 0.3 |
| a3 | b1 | 0.2 |
| a3 | b2 | 0.4 |

- Next: **Learning MRFs with discrete random vars**

CS 3750 Advanced Machine Learning

An example of MRF

- Undirected Graph



- Full joint distribution

$$p(X) = \frac{1}{Z} \psi_1(X_1, X_2) \cdot \psi_2(X_2, X_3) \cdot$$

- Parameters

$$\begin{aligned}
 &\psi_1(X_1 = 0, X_2 = 0), \psi_1(X_1 = 0, X_2 = 1), \\
 &\psi_1(X_1 = 1, X_2 = 0), \psi_1(X_1 = 1, X_2 = 1), \\
 &\psi_2(X_2 = 0, X_3 = 0), \psi_2(X_2 = 0, X_3 = 1), \\
 &\psi_2(X_2 = 1, X_3 = 0), \psi_2(X_2 = 1, X_3 = 1).
 \end{aligned}$$

Assumptions

- Complete data set
 - No hidden variables, no missing value
 - Independent identically distribution (IID)
 - Discrete model
 - Known structure
 - Parameter independency
 - Maximum likelihood estimation
 - More difficult than that of Bayesian network
 - Decomposable or non-decomposable model
-

Notations

- V : set of nodes of the graph.
 - X_u : the random variable associated with $u \in V$
 x_u : an instantiation of X_u
 - C : a subset of V ,
 X_C : set of variables indexed by C
 x_c : an instantiation of X_C
 x_V or x : an instantiation of all random variables
 - N : number of samples in the data set D
 n : Index of data. $n = 1, 2 \dots N$
 - $D : (D_1, D_2, \dots, D_N) = (x_{v,1}, x_{v,2}, \dots, x_{v,N})$
-

Maximum likelihood estimation for MRF

- Full joint distribution

$$p(x_V | \theta) = \frac{1}{Z} \prod_C \psi_C(x_C), \quad Z = \sum_{x_C} \prod_C \psi_C(x_C)$$

- Likelihood

$$p(D_n | \theta) = p(x_{V,n} | \theta) = \prod_{x_V} p(x_V | \theta)^{\delta(x_V, x_{V,n})}$$

$$\delta(x_V, x_{V,n}) = 1 \text{ iff } x_V = x_{V,n}$$

$$p(D | \theta) = \prod_n p(x_{V,n} | \theta) = \prod_n \prod_{x_V} p(x_V | \theta)^{\delta(x_V, x_{V,n})}$$

Maximum likelihood estimation for MRF

- Log likelihood

$$\begin{aligned} l(\theta, D) &= \log p(D | \theta) = \log \left(\prod_n \prod_{x_V} p(x_V | \theta)^{\delta(x_V, x_{V,n})} \right) \\ &= \sum_n \sum_{x_V} \delta(x_V, x_{V,n}) \log p(x_V | \theta) = \sum_{x_V} m(x_V) \log p(x_V | \theta) \end{aligned}$$

- Count: the number of times that configuration x_V is observed is defined as:

$$m(x_V) \equiv \sum_n \delta(x_V, x_{V,n})$$

- And marginal count for clique C :

$$m(x_C) \equiv \sum_{x_V \setminus C} m(x_V)$$

Count and Marginal Count

| X_1 | X_2 | X_3 |
|-------|-------|-------|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |

$$m((X_1=0, X_2=0, X_3=1)) = ?$$

$$m((X_1=1, X_2=0)) = ?$$

Count and Marginal Count

| X_1 | X_2 | X_3 |
|-------|-------|-------|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |

$$m((X_1=0, X_2=0, X_3=1)) = 3$$

$$m((X_1=1, X_2=0)) = ?$$

Count and Marginal Count

| X_1 | X_2 | X_3 |
|-------|-------|-------|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |

$$m((X_1=0, X_2=0, X_3=1)) = 3$$

$$m((X_1=1, X_2=0)) = 3$$

Maximum likelihood estimation for MRF

- Log likelihood

$$\begin{aligned}
 l(\theta, D) &= \sum_n \sum_{x_V} \delta(x_V, x_{V,n}) \log p(x_V | \theta) \\
 &= \sum_{x_V} m(x_V) \log p(x_V | \theta) \\
 &= \sum_{x_V} m(x_V) \log \left(\frac{1}{Z} \prod_C \psi_C(x_C) \right) \\
 &= \sum_{x_V} m(x_V) \sum_C \log \psi_C(x_C) - \sum_{x_V} m(x_V) \log Z \\
 &= \sum_C \sum_{x_C} m(x_C) \log \psi_C(x_C) - N \log Z
 \end{aligned}$$

Bayesian network vs MRF

- Bayesian network

Parameters are decomposed

$$l(\theta, D) = \sum_u \sum_{x_{\{u\} \cup pa(u)}} m(x_{\{u\} \cup pa(u)}) \log \theta_u(x_{\{u\} \cup pa(u)})$$

- MRF

Parameters are **not** decomposed

$$l(\theta, D) = \sum_C \sum_{x_C} m(x_C) \log \psi_C(x_C) - N \log Z$$

$$\log Z = \log \sum_{x_C} \prod_C \psi_C(x_C)$$

Maximum likelihood estimation for MRF

- The derivative of the normalization factor Z

$$\begin{aligned} \frac{\partial \log Z}{\partial \psi_C(x_C)} &= \frac{1}{Z} \frac{\partial}{\partial \psi_C(x_C)} \left(\sum_{\tilde{x}} \prod_D \psi_D(\tilde{x}_D) \right) \\ &= \frac{1}{Z} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \frac{\partial}{\partial \psi_C(x_C)} \left(\prod_D \psi_D(\tilde{x}_D) \right) \\ &= \frac{1}{Z} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \prod_{D \neq C} \psi_D(\tilde{x}_D) \\ &= \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \frac{1}{\psi_C(\tilde{x}_C)} \frac{1}{Z} \prod_D \psi_D(\tilde{x}_D) \\ &= \frac{1}{\psi_C(x_C)} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) p(\tilde{x}) = \frac{p(x_C)}{\psi_C(x_C)} \end{aligned}$$

Maximum likelihood estimation for MRF

- The derivative of the log likelihood

$$\frac{\partial l(\theta, D)}{\partial \psi_C(x_C)} = \frac{m(x_C)}{\psi_C(x_C)} - N \frac{p(x_C)}{\psi_C(x_C)} = 0$$

- Assume $\tilde{p}(x_C) = \frac{1}{N} m(x_C)$ is the empirical marginal
- Then:

$$\frac{\tilde{p}(x_C)}{\psi_C(x_C)} = \frac{p(x_C)}{\psi_C(x_C)}$$

and

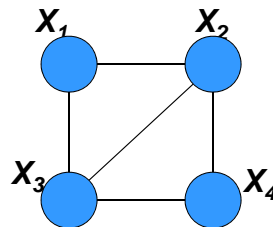
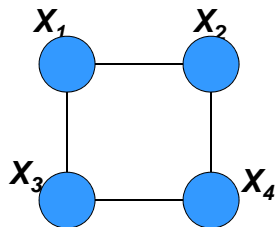
$$\hat{p}_{ML}(x_C) = \tilde{p}(x_C)$$

- An important property of MLE of MRF**

– For each clique C , the *model marginals* $\hat{p}_{ML}(x_C)$
must be equal to the *empirical marginals* $\tilde{p}(x_C)$

Decomposable models

- MRF is decomposable if its underlying graph is chordal

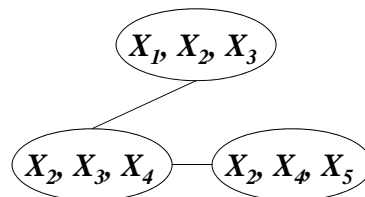
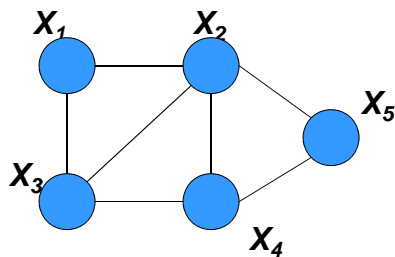


MLE of Decomposable models

- For every clique C , set the clique potential to the empirical marginal for that clique
- For every non-empty intersection between cliques, associate an empirical with that intersection, and divide that empirical marginal into the potential of one of the two cliques that form the intersection.

MLE for decomposable models: example

$$\left. \begin{aligned} \hat{\psi}_{123,ML}(x_1, x_2, x_3) &= \tilde{p}(x_1, x_2, x_3); \\ \hat{\psi}_{234,ML}(x_2, x_3, x_4) &= \frac{\tilde{p}(x_2, x_3, x_4)}{\tilde{p}(x_2, x_3)}; \\ \hat{\psi}_{234,ML}(x_2, x_4, x_5) &= \frac{\tilde{p}(x_2, x_4, x_5)}{\tilde{p}(x_2, x_4)}. \end{aligned} \right\} \Rightarrow Z = 1$$



MLE for decomposable models: example

- MLE of full joint probability

$$\hat{p}_{ML}(x) = \frac{\prod_c \tilde{p}(x_c)}{\prod_s \tilde{p}(x_s)}$$

MLE for non-decomposable models

- We want to find the $\Psi_C(x_C)$

- In MLE we want to satisfy

$$\frac{\tilde{p}(x_C)}{\Psi_C(x_C)} = \frac{p(x_C)}{\Psi_C(x_C)}$$

- Finding $\Psi_C(x_C)$ is not straightforward: it is on both sides

- **Iterative solution:**

$$\psi_C^{(t+1)}(x_C) = \psi_C^{(t)}(x_C) \frac{\tilde{p}(x_C)}{p^{(t)}(x_C)}$$

Iterative proportional fitting (IPF)

IPF update equation (coordinate ascent)

Cycle through all cliques C and keep updating potentials to make the empirical and modeled marginal distributions on the cliques the same using:

$$\psi_C^{(t+1)}(x_C) = \psi_C^{(t)}(x_C) \frac{\tilde{p}(x_C)}{p^{(t)}(x_C)}$$

Properties of IPF

- It works for both decomposable and non-decomposable models
 - It is guaranteed to converge (to a local optima)
 - Log-likelihood is guaranteed to increase or remain the same after the update
-

Two properties of the update equation

- The marginal of updated clique C is equal to its empirical marginal

$$p^{(t+1)}(x_C) = \tilde{p}(x_C)$$

- From the update equation, we can get:

$$p^{(t+1)}(x_C) = \frac{Z^{(t)}}{Z^{(t+1)}} \tilde{p}(x_C)$$

- The normalization factor Z remains constant

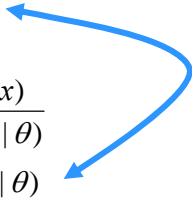
$$Z^{(t+1)} = Z^{(t)}$$

$$\Rightarrow p^{(t+1)}(x_C) = p^{(t)}(x_C) \frac{\tilde{p}(x_C)}{p^{(t)}(x_C)}$$

The relationship between MLE and KL divergence

- MLE

$$\begin{aligned}
 l(\theta, D) &= \sum_n \sum_{x_V} \delta(x_V, x_{V,n}) \log p(x_V | \theta) \\
 &= \sum_{x_V} m(x_V) \log p(x_V | \theta) \\
 &= N \sum_{x_V} \tilde{p}(x_V) \log p(x_V | \theta)
 \end{aligned}$$
 - KL divergence

$$\begin{aligned}
 D(\tilde{p}(x) \| p(x | \theta)) &= \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x | \theta)} \\
 &= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x | \theta)
 \end{aligned}$$
 - Maximizing the likelihood is equivalent to minimizing the KL divergence
- 

Iterative proportional fitting (IPF)

- **IPF update equation (coordinate ascent)**

$$\psi_C^{(t+1)}(x_C) = \psi_C^{(t)}(x_C) \frac{\tilde{p}(x_C)}{p^{(t)}(x_C)}$$
- Updates potentials to make the empirical and modeled marginal distributions the same
- **How hard is it to compute the update?**
 - Empirical clique marginals are typically easy
 - Model clique marginals require inference:
 - Jirousek & Preucil 1995 - More efficient IPF implementation using the junction tree

Gradient ascent

- Alternative to IPF
- Update equation

$$\psi_c^{(t+1)}(x_c) = \psi_c^{(t)}(x_c) + \frac{\lambda}{\psi_c^{(t)}(x_c)} (\tilde{p}(x_c) - p^{(t)}(x_c))$$

- Advantage
 - All parameters can be adjusted simultaneously
 - Disadvantage
 - Have to choose appropriate λ
 - Recalculate Z after each iteration.
-