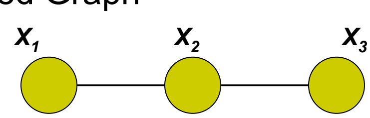
Parameter Estimation of Markov Random Field

Chenhai Xi
chenhai@cs.pitt.edu

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

$$\psi_1(X_1 = 0, X_2 = 0), \psi_1(X_1 = 0, X_2 = 1),$$
$$\psi_1(X_1 = 1, X_2 = 0), \psi_1(X_1 = 1, X_2 = 1),$$
$$\psi_2(X_2 = 0, X_3 = 0), \psi_2(X_2 = 0, X_3 = 1),$$
$$\psi_2(X_2 = 1, X_3 = 0), \psi_2(X_2 = 1, X_3 = 1).$$

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

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

- Likelihood

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

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

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

Maximum likelihood estimation for MRF

- Log likelihood

$$l(\theta, D) = \log p(D \mid \theta) = \log\left(\prod_n \prod_{x_V} p(x_V \mid \theta)^{\delta(x_V, x_{V,n})}\right)$$

$$= \sum_n \sum_{x_V} \delta(x_V, x_{V,n}) \log p(x_v \mid \theta) = \sum_{x_V} m(x_V) \log p(x_V \mid \theta)$$

- 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

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

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

Maximum likelihood estimation for MRF

- Log likelihood

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

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

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

Maximum likelihood estimation for MRF

- The derivative of normalization factor *Z*

$$\frac{\partial \log Z}{\partial \psi_C(x_C)} = \frac{1}{Z} \frac{\partial}{\partial \psi_C(x_C)} \left(\sum_{\tilde{x}} \prod_D \psi_D(\tilde{x}_D) \right)$$

$$= \frac{1}{Z} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \frac{\partial}{\partial \psi_C(x_c)} \left(\prod_D \psi_D(\tilde{x}_D) \right)$$

$$= \frac{1}{Z} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \prod_{D \neq C} \psi_D(\tilde{x}_D)$$

$$= \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) \frac{1}{\psi_C(\tilde{x}_c)} \frac{1}{Z} \prod_D \psi_D(\tilde{x}_D)$$

$$= \frac{1}{\psi_C(x_c)} \sum_{\tilde{x}} \delta(\tilde{x}_C, x_C) p(\tilde{x}) = \frac{p(x_C)}{\psi_C(x_C)}$$

Maximum likelihood estimation for MRF

- The derivative of the log likelihood

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

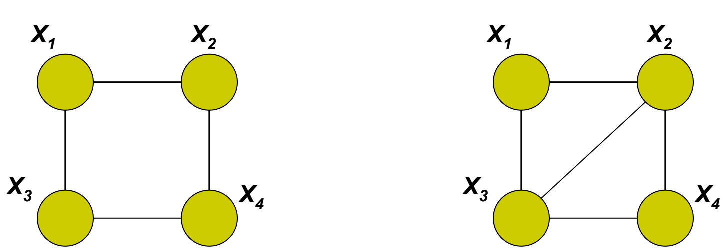
- Set it to zero, we obtain:

$$\hat{p}_{ML}(x_C) = \frac{1}{N} m(x_C) = \widetilde{p}(x_C)$$

- 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* $\widetilde{p}(x_C)$

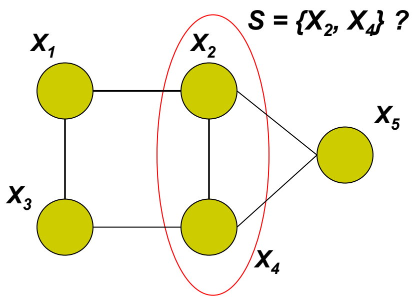
Decomposable models

- Graph *G* is *decomposable* **iff** it can be recursively subdivided into disjoint sets *A*, *B* and *S*, where *S* separates *A* and *B*, and where *S* is complete. The union of *A* and *S* and the union of *B* and *S* are also decomposable



Decomposable models

- Decomposable ⇔triangulated

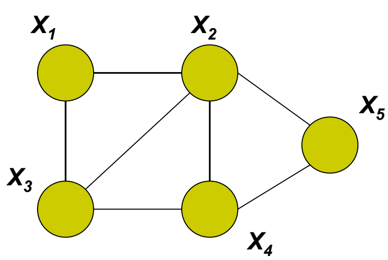


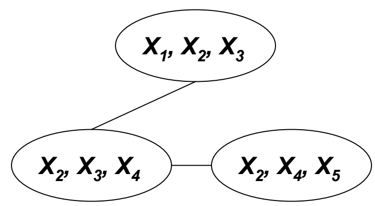
MLE of Decomposable models

- 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.

An example

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





- Could we set ?

$$\hat{\psi}_{123,ML}(x_1,x_2,x_3) = \tilde{p}(x_1,x_2,x_3);$$
$$\hat{\psi}_{234,ML}(x_2,x_3,x_4) = \tilde{p}(x_2,x_3,x_4);$$
$$\hat{\psi}_{345,ML}(x_2,x_4,x_5) = \tilde{p}(x_2,x_4,x_5).$$

- MLE of full joint probability

$$\hat{p}_{ML}(x) = \frac{\prod_C \tilde{p}(x_C)}{\prod_S \tilde{p}(x_S)}$$

Iterative proportional fitting (IPF)

- Properties of IPF
  - It works for both decomposable and non-decomposable models
  - It is guaranteed to converge
  - Log-likelihood is guaranteed to increase or remain the same after
- IPF update equation (coordinate ascent)

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

Two properties of the update equation

- From the update equation, we can get:

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

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

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

- The normalization factor *Z* remains constant

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

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

The relationship between MLE and KL divergence

- MLE

$$l(\theta, D) = \sum_n \sum_{x_V} \delta(x_V, x_{V,n}) \log p(x_v \mid \theta)$$

$$= \sum_{x_V} m(x_v) \log p(x_V \mid \theta)$$

$$= N \sum_{x_V} \tilde{p}(x_v) \log p(x_V \mid \theta)$$

- KL divergence

$$D(\tilde{p}(x) \,\|\, p(x \mid \theta)) = \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x \mid \theta)}$$

$$= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x \mid \theta)$$

- Maximizing the likelihood is equivalent to minimizing the KL divergence

Gradient ascent

- 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.

Exponential family model

- Exponential family model

$$p(x\mid\theta)=\frac{1}{Z}\exp\left\{\sum_{\mathrm{i}}\theta_i f_i(x)\right\},\quad Z=\sum_x \exp\left\{\sum_{\mathrm{i}}\theta_i f_i(x)\right\}$$

- MRF is a specific case of exponential family model

$$p(x\mid\theta)=\frac{1}{Z}\prod_C \psi_C(x_C)$$

$$=\frac{1}{Z}\exp\left(\log\prod_C \psi_C(x_C)\right)=\frac{1}{Z}\exp\left(\sum_C \log\psi_C(x_C)\right)$$

Generalized Iterative scaling (GIS)

- Constraints:

$$f_i(x)\ge 0,\sum_i f_i(x)=1$$

- Update equation

$$p^{(t+1)}(x)=p^{(t)}(x)\prod_i\left(\frac{\sum_x \widetilde{p}(x)f_i(x)}{\sum_x p^{(t)}(x)f_i(x)}\right)^{f_i(x)}$$

- Update equation of **IPF**

$$p^{(t+1)}(x)=p^{(t)}(x)\frac{\widetilde{p}(x_C)}{p^{(t)}(x_C)}$$

Generalized Iterative scaling

- Log likelihood

$$l(\theta, D) = \sum_x \tilde{p}(x) \log p(x \mid \theta)$$
$$= \sum_x \tilde{p}(x) \log p(x \mid \theta) = \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta)$$

- An lower bound *Q* of the log likelihood

$$l(\theta, D) \geq Q(\theta, \theta^{(t)})$$
$$= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_i \exp(\theta_i - \theta^{(t)}) \sum_x f_i(x) p(x \mid \theta^{(t)}) - \log Z(\theta^{(t)}) + 1$$

Generalized Iterative scaling

- Same idea of EM
  - MLE of the original exponential model are difficult
  - MLE of *Q* is relative easy, because the parameters are decoupled.
- Iterative procedure
  - In step *t*, find $\theta^{(t+1)}$ which maximizes the $Q(\theta, \theta^{(t)})$

Generalized Iterative scaling

- The derivative w.r.t θ_i

$$0 = \frac{\partial Q(\theta, \theta^{(t)})}{\partial \theta_i}$$
$$= \sum_x \tilde{p}(x) f_i(x) - \exp(\theta_i - \theta_i^{(t)}) \sum_x p(x \mid \theta^{(t)}) f_i(x)$$

- We obtain

$$\exp(\theta_i^{(t+1)} - \theta_i^{(t)}) = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p(x \mid \theta^{(t)}) f_i(x)}$$
$$\Rightarrow \theta_i^{(t+1)} = \theta_i^{(t)} + \log\left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p(x \mid \theta^{(t)}) f_i(x)} \right)$$

Latent variables

- EM algorithm
  - E-step: Traditional
  - M-step: GIS algorithm

Thank you

Chenhai@cs.pitt.edu