Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2014

# 21 Learning parameters of an undirected graphical model

Today we shall see how to learn parameters or potential for an undirected graphical model from observations for all variables and using knowledge of the graphical structure.

Let $G = (V, E)$ denote the associated undirected graph over $N$ vertices $V = \{1, \ldots, N\}$ and $E \subseteq V \times V$. Let associated random variables $\mathsf{x}_1, \ldots, \mathsf{x}_N$ take value in finite alphabet $\mathcal{X}$. For the purpose of this lecture, we shall restrict attention to $\mathcal{X} = \{0, 1\}$; however, the treatment of this lecture naturally extends for any finite alphabet $\mathcal{X}$. We shall assume that $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^N$. Recall that by the Hammersley-Clifford theorem, stated and proved in an earlier lecture,

$$p(\mathbf{x}) = \exp\big(\sum_{C \in \mathcal{C}(G)} V_C(\mathbf{x}_C)\big), \tag{1}$$

where $\mathcal{C}(G)$ is the set of all cliques of $G$, including the empty set; $\mathbf{x}_C = (x_i)_{i \in C}$, and $V_C(\cdot)$ is the potential function associated with clique $C$ with the form:

$$V_C(\mathbf{x}_C) = \begin{cases} Q(C) & \text{if } \mathbf{x}_C = \mathbf{1} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We observe i.i.d. samples from this distribution, denoted as $\mathcal{D} = \{\mathbf{x}(1), \ldots, \mathbf{x}(S)\}$. The goal is to learn the potential functions associated with the cliques of $G$ from $\mathcal{D}$.

## 21.1 A simple example: binary pair-wise graphical model

Before we consider the general graph, let us consider a graph that does not have triangle. That is, $\mathcal{C}(G)$ consists of all vertices $V$, edges $E$ and the empty set. As before, $\mathcal{X} = \{0, 1\}$. Given this, any such graphical model can be represented as

$$p(\mathbf{x}) = \frac{1}{Z} \exp\big(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j\big), \tag{3}$$

for $\theta_i, \theta_{ij} \in \mathbb{R}$ for all $i \in V$, $(i, j) \in E$; $Z$ being normalization constant. Given this, the question of boils down to learning parameters $\boldsymbol{\theta}$, from $\mathcal{D}$, where

$$\boldsymbol{\theta} = (\theta_i, i \in V; \theta_{ij}, (i, j) \in E).$$

If we had a very large number of samples ($\gg 2^N$) so that each assignment $\mathbf{x} \in \{0,1\}^N$ is sampled enough times, we can estimation empirical probability of each assignment $\mathbf{x} \in \{0,1\}^N$. Now

$$\log p(\mathbf{e}_i) - \log p(\mathbf{0}) = \theta_i,$$

where $\mathbf{e}_i$ is assignment in $\{0,1\}^N$ with $i^{th}$ element 1 and everything else 0, $\mathbf{0}$ is the assignment with all element 0. Similarity,

$$\log p(\mathbf{e}_{ij}) - \log p(\mathbf{e}_i) - \log p(\mathbf{e}_j) + \log p(\mathbf{0}) = \theta_{ij},$$

where $\mathbf{e}_{ij}$ is assignment in $\{0,1\}^N$ with $i^{th}$ and $j^{th}$ elements 1 and everything else 0. Therefore, with potentially exponentially large number of samples, it is possible to learn parameters $\boldsymbol{\theta}$. Clearly this is highly inefficient. The question is, whether it's feasible to learn parameters with a lot fewer samples.

We shall utilize the quintessential property of the undirected graphical model: for any $i \in V$, given $X_{N(i)}$, $X_i$ is independent of $X_{V \backslash N(i) \cup \{i\}}$ where $N(i) = \{j \in V : (i,j) \in E\}$. That is,

$$p(x_i = \cdot | \mathbf{x}_{N(i)} = \mathbf{0}) = p(x_i = \cdot | \mathbf{x}_{V \backslash \{i\}} = \mathbf{0}).$$

Therefore,

$$\begin{aligned} &\log p(x_i = 1 | \mathbf{x}_{N(i)} = \mathbf{0}) - \log p(x_i = 0 | \mathbf{x}_{N(i)} = \mathbf{0}) \\ &= \log p(x_i = 1 | \mathbf{x}_{V \backslash \{i\}} = \mathbf{0}) - \log p(x_i = 0 | \mathbf{x}_{V \backslash \{i\}} = \mathbf{0}) \\ &= \log p(x_i = 1, \mathbf{x}_{V \backslash \{i\}} = \mathbf{0}) - \log p(x_i = 0, \mathbf{x}_{V \backslash \{i\}} = \mathbf{0}) \\ &= \log p(\mathbf{e}_i) - \log p(\mathbf{0}) \\ &= \theta_i. \end{aligned}$$

Thus, by estimating conditional probabilities $p(x_i = \cdot | \mathbf{x}_{N(i)} = \mathbf{0})$ for all $i \in V$, we can learn parameters $\theta_i, i \in V$. Similarly, by estimating $p(x_i = 1, x_j = 1 | \mathbf{x}_{N(i) \cup N(j) \backslash \{i,j\}} = \mathbf{0})$, it's feasible to learn $\theta_{ij}$ for all $(i,j) \in E$.

Therefore, it's feasible to learn parameters $\boldsymbol{\theta}$ with samples scaling as $2^{2d}$, if $|N(i)| \le d$ for all $i \in V$; which is quite efficient when $d \ll N$.

## 21.2 Generic undirected graphical model

For any graphical model with binary valued variables, to learn the associated clique potentials $V_C(\cdot)$ boils down to learning constants $Q(C)$ as in (2). They can be learnt in a very similar manner as in the example of binary pair-wise graphical model.

Concretely, for any $i \in V$,

$$\begin{aligned}
&\log p(x_i = 1|\mathbf{x}_{N(i)} = \mathbf{0}) - \log p(x_i = 0|\mathbf{x}_{N(i)} = \mathbf{0})\\
&= \log p(x_i = 1|\mathbf{x}_{V\setminus\{i\}} = \mathbf{0}) - \log p(x_i = 0|\mathbf{x}_{V\setminus\{i\}} = \mathbf{0})\\
&= \log p(x_i = 1, \mathbf{x}_{V\setminus\{i\}} = \mathbf{0}) - \log p(x_i = 0, \mathbf{x}_{V\setminus\{i\}} = \mathbf{0})\\
&= \log p(\mathbf{e}_i) - \log p(\mathbf{0})\\
&= Q(\{i\}).
\end{aligned}$$

And more generally, for any clique $C \subset V$,

$$\begin{aligned}
&\log p(x_C = \mathbf{1}|\mathbf{x}_{N(C)} = \mathbf{0}) - \log p(x_i = 0|\mathbf{x}_{N(C)} = \mathbf{0})\\
&= \sum_{C' \subset C} Q(C').
\end{aligned}$$

Using the above identity, it's possible to learn constants associated with cliques of increasing sizes iteratively. The number of samples required scale as $2^{d_c}$, where $d_c$ is bound on the number of neighbors of vertices in any clique (including the vertices in the clique) for the graph $G$.

6.438 Algorithms for Inference
Fall 2014