STAT 535 Lecture 12

Estimating the parameters of MRF's

©Marina Meilă
mmp@stat.washington.edu

(after M. Jordan) The estimation will be considered only in the ML framework.

1 The (log)-likelihood

Let, as usual, $P_V = \frac{1}{Z}\prod_C \phi_C(x_C)$ be a joint distribution represented as a MRF, with $\{C\}$ the set of cliques and Z the normalization constant.

The parameter estimation problem calls for the estimation of the entries in all the potential tables ϕ_C. We assume that each entry is a different parameter, and denote it (abusively, perhaps) by $\phi_C(x_C)$.

Denote by $\mathcal{D}$ a data set of complete observations of the variables in V, sampled i.i.d. from an unknown distributions. We denote by $N_C(x_C)$ the number of times configuration x_C over the variables in C appears in the data. We shall see that, just as in the case of the BN, the set of N_C counts are the sufficient statistics for the parameters.

We have that $\sum_{x_C \in \Omega_C} N_C(x_C) = N$.

Example Let $V = \{A, B, C, D\}$ with all variables taking values in $\{0, 1\}$ and

$$P_V = \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{DA}(d,a) \tag{1}$$

with

$$Z = \sum_{a\in\Omega_A}\sum_{b\in\Omega_B}\sum_{c\in\Omega_C}\sum_{d\in\Omega_D} \phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{DA}(d,a) \tag{2}$$

(Thus, Z is a sum over 16 terms.) Let the data set contain $N = 5$ samples.

A	B	C	D
1	1	1	0
1	0	0	1
0	1	1	0
0	0	1	0
1	0	1	1

The log-likelihood is the logarithm of the probability of the data $\mathcal{D}$ under the model P_V, i.e

$$l(\text{parameters}; \mathcal{D}) \;=\; \sum_C \sum_{x_C \in \Omega_C} N_C(x_C) \ln \phi_C(x_C) - N \ln Z \tag{3}$$

The first observation we make is that the data enter the likelihood only via the *sufficient statistics* $N_C(x_C)$. Then, we will find it convenient to normalize both sides of the above equation by the smaple size N. The ratio $\frac{N_c(x_C)}{N} = \hat{P}_C(x_C)$ represents a probability, namely the *empirical distribution* of the variable(s) in clique C, or the *sample marginal* of C w.r.t. the empirical distribution represented by the sample $\mathcal{D}$. After the normalization we obtain

$$\frac{1}{N} l \;=\; \sum_C \sum_{x_C \in \Omega_C} \hat{P}_C(x_C) \ln \phi_C(x_C) - \ln Z \tag{4}$$

The normalization constant Z is a function of all parameters.

For the **example** above, we have

$$N_{AB} \;=\; \begin{array}{r|cc} B: & 0 & 1 \\ \hline A=0 & 1 & 2 \\ 1 & 1 & 1 \end{array} \qquad N_{BC} \;=\; \begin{array}{r|cc} C: & 0 & 1 \\ \hline B=0 & 1 & 2 \\ 1 & 0 & 2 \end{array} \qquad \ldots \tag{5}$$

The log-likelihood of this data is

$$\frac{1}{N} l \;=\; \left(\frac{1}{5} \ln \phi_{AB}(0,0) + \frac{2}{5} \ln \phi_{AB}(0,1) + \frac{1}{5} \ln \phi_{AB}(1,0) + \frac{1}{5} \ln \phi_{AB}(1,1) \right) \tag{6}$$

$$+ \left(\frac{1}{5} \ln \phi_{BC}(0,0) + \frac{2}{5} \ln \phi_{BC}(0,1) + \frac{2}{5} \ln \phi_{BC}(1,1) \right) + \ldots - \ln Z \tag{7}$$

2 Maximizing the likelihood by gradient ascent

To find the maximum value for the parameters, we use the iterative procedure called **gradient ascent**.

GradientAscent

Input sufficient statistics $N_C(x_C)$ (or sample marginals $\hat{P}_C(x_C)$) for all C and all x_C

Initialize ϕ_C with arbitrary values

Repeat

1. $\phi_C(x_C) \leftarrow \phi_C(x_C) + \eta \frac{\partial l/N}{\partial \phi_C(x_C)}$ for all x_C

until "convergence"

In other words, GradientAscent iteratively corrects the current parameter estimates with a correction that will increase the log-likelihood l. The parameter $\eta > 0$ is a *step size* that is sometimes fixed and sometimes estimated at each step (as we shall see in STAT 538).

GradientAscent is a generic optimization algorithm. To use it we need to calculate the expression of the gradient $\frac{\partial l/N}{\partial \phi_C(x_C)}$. We do so now.

$$\frac{\partial l/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\partial}{\partial \phi_{C_0}(x^*_{C_0})} \sum_C \sum_{x_C \in \Omega_C} \frac{N_C(x_C)}{N} \ln \phi_C(x_C) - \frac{\partial}{\partial \phi_{C_0}(x_{C_0^*})} \ln Z \quad (8)$$

$$= \underbrace{\frac{N_{C_0}(x^*_{C_0})}{N}}_{\hat{P}_{C_0}(x^*_{C_0})} \frac{1}{\phi_{C_0}(x^*_{C_0})} - \frac{\frac{\partial Z}{\partial \phi_{C_0}(x^*_{C_0})}}{Z} \quad (9)$$

$$\frac{\partial Z}{\partial \phi_{C_0}(x_{C_0^*})} = \sum_{x_V} \frac{\partial}{\partial \phi_{C_0}(x^*_{C_0})} \left[\prod_{C' \neq C_0} \phi_{C'}(x_{C'}) \phi_{C_0}(x^*_{C_0}) \right] \quad (10)$$

$$= \sum_{x_{V \setminus C_0}} \prod_{C' \neq C_0} \phi_{C'}(x_{C' \setminus C_0}, x^*_{C' \cap C_0}) \quad (11)$$

$$= \frac{Z}{\phi_{C_0}(x_{C_0^*})} \sum_{x_{V \setminus C_0}} \prod_{C' \neq C_0} \phi_{C'}(x_{C' \setminus C_0}, x^*_{C' \cap C_0}) \frac{\phi_{C_0}(x_{C_0^*})}{Z} \quad (12)$$

$$= \frac{Z}{\phi_{C_0}(x_{C_0^*})} \sum_{x_{V \setminus C_0}} P_V(x_{V \setminus C_0}, x^*_{C_0}) \quad (13)$$

$$= Z \frac{P_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} \quad (14)$$

$$\frac{\partial l/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0}) - P_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} \quad (15)$$

Thus, the gradient w.r.t a parameter $\phi_{C_0}(x^*_{C_0})$ depends on: (1) the current value of the parameter $\phi_{C_0}(x^*_{C_0})$, (2) the *empirical marginal* probability $\hat{P}$ of the configuration $x^*_{C_0}$, and (3) the marginal of the respective configuration as computed by the model P_V, i.e. $P_{C_0}(x_{C_0^*})$.

The first two quantities are readily available. The marginal $P_{C_0}(x_{C_0^*})$, however, must be computed. As we know, computing marginals is *inference*, thus we must

be able to perform inference in the MRF in order to estimate the parameters.

For inference we can use

- variable elimination
- triangulation and Junction Tree algorithm
- MCMC (aka Markov Chain Monte Carlo), which produces approximate marginals

A more general observation related to equation (**??**). We see from it that $\frac{\partial \ln Z}{\partial \phi_C} = \frac{1}{\phi_C} P_C$. Assume there is another variable $\theta_C(x_C)$ for every $\phi_C(x_C)$ so that $\frac{d\theta_C(x_C)}{d\phi_C(x_C)} = \frac{1}{\phi_C(x_C)}$. Then, obviously, we would have

$$\frac{\partial \ln Z}{\partial \theta_C} \;=\; P_C \tag{16}$$

with $\theta_C = \ln \phi_C$. How would P_V look like in the parametrization θ?

$$P_V(x) \;=\; \frac{1}{Z} \prod_C e^{\theta_C(x_C)} \;=\; \frac{1}{Z} e^{\sum_C \theta_C(x_C)} \tag{17}$$

Equation (16) and representation (**??**) are typical of the more general class of *exponential family models*, or which MRF's are an example.

3 Iterative Proportional Fitting (IPF)

IPF is an alternative to gradient ascent, that does not require setting the step size.

The idea is that, at the optimum, the gradient will be zero. Hence we will have

$$\frac{\hat{P}_C(x_C)}{\phi_C(x_C)} \;=\; \frac{P_C(x_C)}{\phi_C(x_C)} \tag{18}$$

or

$$\phi_C(x_C) \;=\; \phi_C(x_C) \frac{\hat{P}_C(x_C)}{P_C(x_C)} \tag{19}$$

for all cliques C and configurations x_C. The IPF algorithm tries to reach this equilibrium point by multiplicative updates to the parameter $\phi_C(x_C)$ (while gradient ascent performs additive updates).

IPF **Algorithm**

Repeat

for every clique $C \in \mathcal{C}$

$$\phi_C(x_C) \leftarrow \phi_C(x_C)\frac{\hat{P}_C(x_C)}{P_C(x_C)} \tag{20}$$

until convergence

Proposition 1 The IPF algorithm preserves the value of the normalization constant Z.

Proof Assume that at step t the parameters of clique C are updated while the other cliques' parameters stay the same.

$$\begin{aligned}
P_C^{(t+1)}(x_C) &= \sum_{x_{V\setminus C}} P_V^{(t+1)}(x_V) && (21)\\
&= \sum_{x_{V\setminus C}} \prod_{C'} \phi_{C'}^{(t+1)}(x_{C'})/Z^{(t+1)} && (22)\\
&= \sum_{x_{V\setminus C}} \phi_C(x_C)^{(t+1)} \prod_{C'\neq C} \phi_{C'}^{(t)}(x_{C'})/Z^{(t+1)} && (23)\\
&= \sum_{x_{V\setminus C}} \phi_C^{(t)}(x_C)\frac{\hat{P}_C(x_C)}{P_C^{(t)}(x_C)} \prod_{C'\neq C} \phi_{C'}^{(t)}(x_{C'})/Z^{(t+1)} && (24)\\
&= \frac{\hat{P}_C(x_C)}{Z^{(t+1)}P_C(x_C)} \sum_{x_{V\setminus C}} \underbrace{\phi_C(x_C)^{(t)} \prod_{C'\neq C} \phi_{C'}^{(t)}(x_{C'})}_{P_V^{(t)}(x)Z^{(t)}} && (25)\\
&= \frac{Z^{(t)}\hat{P}_C^{(t)}(x_C)}{Z^{(t+1)}P_C(x_C)} \sum_{x_{V\setminus C}} P_V^{(t)}(x) && (26)\\
&= \frac{Z^{(t)}\hat{P}_C(x_C)}{Z^{(t+1)}P_C^{(t)}(x_C)} P_C^{(t)}(x_C) && (27)\\
&= \frac{Z^{(t)}}{Z^{(t+1)}}\hat{P}_C(x_C) && (28)
\end{aligned}$$

Summing now both sides over $x_C \in \Omega_C$ and noting that $\sum_{\Omega_C} \hat{P}_C = \sum_{\Omega_C} P_C = 1$ we obtain $\frac{Z^{(t)}}{Z^{(t+1)}} = 1$, which completes the proof.

From $Z^{(t)} = Z^{(t+1)}$ and (28) we can immediately derive:

Proposition 2 After updating the parameters ϕ_C we have that $P_C = \hat{P}_C$.

Hence, the IPF updates can be thought of as updating each clique iteratively in a way that makes its marginal equal to the data marginal. Next, we will give an interpretation for IPF as gradient ascent.

Let us consider the gradient (9) with the second term expressed by equation (13).

$$\frac{\partial l^{(t)}/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} - \frac{P^{(t)}_{C_0}(x^*_{C_0})}{\phi^{(t)}_{C_0}(x^*_{C_0})} \tag{29}$$

Note that the first occurence of $\phi_{C_0}(x^*_{C_0})$ is as a *function argument*, while its second occurence is as a *parameter value*. We evaluate this gradient now at the *next* parameter value, $\phi^{(t+1)}_{C_0}(x^*_{C_0})$.

$$\left.\frac{\partial l^{(t)}/N}{\partial \phi_{C_0}(x^*_{C_0})}\right|_{\phi^{(t+1)}_{C_0}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0})}{\phi^{(t+1)}_{C_0}(x^*_{C_0})} - \frac{P^{(t)}_{C_0}(x^*_{C_0})}{\phi^{(t)}_{C_0}(x^*_{C_0})} \tag{30}$$

If we set the condition that the gradient in this point is zero, we obtain equation (20). Setting this condition implies that we move in parameter space in the direction of the gradient until we (approximately) find a point where the gradient is zero, i.e. the point of the maximum increase along the gradient direction.