6.892 Machine learning seminar

Prof. Tommi S. Jaakkola
tommi@ai.mit.edu

February 24, 1999

ML parameter estimation in undirected models

Suppose we have an undirected model with a joint distribution

$$P(x_1, \ldots, x_n) = \prod_c \phi_c(x_c) \tag{1}$$

where the product is over the cliques in the graph. Given a set of observations $D_T = \{x^1, \ldots, x^T\}$ we would like to find the parameters, the clique potentials, so as to maximize the (average) likelihood of these observations

$$\frac{1}{T}\sum_{t=1}^{T} \log P(x_1^t, \ldots, x_n^t) = \sum_x \hat{P}(x) \log P(x_1, \ldots, x_n) \tag{2}$$

where $\hat{P}(x)$ is the empirical distribution

$$\hat{P}(x) = \sum_{t=1}^{T} \delta(x = x^t) \tag{3}$$

where $\delta(x = x^t) = 1$ whenever $x = x^t$ and zero otherwise. Taking into account the structure of the joint distribution gives

$$\sum_x \hat{P}(x) \log P(x_1, \ldots, x_n) = \sum_c \sum_{x_c} \hat{P}(x_c) \log \phi_c(x_c) \tag{4}$$

Here $\hat{P}(x_c)$ are the marginals computed from the empirical distribution. It is clear from this form that the marginals $\{\hat{P}(x_c)\}$ are sufficient statistics for estimating the clique potentials $\{\phi_c(x_c)\}$. Since the joint distribution P must sum to one, the estimation of these potential functions is subject to the constraint

$$\sum_x \prod_c \phi_c(x_c) = 1 \tag{5}$$

making it less apparent how the potential functions ought to be chosen.

If we have to base our estimation of the potential functions on a collection of empirical marginals $\{\hat{P}(x_c)\}$ then it seems reasonable to choose the potentials so that the marginals computed from the model agree with these empirical marginals. This is indeed what the maximum likelihood objective forces us to do. To see this, consider the following modified likelihood objective

$$l = \sum_x \hat{P}(x) \log \frac{\prod_c \phi_c(x_c)}{\sum_{x'} \prod_{c'} \phi_c(x'_{c'})} = \sum_c \sum_{x_c} \hat{P}(x_c) \log \phi_c(x_c) - \log \sum_{x'} \prod_{c'} \phi_c(x'_{c'}) \tag{6}$$

where the denominator is the normalizing constant (partition function) which is introduced here merely to be able to take derivatives with respect to the potential functions subject to the overall constraint that the distribution sums to one (the value of the denominator at the current setting of the clique potentials is assumed to be one). At the maximum likelihood setting the derivatives of the likelihood objective with respect to these potentials must be zero. Thus

$$\begin{aligned}
\frac{\partial l}{\partial \phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} &= \sum_c \sum_{x_c} \hat{P}(x_c) \frac{\partial}{\partial \phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} \log \phi_c(x_c) - \frac{1}{\sum_{x'} \prod_{c'} \phi_c(x'_{c'})} \sum_{x'} \frac{\partial}{\partial \phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} \prod_{c'} \phi_{c'}(x'_{c'}) && (7)\\
&= \frac{\hat{P}_{\tilde{c}}(\tilde{x}_{\tilde{c}})}{\phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} - \sum_{x'} \delta(x'_{\tilde{c}} = \tilde{x}_{\tilde{c}}) \prod_{c' \neq \tilde{c}} \phi_{c'}(x'_{c'}) && (8)\\
&= \frac{\hat{P}_{\tilde{c}}(\tilde{x}_{\tilde{c}})}{\phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} - \sum_{x'} \delta(x'_{\tilde{c}} = \tilde{x}_{\tilde{c}}) \frac{1}{\phi_c(x'_{\tilde{c}})} \prod_{c'} \phi_{c'}(x'_{c'}) && (9)\\
&= \frac{\hat{P}_{\tilde{c}}(\tilde{x}_{\tilde{c}})}{\phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} - \frac{1}{\phi_c(\tilde{x}_{\tilde{c}})} \sum_{x'} \delta(x'_{\tilde{c}} = \tilde{x}_{\tilde{c}}) \prod_{c'} \phi_{c'}(x'_{c'}) && (10)\\
&= \frac{\hat{P}_{\tilde{c}}(\tilde{x}_{\tilde{c}})}{\phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} - \frac{1}{\phi_c(\tilde{x}_{\tilde{c}})} \sum_{x'} \delta(x'_{\tilde{c}} = \tilde{x}_{\tilde{c}}) P(x') && (11)\\
&= \frac{\hat{P}_{\tilde{c}}(\tilde{x}_{\tilde{c}})}{\phi_{\tilde{c}}(\tilde{x}_{\tilde{c}})} - \frac{P(\tilde{x}_{\tilde{c}})}{\phi_c(\tilde{x}_{\tilde{c}})} = 0 && (12)
\end{aligned}$$

which clearly implies that the marginals computed from the model must agree with the (available) empirical marginals. To refine the potential functions so that this condition holds for all the empirical marginals, we resort to a simple algorithm known as *Iterative Proportial Fitting* or IPF for short.

Iterative proportional fitting

Given a consistent set of observed marginals $\{\hat{P}(x_c)\}$, we want to set the potential functions in the undirected probability model so that the model agrees with these marginals. This can be done through the following iterative scheme:

1. Choose a clique c and compute the marginal $P(x_c)$ from the current distribution.
2. Update the corresponding clique potential according to

$$\phi_c(x_c) \leftarrow \phi_c(x_c)\, \frac{\hat{P}(x_c)}{P(x_c)}, \quad \forall x_c \tag{13}$$

3. Iterate until $P(x_c) \cong \hat{P}(x_c)$ for all x_c and c.

Note that since we are updating only clique potentials the structure of the model cannot change. The observed marginals $\hat{P}(x_c)$ need not be over the whole cliques so long as they are contained in at least one. To show that this procedure actually converges let $P(x)$ be the probability distribution corresponding to the current setting of the potential functions and let $P'(x)$ be the distribution after having updated one of the clique potentials, say $\phi_c(x_c)$. Thus

$$P'(x) = \frac{\hat{P}(x_c)}{P(x_c)} \phi_c(x_c) \prod_{c' \neq c} \phi_{c'}(x_{c'}) = \frac{\hat{P}(x_c)}{P(x_c)} P(x) \tag{14}$$

We aim to show that the relative entropy between the empirical distribution $\hat{P}$ (which has the correct marginals) and our distribution P decreases after each update so long as the marginals $\hat{P}(x_c)$ and $P(x_c)$ disagree. Consider therefore

$$\begin{aligned} D(\hat{P} \| P') = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P'(x)} &= \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P(x)\frac{\hat{P}(x_c)}{P(x_c)}} & (15) \\ &= \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P(x)} - \sum_{x_c} \hat{P}(x_c) \log \frac{\hat{P}(x_c)}{P(x_c)} & (16) \\ &= D(\hat{P} \| P) - D(\hat{P}_c \| P_c) & (17) \end{aligned}$$

The fact that relative entropy cannot become negative guarantees that $D(\hat{P}_c \| P_c) \to 0$ as the iteration proceeds. This implies that $P(x_c) \to \hat{P}_c(x_c)$ as desired.