6.892 Machine learning seminar

http://www.ai.mit.edu/courses/6.892-ml

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

March 17, 1999

1 Beyond simple mean field

- We have transformed the inference problem in graphical models as an optimization problem within the maximum entropy framework.
- Let $x = x_v \cup x_h$ where x_v is the set of "visible" variables whose probability we are interested in computing and x_h is the set of remaining or "hidden" variables.

  (Note that h is a set of nodes in the graph while x_h is the set of variables associated with these nodes)

  A general inference problem is to compute the marginal probability over the variables of interest (here on log-scale):

$$\log P_0(x_v|\theta) = \log \sum_{x_h} P_0(x_v, x_h|\theta) \tag{1}$$

- As discussed earlier in the class we can reduce this calculation into an optimization problem:

$$\log P_0(x_v|\theta) = \max_P \left\{ \sum_{x_h} P(x_h) \log P_0(x_v, x_h|\theta) + H(P) \right\} \tag{2}$$

  P is known as the variational distribution.

  Alternative proof of this (in case you didn't follow the physics point of view): for any distribution P over the hidden variables x_h we have

$$\sum_{x_h} P(x_h) \log P_0(x_v, x_h|\theta) + H(P) = \sum_{x_h} P(x_h) \log \frac{P_0(x_v, x_h|\theta)}{P(x_h)} \tag{3}$$

$$\overset{\text{Jensen's ineq.}}{\leq} \quad \log \sum_{x_h} P(x_h) \frac{P_0(x_v, x_h|\theta)}{P(x_h)} \tag{4}$$

$$= \quad \log \sum_{x_h} P_0(x_v, x_h|\theta) = \log P_0(x_v) \tag{5}$$

Note that you'd still have to show that the maximum is attained for some P. This is the P for which Jensen's inequality gives an equality. In general, this happens when the argument to the log does not vary as a function of x. This property holds whenever P is equal to the posterior probability over x_h, i.e, $P(x_h) = P_0(x_h|x_v, \theta)$.

- There is no magic here: to compute the desired marginal probability over the visible variables x_v, we have to set P equal to the true posterior probability over the hidden variables in evaluating the bound. This is no easier than computing the marginal probability directly.

- The advantage of this variational formulation lies in the fact that we can now easily approximate the marginal probability. For any P, no matter how poorly it approximates the posterior, we nevertheless have a lower bound:

  $$\log P_0(x_v|\theta) \geq \sum_{x_h} P(x_h) \log P_0(x_v, x_h|\theta) + H(P) \tag{6}$$

  Our task is now to find a class of distributions such that the lower bound can be computed and optimized (maximized) easily for P in this class.

- In mean field approximation we assume that the variational distribution P is completely factorized:

  $$P(x_h) = \prod_{i \in h} P_i(x_i) \tag{7}$$

  This is a crude approximation to the true posterior over x_h (which would attain the maximum of the lower bound). It nevertheless allows us to a) evaluate the bound easility and b) iteratively optimize the component distributions in the above factorization.

  The set of factorized distributions is not convex, i.e., the a convex combination of two different factorized distributions is rarely factorized. This means that we cannot expect to find the globally optimal P in the sense of maximizing the lower bound within the class of factorized distributions. Locally optimal solutions are often useful, however.

- It is important to realize that the bound is typically optimized separately for each "query" configuration x_v. This increases the accuracy of the approximation considerably.

1.1 On factorizations

- It is essential to differentiate two different factorizations:

  1) that of the original model $P_0(x_v, x_h|\theta)$
  2) that of the variational distribution $P(x_h)$

  These factorizations need not be related. It is, however, natural and useful to factorize the variational distribution P according to the orginal distribution P_0 so long as P still remains tractable. After all the best variational distribution is the posterior $P(x_h) = P_0(x_h|x_v, \theta)$.

1.2 Structured mean field

- Let us now generalize the mean field approximation a little to allow more general factorizations.

  Let $h = h_1 \cup h_2 \cup \ldots \cup h_p$ be a disjoint partition of the hidden nodes. In other words, $h_i \cap h_j = \phi$ for $i \neq j$. We assume that variational distribution P factorizes accordingly:

$$P(x_h) = \prod_{k=1}^{p} P_{h_k}(x_{h_k}) \tag{8}$$

  (note that "factorization" here implies that the component distributions have no variables in common with each other. This is a stricter form of factorization than e.g. the generic factorization of a joint distribution consistent with a graph structure. We will generalize this later on).

  The entropy of this distribution is simply the sum of the entropies of the component distribution:

$$H(P) = \sum_{k} H(P_{h_k}) \tag{9}$$

- Notation:

  Let us again use the notation $\langle \cdot \rangle$ to denote the average with respect to the above distribution: for any $f(x_h)$

$$\langle f(x_h) \rangle = \sum_{x_h} \left[\prod_{k=1}^{p} P_{h_k}(x_{h_k}) \right] f(x_h) \tag{10}$$

Also for an average over x_h except x_{h_j} we use

$$\langle f(x_h)\rangle_j = \sum_{x_h \backslash x_{h_j}} \left[\prod_{k\neq j}^{p} P_{h_k}(x_{h_k})\right] f(x_h) \tag{11}$$

It is important to realize that while $\langle f(x_h)\rangle$ is a constant, i.e., doesn't depend on any of the variables x_h, $\langle f(x_h)\rangle_j$ is in general a function of x_{h_j}. Note also that

$$\sum_{x_{h_j}} P_{h_j}(x_{h_j})\langle f(x_h)\rangle_j = \langle f(x_h)\rangle \tag{12}$$

- We can now write the lower bound as before

$$\begin{aligned}
\log P_0(x_v|\theta) &\geq \langle \log P_0(x_v, x_h|\theta)\rangle + H(P) && (13)\\
&= \sum_{x_{h_j}} P_{h_j}(x_{h_j})\langle \log P_0(x_v, x_h|\theta)\rangle_j \\
&\quad + H(P_{h_j}) + \sum_{k\neq j} H(P_{h_k}) && (14)
\end{aligned}$$

where we have pulled out a particular component distribution $P_{h_j}(x_{h_j})$. We can do this trick for any of the component distributions. Here

$$\langle \log P_0(x_v, x_h|\theta)\rangle_j \tag{15}$$

is a function of x_v and x_{h_j} only.

- The optimal setting of the component distribution $P_{h_j}(x_{h_j})$ when the other components of the variational distribution are held fixed is the maximum entropy style solution:

$$P_{h_j}(x_{h_j}) \leftarrow \frac{1}{Z} e^{\langle \log P_0(x_v, x_h|\theta)\rangle_j} \tag{16}$$

where Z normalizes the distribution over x_{h_j}.

In evaluating the exponent $\langle \log P_0(x_v, x_h|\theta)\rangle_j$ we have to include only the components in the factorization of $P_0(x_v, x_h|\theta)$ that depend on some of the variables in x_{h_j}. All other terms would cancel during the normalization. In other words, if

$$P_0(x_v, x_h|\theta) = \prod_c \phi_c(x_c) = \left[\prod_{c: c\cap h_j = \emptyset} \phi_c(x_c)\right]\left[\prod_{c': c'\cap h_j \neq \emptyset} \phi_{c'}(x_{c'})\right] \tag{17}$$

Then

$$\langle \log P_0(x_v, x_h|\theta)\rangle_j = \sum_{c': c' \cap h_j \neq \emptyset} \langle \log \phi_{c'}(x_{c'})\rangle_j + \text{constant} \tag{18}$$

where the "constant" is not a function of any of the variables in x_{h_j} and we can safely ignore it in updating P_{h_j}.

1.3 Directed graphs as variational distributions

- The approximate distribution need not be factorized disjointly as we assumed in the previous section. We can, for example, consider tree structured variational distributions or more generally variational distributions P that are themselves directed graphical models.

- Consider a tractable[1] variational distribution of the form

$$P(x_h) = \prod_{i \in h} P(x_i|x_{pa_i}) \tag{19}$$

where necessarily $\{x_i, x_{pa_i}\} \in x_h$. If this defines a tree then the set of parents pa_i contains at most one node.

Some notation: for any $f(x_h)$ let

$$\langle f(x_h)\rangle = \sum_{x_h} \left[\prod_{i \in h} P(x_i|x_{pa_i})\right] f(x_h) \tag{20}$$

and we also define a partially completed average over x_h where x_j and x_{pa_j} are assumed to be fixed and the corresponding conditional probability $P(x_j|x_{pa_j})$ has been surgically removed:

$$\langle f(x_h)\rangle_{j,pa_j} = \sum_{x_h \setminus \{x_j, x_{pa_j}\}} \left[\prod_{i \neq j} P(x_i|x_{pa_i})\right] f(x_h) \tag{21}$$

Note that this is *not* a conditional average but merely a partially completed average. By design, this "average" is not a function of the conditional probability $P(x_j|x_{pa_j})$. As before, we have the convenient property that

$$\sum_{x_j, x_{pa_j}} P(x_j|x_{pa_j}) \langle f(x_h)\rangle_{j,pa_j} = \langle f(x_h)\rangle \tag{22}$$

[1]The structure of this distribution is at this point given by the user and we may assume that the distribution is tractable. Otherwise it would be of no use.

- Since in this general case the factorization is not disjoint, the entropy term no longer reduces to a sum of component entropies. We can nevertheless explicate the dependence of the entropy on any particular conditional probability $P(x_j|x_{pa_j})$:

$$\begin{aligned}
H(P) &= -\sum_{x_h} P(x_h) \log \prod_{i \in h} P(x_i|x_{pa_i}) && (23)\\
&= -\sum_i \sum_{x_{pa_i}} P(x_{pa_i}) \sum_{x_i} P(x_i|x_{pa_i}) \log P(x_i|x_{pa_i}) && (24)\\
&= \sum_i \sum_{x_{pa_i}} P(x_{pa_i}) H(P_{x_i|x_{pa_i}}) && (25)\\
&= \sum_{i<j} \sum_{x_{pa_i}} P(x_{pa_i}) H(P_{x_i|x_{pa_i}}) + \sum_{x_{pa_j}} P(x_{pa_j}) H(P_{x_j|x_{pa_j}}) \\
&\quad + \sum_{x_j, x_{pa_j}} P(x_j|x_{pa_j}) \langle \sum_{i>j} H(P_{x_i|x_{pa_i}}) \rangle_{j,pa_j} && (26)
\end{aligned}$$

  where we have assumed that the nodes are ordered in such as way that pa_i always precedes i in the ordering ("parents" come before "children"). The first term in the above result does not depend on the conditional $P(x_j|x_{pa_j})$, the second term contains the entropy of $P(x_j|x_{pa_j})$, and the last term depends linearly on $P(x_j|x_{pa_j})$.

- The solution to any maximization problem in the space of probability distributions that involves only linear terms in addition to the entropy of the distribution (possibly weighted with a positive coefficient) reduces to the same maximum entropy form as we have seen before.

- Omitting the details, we can write the update equations for any variational distribution that factorizes according to a directed graph as:

$$P(x_j|x_{pa_j}) \leftarrow \frac{1}{Z(x_{pa_j})} \exp\left\{ \frac{1}{P(x_{pa_j})} \left\langle \log P_0(x_v, x_h|\theta) + \sum_{i>j} H(P_{x_i|x_{pa_i}}) \right\rangle_{j,pa_j} \right\} \quad (27)$$

  where for each fixed x_{pa_j}, $Z(x_{pa_j})$ normalizes the distribution over x_j.

2 Parameter estimation and the EM algorithm

- We can use the above approximate inference formalism for estimating the parameters of $P_0(x_v, x_h|\theta)$.

  Suppose we have a set of observations $\{x_v^t\}$, $t = 1, \ldots, T$. We assume here merely for simplicity of notation that all the observations are over the same set of "visible" variables x_v. We want to maximize the likelihood of these observations with respect to θ:

$$\max_\theta \sum_{t=1}^{T} \log P_0(x_v^t|\theta) \tag{28}$$

- Whenever our model P_0 is tractable we can use the EM-algorithm

  E-step: compute

  $$Q(\theta|\theta_{old}) = \sum_{t=1}^{T} \sum_{x_h} P_0(x_h|x_v^t, \theta_{old}) \log P_0(x_v^t, x_h|\theta) \tag{29}$$

  M-step: find the next set of parameters θ that maximize (increase)

  $$Q(\theta|\theta_{old}) \tag{30}$$

  To formalized the EM-algorithm in terms of the variational lower bound introduced earlier, consider

  $$L(x_v, P, \theta) = \sum_{x_h} P(x_h) \log P_0(x_v, x_h|\theta) + H(P) \tag{31}$$

  which is the lower bound for a particular setting of of the visible configuration x_v, the variational distribution P and the parameters θ of the model P_0. We have already seen that for this variational bound

  $$\log P_0(x_v|\theta) = \max_P L(x_v, P, \theta) = L(x_v, P^*, \theta) \geq L(x_v, P, \theta) \tag{32}$$

  where $P^*(x_h) = P_0(x_h|x_v, \theta)$, the posterior distribution, and P is any other distribution over x_h.

  Define a max–max algorithm as follows:

  M1-step: find $P^t(x_h)$, $t = 1, \ldots, T$ that maximize

  $$\sum_t L(x_v^t, P^t, \theta_{old}) \tag{33}$$

  M2-step: find θ that maximize

  $$\sum_t L(x_v^t, P^t, \theta) \tag{34}$$

  where the variational distributions $\{P^t\}$ are held fixed.

To see that this is indeed the EM-algorithm note that the first maximization step yields $P^t(x_h) = P_0(x_h|x_v^t, \theta_{old})$ for all t (see above). Thus the second step reduces to maximizing

$$
\begin{aligned}
& \sum_t L(x_v^t, P^t, \theta) \\
&= \sum_t \left[\sum_{x_h} P_0(x_h|x_v^t, \theta_{old}) \log P_0(x_v^t, x_h|\theta) + H(P_{old}^t) \right] \qquad (35) \\
&= Q(\theta|\theta_{old}) + \sum_t H(P_{old}^t) \qquad (36)
\end{aligned}
$$

with respect to θ. The second term does not depend on θ, and thus this maximization is equivalent to the M-step of the EM-algorithm.

- In a (generalized) mean field version of the above algorithm, we simply carry out the first maximization with respect to $\{P^t\}$ within the class of (partially) factorized distributions, discussed earlier.

- Whenever the first maximization is not attained, the algorithm increases a lower bound on the log-likelihood of the observations rather than the log-likelihood itself. This is the price that one pays for the ability to deal with intractable distributions.

  [Figure: true log-likelihood as a function of the EM iterations and the increasing lower bound]

References

Barber and Wiegerinck (1998). Variational approximations of graphical models using undirected graphs. NIPS 11.

Dempster, Laird, and Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**:1-38.

Haft, Hoffmann and Tresp (1998). Mean field inference in general probabilistic setting. Proceedings of the seventh international workshop on Artificial Intelligence and Statistics.

Neal and Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. Learning in graphical models. M. I. Jordan (ed.). MIT Press.

Peterson and Anderson (1987). A mean field theory learning algorithm for neural networks. Complex Systems 1, 995-1019.

Saul, Jaakkola, Jordan (1996). Mean field theory for sigmoid belief networks. Availabe from http://www.cs.washington.edu/research/jair/abstracts/saul96a.html

Saul and Jordan (1996). Exploiting tractable substructures in intractable networks. In Advances of Neural Information Processing Systems 8.