PROBABILISTIC GRAPHICAL MODELS.

Lecture 4 : LEARNING PARAMETERS WITH HIDDEN VARIABLES.

In this chapter we consider learning parameters from iid samples of the form

$$\{ \underbrace{\underline{v}^{(1)}\ \underline{h}^{(1)}} ;\ \underbrace{\underline{v}^{(2)}\ \underline{h}^{(2)}} ;\ \ldots ;\ \underbrace{\underline{v}^{(N)}\ \underline{h}^{(N)}} \}$$

$$\{ \underline{x}^{(1)} ;\ \underline{x}^{(2)} ;\ \ldots ;\ \underline{x}^{(N)} \}.$$

where $v^{(1)} \ldots v^{(N)}$ are accessible or <u>visible</u> and $\underline{h}^{(1)} \ldots \underline{h}^{(N)}$ are not accessible or <u>hidden</u> (or erased).

It turns out that directly maximizing the log-likelihood is intractable because the marginalization over hidden variables is difficult. We will use so-called <u>variational methods</u>.

<u>The setting</u> is as follows: we suppose the data comes from a probabilistic model with visible and hidden variables

$$p(\underline{v}, \underline{h}, \vartheta) = p(\underline{v}, \underline{h} \mid \vartheta)\ \underbrace{p_0(\vartheta)}_{\text{prior.}}$$

visible hidden

Here we assume a prior $p_0(\vartheta)$ over parameters ϑ. The case of ML training corresponds a flat-prior $p_0(\vartheta) \sim$ constant.

Examples of model: RBM, HMM, Gaussian Mixture Model,

The variational method replaces the distribution

$p(\underline{h}^{(1)} \ldots \underline{h}^{(N)} \vartheta \mid \underline{v}^{(1)} \ldots \underline{v}^{(N)})$ which is not

available because the $\underline{h}$'s are not observed (i.e empirical frequencies are not available) by a "variational class" of distributions of the form

$$\left\{ \prod_{m=1}^{N} q_m(\underline{h}^{(m)}) \right\} q(\vartheta)$$

and optimize $L(\vartheta)$ over q_m's and $q(\vartheta)$.

This will lead us to the "variational Bayes Expectation-Maximization algorithm". For the class $q(\vartheta) = \delta_{\vartheta \vartheta_*}$ and flat $p_0(\vartheta)$ this reduces to the

standard Expectation-Maximization (EM) algorithm.

Remark: $\underline{h}^{(m)}, m=1 \ldots N$ are dummy variables that are NOT observed - Only $\underline{v}^{(m)}$ are known.

I. VARIATIONAL BOUND.

Assume iid data samples $\underline{v}^{(1)} \underline{h}^{(1)} \dots \underline{v}^{(N)} \underline{h}^{(N)}$ where $\underline{v}^{(1)} \dots \underline{v}^{(N)}$ are visible and $\underline{h}^{(1)} \dots \underline{h}^{(N)}$ are hidden. We have

$$\boxed{\log \mathbb{P}(\underline{v}^{(1)} \dots \underline{v}^{(N)}) \geq \sum_{m=1}^{N} \mathbb{E}_{q_m(\underline{h}^m) q(\theta)} \left[\log p(\underline{h}^m, \underline{v}^m | \theta) \right] - \sum_{m=1}^{N} \mathbb{E}_{q_m(\underline{h}^m)} \left[\log q_m(\underline{h}^m) \right] - \mathbb{E}_{q(\theta)} \left[\log \frac{q(\theta)}{p_0(\theta)} \right]}$$

- first term is called energy term.
- second " " entropy term.
- optimization of the lower bound on $q_m(h^m)$ & $q(\theta)$ will give the variational Bayes EM algorithm.

Proof of the variational bound:

The starting point is to write

$$KL\left(\underbrace{\prod_{m=1}^{N} q_m(h^m)\, q(\theta)}_{\text{variational distr}} \,\Big\|\, p(h^{(1)} \dots h^{(N)}, \theta \,|\, v^{(1)} \dots v^{(N)}) \right) \geq 0$$

$$KL = \mathbb{E}_{\left\{\prod_{m=1}^{N} q_m(h^m)\right\} q(\vartheta)} \left[\log \left\{ \prod_{m=1}^{N} q_m(h^m) \right\} q(\vartheta) \right]$$

$$- \mathbb{E}_{\left\{\prod_{m=1}^{N} q_m(h^m)\right\} q(\vartheta)} \left[\log p(h^{(1)} \dots h^{(N)}, \vartheta \mid v^{(1)} \dots v^{(N)}) \right]$$

$$= \sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \left[\log q_m(h^m) \right] + \mathbb{E}_{q(\vartheta)} \left[\log q(\vartheta) \right]$$

$$- \mathbb{E}_{\left\{\prod_{m=1}^{N} q_m(h^m)\right\} q(\vartheta)} \left[\log \frac{p(h^{(1)} \dots h^{(N)}, \vartheta, v^{(1)} \dots v^{(N)})}{p(v^{(1)} \dots v^{(N)})} \right]$$

Then we get for the log-likelihood of the visible data:

$$\log p(v^{(1)} \dots v^{(N)}) = KL\left(\prod_{m=1}^{N} q_m(h^m) q(\vartheta) \,\|\, p(h^{(1)} \dots h^{(N)} \vartheta \mid v^{(1)} \dots v^{(N)}) \right)$$

$$+ \mathbb{E}_{\prod_{m=1}^{N} q_m(h^m) q(\vartheta)} \left(\log p(h^{(1)} \dots h^{(N)}, \vartheta, v^{(1)} \dots v^{(N)}) \right)$$

$$- \sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \left(\log q_m(h^m) \right) - \mathbb{E}_{q(\vartheta)} \left[\log q(\vartheta) \right]$$

Assume now that the Data is iid;

$$p(h^{(1)} \dots h^{(N)}, \vartheta, v^{(1)} \dots v^{(N)}) = p(h^{(1)} \dots h^{(N)} v^{(1)} \dots v^{(N)} | \vartheta) p_0(\vartheta)$$

$$\overset{iid}{=} \left\{ \prod_{n=1}^{N} p(h^{(n)}, v^{(n)} | \vartheta) \right\} p_0(\vartheta).$$

We then get:

$$\log p(v^{(1)} \dots v^{(N)}) = KL\left(\prod_{n=1}^{N} q_n(h^n)\, q(\vartheta) \,//\, p(h^{(1)} \dots h^{(N)} \vartheta | v^{(1)} \dots v^{(N)}) \right)$$

$$+ \sum_{n=1}^{N} \mathbb{E}_{q_n(h^n) q(\vartheta)} \left[\log p(h^n, v^n | \vartheta) \right] \quad \leftarrow \text{energy term}$$

$$- \sum_{n=1}^{N} \mathbb{E}_{q_n(h^n)} \left[\log q_n(h^n) \right] \quad \leftarrow \text{entropy term.}$$

$$- \mathbb{E}_{q(\vartheta)} \left[\log \frac{q(\vartheta)}{p_0(\vartheta)} \right]$$

The variational bound simply follows from $KL \geq 0$.

■

II. Optimization of lower bound and the variational Bayes EM algorithm.

We do a "coordinate-wise" optimization over $\{q_m(h),\ m=1\cdots N\}$ and $q(\theta)$ of the lower bound (on page 55).

E-step: **Fix $q(\theta)$ and optimize over $\{q_m(h),\ m=1\cdots N\}$**

We look at terms which depend on q_m:

$$(*) = \sum_{m=1}^{N} \underbrace{\mathbb{E}_{q_m(h^m)\cdot q(\theta)}\left[\log p(h^m, v^m, \theta)\right]} - \mathbb{E}_{q_m(h^m)}\left[\log q_m(h^m)\right]$$

$$\mathbb{E}_{q(\theta)}\left[\log p(h^m, v^m, \theta)\right] = \log\left\{\frac{1}{Z}\exp \mathbb{E}_{q(\theta)}\left[\log p(h^m, v^m, \theta)\right]\right\} + \log Z$$

Thus:

$$(*) = \sum_{m=1}^{N}\left\{\mathbb{E}_{q_m(h^m)}\left[\log \frac{1}{Z}\exp \mathbb{E}_{q(\theta)}\left[\log p(h^m, v^m | \theta)\right]\right] - \mathbb{E}_{q_m(h^m)}\left[\log q_m(h^m)\right]\right\}$$

$$= -\sum_{m=1}^{N} KL\left(q_m(h^m)\ \Big\|\ \frac{1}{Z}\exp \mathbb{E}_{q(\theta)}\left(\log p(h^m v^m | \theta)\right)\right).$$

Therefore to maximize (*) over q_m we have to minimize the KL, i.e, we set:

$$q_m(h^m) = \frac{1}{Z} \exp \mathbb{E}_{q(\theta)} \left[\log p(h^m v^m \mid \theta) \right]$$

where Z = Normalization factor

$$= \sum_{h^m} \exp \mathbb{E}_{q(\theta)} \left[\log p(h^m v^m \mid \theta) \right].$$

<u>The E-step is an iterative step</u>; at time t we do

$$\begin{cases} q_m^{t+1}(h^m) = \dfrac{1}{Z_t} \exp \mathbb{E}_{q_t(\theta)} \left[\log p(h^m, v^m \mid \theta) \right] \\ Z_t = \displaystyle\sum_{h^m} \exp \mathbb{E}_{q_t(\theta)} \left[\log p(h^m, v^m \mid \theta) \right] \end{cases}$$

M-step: Fix $\{q_m(h^m), m=1 \dots N\}$ and optimize over $q(\theta)$.

We must maximize terms in lower bound (on page 55) which depend on $q(\theta)$:

$$(**) = \sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)\, q(\theta)} \left[\log p(h^m, v^m | \theta) \right] - \mathbb{E}_{q(\theta)} \left[\log \frac{q(\theta)}{p_0(\theta)} \right]$$

$$= - \mathbb{E}_{q(\theta)} \{ \log q(\theta) \} + \mathbb{E}_{q(\theta)} \Big\{ \underbrace{\sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \left[\log p(h^m v^m | \theta) \right] + \log p_0(\theta)}_{} \Big\}.$$

$$+ \log \tilde{Z} + \log \left[\frac{1}{\tilde{Z}} \exp \left\{ \sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \left[\log p(h^m v^m | \theta) \right] + \log p_0(\theta) \right\} \right]$$

$$= + \log \tilde{Z} + \log \left[\frac{p_0(\theta)}{\tilde{Z}} \exp \left\{ \sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \left[\log p(h^m v^m | \theta) \right] \right\} \right]$$

with the normalization factor

$$\tilde{Z} = \int d\theta\, p_0(\theta) \exp \left[\sum_{m=1}^{N} \mathbb{E}_{q_m(h^m)} \log p(h^m v^m | \theta) \right].$$

Note that $\tilde{Z}$ is independent of $q(\theta)$. We see that maximizing (**) over $q(\theta)$ is equivalent to maximizing

$$-\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right] + \mathbb{E}_{q(\theta)}\left[\log \frac{p_0(\theta)}{\tilde{Z}} \prod_{n=1}^{N} \exp\left(\mathbb{E}_{q_n(h^n)} \log p(h^n v^n|\theta)\right)\right]$$

$$= -KL\left(q(\theta) \,\|\, \frac{p_0(\theta)}{\tilde{Z}} \prod_{n=1}^{N} \exp\left(\mathbb{E}_{q_n(h^n)} \log p(h^n v^n|\theta)\right)\right)$$

This is achieved by setting

$$q(\theta) = \frac{p_0(\theta)}{\tilde{Z}} \prod_{n=1}^{N} \exp\left(\mathbb{E}_{q_n(h^n)} \log p(h^n v^n|\theta)\right).$$

M - step iteration: at time t we do:

$$\begin{cases} q_{t+1}(\theta) = \dfrac{p_0(\theta)}{\tilde{Z}_{t+1}} \displaystyle\prod_{n=1}^{N} \exp\left(\mathbb{E}_{q_n^{t+1}(h^n)} \log p(h^n v^n|\theta)\right) \\[2ex] \tilde{Z}_{t+1} = \displaystyle\int d\theta\, p_0(\theta) \exp\left\{\sum_{n=1}^{N} \mathbb{E}_{q_n^{t+1}(h^n)} \log p(h^n v^n|\theta)\right\} \end{cases}$$

<u>Remarks:</u> The iterative E & M steps can only increase the lower bound and will converge to some local maximum. There is no guarantee that the global maximum of the lower bound is found. Also, there is no guarantee that the log-likelihood $\log p(v^{(1)} \dots v^{(N)})$ increases along E & M steps.

For the simplest possible variational class $q(\theta) = \delta_{\theta, \theta_*}$ where we assume that θ takes a single value θ_* and with a flat prior $p_0(\theta) \sim$ constant the algorithm reduces to the <u>standard EM algorithm</u>. We will show that in this case not only the lower bound will only increase (until we reach a local maximum) but remarquably <u>also the likelihood itself increases.</u> This is a very nice feature of the EM standard algo-

III. Reduction to standard EM algorithm.

If we assume that θ is characterized by not a full distribution but by a single numbers θ_t we set $q_t(\theta) = \delta_{\theta, \theta_t}$ and the E-step becomes:

$$\begin{cases} q_n^{t+1}(h^n) = \dfrac{1}{Z_t} \cdot p(h^n, v^n \mid \theta_t) = p(h^n \mid v^n \theta_t). \\ Z_t = \sum_{h^n} p(h^n, v^n \mid \theta_t). \end{cases}$$

To get the M-step we note that $\theta_{t+1} = \arg\max_{\theta} \delta_{\theta, \theta_{t+1}}$ so with $p_0(\theta) = \text{constant}$:

$$\theta_{t+1} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{q_n^{t+1}(h^n)} \left[\log p(h^n, v^n \mid \theta) \right].$$

Standard EM algorithm:

E-step : $q_n^{t+1}(h^n) = p(h^n \mid v^n \theta_t)$

M-step : $\theta_{t+1} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{q_n^{t+1}(h^n)} \left[\log p(h^n v^n \mid \theta) \right]$

A closer look at EM lower bound:

In standard ML training we seek to maximize $\log p(v^{(1)} \dots v^{(N)} | \vartheta)$. Since we do not have access to hidden variable and cannot estimate $p(h^{(1)} \dots h^{(N)} | v^{(1)} \dots v^{(N)} \vartheta)$ we replace it by the variational class $\prod_{n=1}^{N} q_n(h^n)$.

Starting from

$$KL\left(\prod_{n=1}^{N} q_n(h^n) \,\|\, p(h^{(1)} \dots h^{(N)} | v^{(1)} \dots v^{(N)} \vartheta) \right) \geq 0$$

we get by identical (and simpler) calculations than before the variational lower bound (iid samples):

$$L(\vartheta) = \log p(v^{(1)} \dots v^{(N)} | \vartheta) = KL\left(\prod_{n=1}^{N} q_n(h^n) \,\|\, p(h^{(1)} \dots h^{(N)} | v^{(1)} \dots v^{(N)} \vartheta) \right)$$

$$+ \sum_{n=1}^{N} \mathbb{E}_{q_n(h^n)} \left[\log p(h^n, v^n | \vartheta) \right] \leftarrow \text{energy term}$$

$$- \sum_{n=1}^{N} \mathbb{E}_{q_n(h^n)} \left[\log q_n(h^n) \right] \leftarrow \text{entropy term}$$

$$\geq (\text{energy term}) + (\text{entropy term}).$$

We have the nice properties:

(i) During EM steps lower bound only increases until it reaches a local stat pt. This is obvious because E M are coordinate-wise optimization of lower bound

(ii) Less obviously, also $\log p(v^{(1)} \ldots v^{(N)} | \theta_t) = L(\theta_t)$ increases ($\rightsquigarrow$ popular algo.).

Proof of (ii).

Note that the formula for $L(\theta_t)$ is valid for any variational distr $q_n(h^n)$ $\rightarrow$ so it is valid when we use $q_n^t(h^n)$ as well as $q_n^{t+1}(h^n)$!

At time t of algorithm:

$L(\theta_t) = \text{expression}(\theta_t, q_n^t) = \text{expression}(\theta_t, q_n^{t+1})$

but here E-step: $p_n(h^n | v^n \theta_t) = q_n^{t+1}$

Thus

$KL\left(\prod_{n=1}^{N} q_n^{t+1} \,\|\, p(h^1 \ldots h^N | v^1 \ldots v^N \theta)\right) = 0$

$\Rightarrow$ $\text{expression}(\theta_t, q_n^{t+1}) = \text{expression of Lower Bound}(\theta_t, q_n^{t+1})$

(M-step) $\rightsquigarrow$ $\leq \text{expression of Lower Bound}(\theta_{t+1}, q_n^{t+1})$

$\leq L(\theta_{t+1})$.