Lecture 4. Free Energies & Mean Field Theory.

Note Title 1/17/2010

The next few lectures describe methods for performing inference on MRF's.
This lecture introduces free energies and mean field theory approximations to them.

But first, we describe steepest descent. (many variables here in the subscript.)
If we have a function $E[\underline{x}]$ then we can perform steepest descent by

$$\underline{x}^{t+1} = \underline{x}^t - \epsilon \underline{\nabla} E[\underline{x}^t]$$

provided $E[\underline{x}]$ is differentiable – i.e. the state variables $\underline{x}$ take continuous values.
In MRFs, the variables are often discrete but can be treated as if they are continuous. For example, in the weak membrane model the ω_i take values $0, 1, \ldots, 255$. But it is convenient to approximate them as continuous values. Alternatively we can think of the ω_i being real values (this can give technical problems for some distributions – but we ignore this.)

For the MRF $P(\underline{\omega}|I) = \frac{1}{Z} e^{-\left\{\sum_i \phi_i(\omega_i, I) + \sum_{i,j} \psi_{ij}(\omega_i, \omega_j)\right\}}$

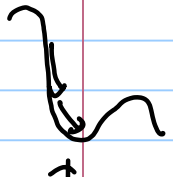
$$\hat{\underline{\omega}} = \arg\max_{\underline{\omega}} P(\underline{\omega}|I)$$

is equivalent to minimizing $E[\underline{\omega}] = \sum_i \phi_i(\omega_i, I) + \sum_{i,j} \psi_{ij}(\omega_i, \omega_j)$

Steepest descent: $\underline{\nabla}_i E[\underline{\omega}] = \frac{\partial \phi_i}{\partial \omega_i} + \sum_{j \in \mathcal{N}(i)} \frac{\partial \psi_{ij}}{\partial \omega_i}(\omega_i, \omega_j)$

$$\omega_i^{t+1} = \omega_i^t - \epsilon \underline{\nabla}_i E[\underline{\omega}^t]$$

Steepest descent is guaranteed to decrease the energy monotonically – provided ϵ is small enough.
The problems are:
(i) how to pick the best size of ϵ.
Too small – slow convergence, Too large – unstable algorithm
(ii) algorithm hugs the energy surface – it cannot make large moves or avoid even small local minima



(2) Variational Free Energies:

$p(\underline{x}|I)$ → suppose we want to estimate the marginals $p(x_i|I) = \sum_{x_j : j \neq i} p(\underline{x}|I)$.

(Later we will extend this to estimate the $\hat{\underline{x}} = \text{ARG MAX}_{\underline{x}} p(\underline{x}|I)$)

Define pseudo marginals $b_i(x_i)$ $\quad b_i(x_i) \geq 0$, $\sum_{x_i} b_i(x_i) = 1$

these are 'variables' that will be used to approximate the $p(x_i|I)$.

Full distribution $B(\underline{x}) = \prod_{i \in D} b_i(x_i)$

Define a similarity measure between $B(\underline{x})$ and the full distribution $p(\underline{x}|I)$

Kullback-Leibler divergence

$$KL(B,P) = \sum_{\underline{x}} B(\underline{x}) \log \frac{B(\underline{x})}{p(\underline{x}|I)}$$

Propation $KL(B,P) \geq 0$, $= 0$, only if $B(\underline{x}) = p(\underline{x}|I)$.

Note: $B^*(\underline{x})$ which minimizes $KL(B,P)$ is the factorizable distribution which best approximates $p(\underline{x}|I)$.

Substituting the form of $B(\underline{x})$ into $KL(B,P)$ yields

$$KL(B,P) = \mathcal{F}_{MFT}(B) + \log Z.$$

where the 'free energy' $\mathcal{F}_{MFT}(B)$ is:

$$\mathcal{F}_{MFT}(B) = \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j} b_i(x_i) b_j(x_j) \psi_{ij}(x_i, x_j) + \sum_{i \in V} \sum_{x_i} b_i(x_i) \phi_i(x_i, I) + \sum_i \sum_{x_i} b_i(x_i) \log b_i(x_i)$$

First term is $\sum_{\underline{x}} B(\underline{x}) E(\underline{x}, I)$ Second is $\sum_{\underline{x}} B(\underline{x}) \log B(\underline{x})$

expected energy — negative entropy

Minimizing $\mathcal{F}_{MFT}(B)$ w.r.t. B gives

(i) an upper bound for $\log Z \geq -\mathcal{F}_{MFT}(B^*)$ (relates to model evidence)

(ii) the best factorized approximation to $p(\underline{x}|I)$

(iii) an estimate of the marginals of the distribution — which relate to the estimate $\hat{\underline{x}} = \text{ARG MAX}_{\underline{x}} p(\underline{x}|I)$

(3) We still need an algorithm to minimize the free energy. Also the free energy function might have local minima so it is possible that the algorithm gets trapped in a local minimum.

But, the free energy formulation replaces a function $E[\underline{x};\underline{I}]$ of discrete variables by a function $\mathcal{F}[B]$ of continuous variables. This enables us to compute derivatives, do steepest descent and other algorithms (see later).

For the weak membrane model we treat the variables $\{w_{ij}\}$ as if it is continuous, and define pseudomarginals over the process variables $\{y_{ij}\}$

$$\mathcal{F}_{EM}(B,\underline{x}) = -\sum_{\underline{y}} B(\underline{y}) \log P(\underline{x},\underline{y}|\underline{I}) + \sum_{\underline{y}} B(\underline{y}) \log B(\underline{y})$$

Also expected energy + negative entropy.

It can be shown (easily) that if we can minimize $\mathcal{F}_{EM}(B,\underline{x})$ wrt. $B,\underline{x}$ to get $(B^*,\underline{x}^*) = \underset{(B,\underline{x})}{\text{ARG MAX}}\ \mathcal{F}_{EM}(B,\underline{x})$

then $\underline{x}^* = \underset{\underline{x}}{\text{ARG MAX}}\ P(\underline{x}|\underline{I})$ with $P(\underline{x}|\underline{I}) \propto \sum_{\underline{y}} P(\underline{x},\underline{y}|\underline{I})$

This is the theoretical background for the Expectation Maximization (EM) algorithm.

The E-step of EM is equivalent to minimizing $\mathcal{F}_{EM}(B,\underline{x})$ with respect to B

M-step is equivalent to minimizing $\mathcal{F}_{EM}(B,\underline{x})$ with respect to $\underline{x}$

This free energy formulation shows that the EM algorithm is guaranteed to converge since each step reduces the free energy.

But this formulation is richer than standard EM because it also shows convergence for any algorithm which reduces the free energy at each time step.

Note: this still does not guarantee convergence to the global minimum.

(4) Free energy for the weak membrane.

$$F_T(\{b_{ij}\}) = T\sum_i (x_i - I_i)^2 + A\sum_{ij}(1-b_{ij})(x_i - x_j)^2$$
$$+ B\sum_{ij} b_{ij} + \sum_{ij}\{b_{ij}\log b_{ij} + (1-b_{ij})\log(1-b_{ij})\}$$

where $b_{ij} = b_{ij}(y_{ij}=1)$.

The M-step:

min w.r.t. $\{x_i\}$ $\quad T\sum_i (x_i - I_i)^2 + A\sum_{ij}(1-b_{ij})(x_i - x_j)^2$.

quadratic problem – solution by linear algebra.

note: non-zero b_{ij} weakens the smoothing between x_i & x_j. if $b_{ij}=1$, then no smoothing occurs

E-step:

min w.r.t. b_{ij} → give analytic solution

$$b_{ij} = \frac{1}{1+e^{\{-A(x_i-x_j)^2+B\}}}$$

if $(x_i - x_j)$ is large then $b_{ij} \to 1$
$|x_i - x_j|$ is small then $b_{ij} \to 0$

The iterative algorithm is like smoothing the input $\{I_i\}$, but adapting the weights of the smoothness functions (the b_{ij}) so that they become large (close to 1) if $|x_i - x_j|$ is big.

Note: initial conditions must be well chosen to help converge to the global minimum.

E.G. initialize so that $x_i = I_i$ at time $t=0$, then update the b_{ij}... → likely to converge (but no guarantees).

Note: it is unusual that the steps of an EM algorithm are so simple.

In particular, the simplicity of the E-step is because the distribution $P(\underline{x}, \underline{y}|I)$ is factorizable in the $\underline{y}$-variables.

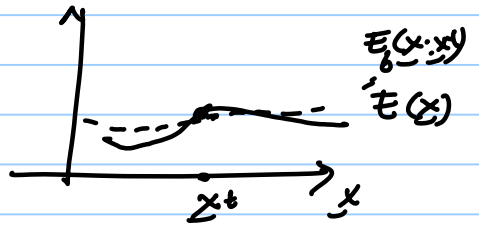
Suppose we modify the energy $E(\underline{x}, \underline{y}; I)$ by adding terms like $C\sum y_{ij} y_{kl}$, which encourages edges to be continuous spatially.

In this case, we cannot solve the E-step analytically but we can modify the free energy to approximate $P(\underline{y})$ by a factorized distribution.

(5) # Discrete Iterative Algorithms

We give procedures to construct iterative algorithms that can decrease energy functions by making non-local steps and without needing a step size parameter (as for steepest descent).

Variational Bounding



current state $\underline{x}^t$

define a bounding energy $E_b(\underline{x}:\underline{x}^t)$ such that:

$$E_b(\underline{x}^t:\underline{x}^t) = E(\underline{x}^t)$$

& $E(\underline{x}) \leq E_b(\underline{x}:\underline{x}^t) \quad \forall \underline{x}$

Then choose $\underline{x}^{t+1}$ s.t. $E_b(\underline{x}^{t+1}:\underline{x}^t) \leq E_b(\underline{x}^t:\underline{x}^t)$

which implies $E(\underline{x}^{t+1}) \leq E(\underline{x}^t)$.

Intuition → pick the bounding energy $E_b(\underline{x}:\underline{x}^t)$ so that it is easy to minimize

CCCP → Concave Convex Procedure.

This is a special case of variational bounding which gives a special set of bounding energies.

Claim: for almost all energies, $E(\underline{x})$, we can write it as $E(\underline{x}) = E_{vex}(\underline{x}) + E_{conc}(\underline{x})$ → seems paradoxical

(to prove, consider the Hessians of E, E_{vex}, E_{conc})

$$\nabla\nabla E = \nabla\nabla E_{vex} + \nabla\nabla E_{conc}$$

for any $\nabla\nabla$.

Then $E_b(\underline{x}:\underline{x}^t) = E_{vex}(\underline{x}) + E_{conc}(\underline{x}^t) + (\underline{x} - \underline{x}^t) \cdot \frac{\partial E_{conc}(\underline{x}^t)}{\partial \underline{x}}$

this is a convex function - so usually possible to minimize it.

corresponds to an update $\frac{\partial E_{vex}(\underline{x}^{t+1})}{\partial \underline{x}} = -\frac{\partial E_{conc}(\underline{x}^t)}{\partial \underline{x}}$

Note: this is a non-local method, so it may avoid local minima that a steepest descent algorithm can get trapped in. But no guarantee that you get to the local minima.

(6) Note: Researchers have invented Discrete Iterative Algorithms for problems on a one by one basis — E.M., Generalized Iterative Scaling, Sinkhorn — but this can be shown to all be equivalent to CCCP.

Example: $E[\omega] = \frac{1}{2}\sum_{ij} T_{ij}\omega_i\omega_j - \sum_i \theta_i\omega_i$ $\quad \omega_i \in [0,1]$

Mean Field Theory

$$\mathcal{T}[b] = \frac{1}{2}\sum_{ij} T_{ij} b_i b_j + \sum_i \theta_i b_i + \sum_i \{ b_i \log b_i + (1-b_i)\log(1-b_i)\}$$

Assume T_{ij} is negative definite matrix

convex: $\sum_i \{ b_i \log b_i + (1-b_i)\log(1-b_i)\}$; concave: $\frac{1}{2}\sum_{ij} T_{ij} b_i b_j + \sum_i \theta_i b_i$

$$\frac{\partial \mathcal{T}_{vex}}{\partial b_i} = \log\frac{b_i}{1-b_i}, \quad \frac{\partial \mathcal{T}_{cave}}{\partial b_i} = \sum_j T_{ij} + \theta_i .$$

yields $$b_i^{t+1} = \frac{e^{-\{\sum_j T_{ij} b_j^t + \theta_i\}}}{1 + e^{-\{\sum_j T_{ij} b_j^t + \theta_i\}}} .$$

non-local MA.

Note: steepest descent can be derived as a special case.

$$E(x) = \frac{1}{2}Ax^2 + E(x) - \frac{1}{2}Ax^2$$

convex; concave $\Rightarrow$ if A sufficiently large

$$A x^{t+1} = -\frac{\partial E}{\partial x}(x^t) + A x^t$$

$$x^{t+1} = x^t - \frac{1}{A}\frac{\partial E}{\partial x}(x^t) \qquad \epsilon = \frac{1}{A}$$

gives a size for the ϵ-parameter (to ensure convexity of $E(x) - \frac{1}{2}Ax^2$)

this can be used to help design stable steepest descent algorithms (e.g. Eyre's method)