4/3/14

MRF INFERENCE: (FINISHING) VARIATIONAL INFERENCE

1) Variational Inference: the "incorrect" direction

$$\underbrace{p(\vec{x}) = \frac{1}{Z} \prod_c \psi_c(x_c)}_{\substack{\text{General Gibbs distribution} \\ \text{with Factors } F = \{\psi_c(x_c)\}}} \qquad q(\vec{x}) = \prod_i q_i(x_i)$$

Goal: $\underset{q}{\operatorname{argmin}} \; KL(q \| p) = \sum_{\vec{x}} q(\vec{x}) \log \frac{q(\vec{x})}{p(\vec{x})}$

$$= \underbrace{\sum_{\vec{x}} q(\vec{x}) \log q(\vec{x})}_{-H_q(\vec{X})} - \sum_{\vec{x}} q(\vec{x}) \log p(\vec{x})$$

term #1: $-H_q(\vec{X}) = -\sum_i H_{q_i}(X_i)$ [Why? $\because q \models X_i \perp X_j$]

$$= \sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i)$$

term #2: $\sum_{\vec{x}} q(\vec{x}) \log p(\vec{x}) = \sum_{\vec{x}} q(\vec{x}) \log \left[\frac{1}{Z} \prod_c \psi_c(x_c) \right]$

$$= \sum_{\vec{x}} q(\vec{x}) \sum_c \log \psi_c(x_c) - \left(\xcancel{\sum_{\vec{x}} q(\vec{x})} \right)^{1} \log Z$$

$\log \psi_c(x_c)$ are log-potentials or log-factors. They will play a very important role in MAP Inference later where it is common to always work in log-domain.

Notation $\theta_c(x_c) \equiv \log \psi_c(x_c)$

So in a pairwise MRF $\theta_i(x_i) \equiv \log \psi_i(x_i)$
& $\theta_{ij}(x_i, x_j) \equiv \log \psi_{ij}(x_i, x_j)$

Node to edge "Scores"

In some papers people also use negative-log-potentials or "Energies". All algorithms have equivalent formulations in "Scores" & "Energies"

$$S(\vec{x}) = \sum_c \theta_c(x_c) = \sum_c \log \psi_c(x_c)$$
$$E(\vec{x}) = \sum_c E_c(x_c) = \sum_c -\log \psi_c(x_c)$$

Back to term #2

$$= \sum_{\vec{x}} q(\vec{x}) \sum_c \theta_c(x_c) - \log Z$$

$$= \sum_c \sum_{\vec{x}} q(\vec{x}) \theta_c(x_c) - \log Z$$

$$= \sum_c \underbrace{\sum_{x_c} q_c(x_c)}\, \theta_c(x_c) - \log Z$$

where $q_c(x_c) = \prod_{i \in c} q_i(x_i)$

Putting it all together

$$KL(q \| p) = -H_q(\vec{X}) - \sum_c \sum_{x_c} q_c(x_c)\,\theta_c(x_c) + \log Z$$

$$\Rightarrow \log Z = KL(q \| p) + \underbrace{H_q(\vec{X}) + \sum_c \sum q_c(x_c)\,\theta_c(x_c)}_{\text{Called "Gibbs Free Energy" } F[p,q]}$$

$$\Rightarrow \log Z = KL(q \| p) + F[p,q]$$

Note 1: $\log Z$ = constant w.r.t q

$$\Rightarrow \arg\min_q KL(q \| p) = \arg\max_q F[p,q]$$

Note 2: $KL(q \| p) \geq 0$ [why? See HW0 in Intro to ML]

$$\Rightarrow \log Z \geq F[p,q]$$

Ah, so we're maximizing a lower-bound on $\log Z$!

Note 3: Maximizing $F[p,q]$ "makes sense"

$$F[p,q] = H_q(\vec{X}) + \sum_c \underbrace{\sum_{x_c} q_c(x_c)\,\theta_c(x_c)}_{E_{q_c}[\theta_c(X_c)]}$$

So find me a simple q that puts heavy mass on high scoring configurations AND is as uncertain as possible!

Note 4: There will usually be a trade-off

If q is not simple enough, we can always get $KL(q \| p) = 0$ by setting $q = p$
so lower-bound is tight
but can't perform $\max_q F[p,q]$ with

complicated distributions!

2 Mean-Field Update Rules

So how do we maximize $F[p,q]$?
Let's write the problem:

$$\max_{\{q_i\}} -\sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i) + \sum_c \sum_{x_c} \prod_{i \in c} q_i(x_i)\, \theta_c(x_c)$$

$$\text{s.t} \qquad \sum_{x_i} q_i(x_i) = 1 \quad : \lambda_i \qquad \forall i$$

$$q_i(x_i) \geq 0$$

Lagrangian Multiplier

Lagrangian Multiplier Method!

[Assume pairwise MRF for simplicity. Derivation generalizes to arbitrary MRFs]

$$\Rightarrow \sum_c \sum_{x_c} q_c(x_c)\theta_c(x_c) = \sum_i \sum_{x_i} q_i(x_i)\theta_i(x_i)$$

$$+ \sum_{(i,j) \in E} \sum_{x_i} \sum_{x_j} q_i(x_i) q_j(x_j)\, \theta_{ij}(x_i, x_j)$$

$$\mathcal{L}(\{q_i(\cdot)\}, \{\lambda_i\}) = -\sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i)$$
$$+ \sum_i \sum_{x_i} q_i(x_i)\,\theta_i(x_i)$$
$$+ \sum_{ij} \sum_{x_i} \sum_{x_j} q_i(x_i)\, q_j(x_j)\, \theta_{ij}(x_i, x_j)$$
$$- \sum_i \lambda_i \Big[\sum_{x_i} q_i(x_i) - 1 \Big]$$

Now, $\dfrac{\partial \mathcal{L}}{\partial q_s(x_s)} = 0$ [to find stationary points]

$$\Rightarrow \quad -\frac{q_s(x_s)}{q_s(x_s)} - \log q_s(x_s)$$
$$+ \theta_s(x_s) \qquad\qquad = 0$$
$$+ \sum_{t \in N(s)} \sum_{x_t} q_t(x_t)\, \theta_{st}(x_s, x_t)$$
$$- \lambda_s$$

$$\Rightarrow \quad \log q_s(x_s) = -1 - \lambda_s + \theta_s(x_s) + \sum_{t \in N(s)} \sum_{x_t} q_t(x_t)\, \theta_{st}(x_s, x_t)$$

$$\Rightarrow q_s(x_s) = \underbrace{e^{-1-\lambda_s}}_{\substack{\text{constant} \\ \text{w.r.t } x_s}} \cdot \underbrace{e^{\theta_s(x_s)}}_{\psi_s(x_s)} \cdot \underbrace{\prod_{t \in N(s)} e^{\sum_{x_t} \theta_{st}(x_s, x_t)\, q_t(x_t)}}_{\text{"messages"}}$$

$$\Rightarrow \boxed{q_s(x_s) = \frac{1}{Z_s} \psi_s(x_s) \cdot \prod_{t \in N(s)} e^{\sum_{x_t} \theta_{st}(x_s, x_t) q_t(x_t)}}$$

~~The~~ This is a fixed point equation.
Any q satisfying this is a fixed point.

How do we reach such a q?
Iterative update rules:

$$q_s^{(t+1)}(x_s) = \frac{1}{Z_s} \psi_s(x_s) \cdot \prod_{t \in N(s)} e^{\sum_{x_t} \theta_{st}(x_s, x_t) q^{(t)}(x_t)}$$

This is guaranteed to converge to a fixed point.
But that fixed point may not be the optimum maxima! Non-convex objective.

Update rule looks like BP but isn't.

Need a different assumption on q to get BP updates!
[Advanced topic: won't cover in class]
Maybe Final? :)