Variational Bayes

- find analytic approximation to intractable posterior.

$p(x|y)$ - exact/true posterior

$q(x)$ - approximate posterior

Idea: 1. can we find parameters of $q(x)$, called variational parameters, such that $q(x)$ is as similar to $p(x)$ as possible?

2. Find parameters such that the KL-divergence

$$KL(q(x)\|p(x|y)) = \int q(x) \cdot \log \frac{q(x)}{p(x|D)} dx =$$

$$= \left\{ p(x|y) = \frac{p(x,y)}{p(y)} \right\} =$$

$$= \int q(x) \cdot \log \frac{q(x) \cdot p(y)}{p(x,y)} dx =$$

$$= \int q(x) \cdot \log \frac{q(x)}{p(x,y)} dx + \underbrace{\int q(x) \cdot dx}_{1} \cdot \log p(y) =$$

$$= \int q(x) \cdot \log \frac{q(x)}{p(x,y)} dx + \log p(y)$$

$$\Rightarrow \log p(y) = KL\left(q(x) \| p(x|y)\right) - \int q(x) \cdot \log \frac{q(x)}{p(x,y)} dx =$$

$$= KL\left(q(x) \| p(x|y)\right) + \int q(x) \cdot \log \frac{p(x,y)}{q(x)} dx$$

$$\Rightarrow \log p(y) \geq \int q(x) \cdot \log \frac{p(x,y)}{q(x)} dx = \mathcal{L}(q(x))$$

$\Rightarrow$ maximise $\mathcal{L}(q(x))$ to minimise KL

equality only when $KL(q(x) \| p(x|y)) = 0$

$\Rightarrow q(x) = p(x|y)$

Mean Field

$$q(x) = \prod_i q_i(x_i)$$

$\Rightarrow$ we fit the marginals of the distribution

$$\mathcal{L}(q) + KL(q \| p) = \log p(y)$$

$$\Rightarrow \quad \mathcal{L}(q) \leq \log p(y)$$

$$\mathcal{L}(q) = \int q(x) \cdot \log \frac{p(y,x)}{q(x)} \, dx = \left\{ q(x) = \prod_i q_i(x_i) \right\} =$$

$$= \int \prod_i q(x_i) \cdot \log \frac{p(y,x)}{\prod_k q_k(x_k)} \, dx =$$

$$= \int \prod_i q(x_i) \cdot \left(\log p(y,x) - \sum_k \log q_k(x_k) \right) dx =$$

We want to have a scheme where we optimise each component in turn therefor we would like to re-write $\mathcal{L}(q)$ to single out each component as $\mathcal{L}(q_j) = \mathcal{L}'(q_j) + \mathcal{L}(q_{\neg j})$

$$\mathcal{L}(q_j) = \int \prod_i q(x_i) \cdot \left(\log p(y,x) - \sum_k \log q_k(x_k) \right) dx =$$

$$= \{\text{split integral}\} =$$

$$= \int_j \int_{\neg j} q_j(x_j) \prod_{i \neq j} q_i(x_i) \left(\log p(x,y) - \sum_k \log q_k(x_k) \right) dx_{\neg j} dx_j$$

$$= \{\text{move integral inside and collect terms}\} =$$

$$= \int_j q_j(x_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(x_i) \cdot \log p(y,x) \, dx_{\neg j}}_{\log f_j(x_j)} dx_j -$$

$$- \int_j q_j(x_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(x_i) \left(\sum_{k \neq j} \log q_k(x_k) + \log q_j(x_j) \right) dx_{\neg j}}_{\log q_j(x_j) \cdot \underbrace{\int \prod_{i \neq j} q_i(x_i) \, dx_{\neg j}}_{1} \; + \; \underbrace{\int_{\neg j} \prod q_i(x_i) \sum_{k \neq j} \log q_k(x_k) \, dx_{\neg j}}_{\text{const. when we fit } q_j}} dx_j =$$

$$= \int_j q_j(x_j) \log f_j(x_j) \, dx_j - \int_j q_j(x_j) \cdot \log q_j(x_j) \, dx_j + \text{const.} =$$

$$= \int_j q_j(x_j) \cdot \log \frac{f_j(x_j)}{q_j(x_j)} \, dx_j + \text{const.}$$

$$= \underbrace{-\int q_j(x_j) \cdot \log \frac{q_j(x_j)}{f_j(x_j)} \, dx_j}_{KL(q_j(x_j) \| f_j(x_j))} + \text{const.}$$

$$\Rightarrow -KL(q_j(x_j) \| f_j(x_j)) + c$$

$\Rightarrow$ KL - divergence between the factor $q_j(x_j)$ and the distribution when all other factors have been averaged out.

$\Rightarrow$ maximise $\mathcal{L}(q_j(x_j))$ by minimising $KL(q_j(x_j) \| f_j(x_j))$

- KL - divergence always positive

  $\Rightarrow$ minimum $= \emptyset$

  $KL(q_j(x_j) \| f_j(x_j)) = 0$

  $\Rightarrow q_j(x_j) = f_j(x_j)$

$$\log f_j(x_j) = \int_{\mathcal{J}} \underbrace{\prod_{i \neq j} q_i(x_i)}_{q_{\mathcal{J}}(x_{\mathcal{J}})} \log p(x,y) \, dx_{\mathcal{J}} =$$

$$= \mathbb{E}_{q_{\mathcal{J}}(x_{\mathcal{J}})}\left[\log p(x,y)\right]$$

- We need to pick q so that we can compute the above expectation

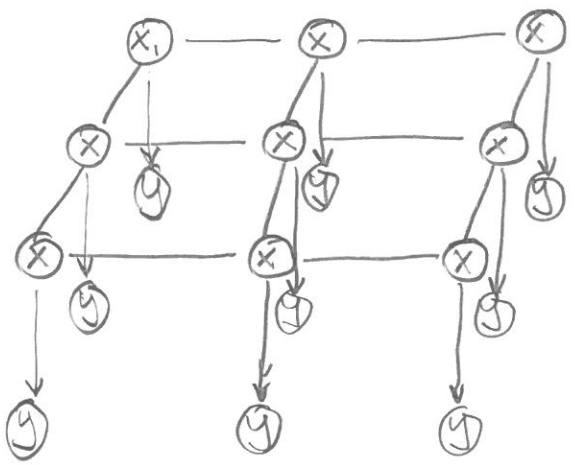
$$q(x_{\bar{j}}) = \frac{1}{Z} \cdot \exp\left(\mathbb{E}_{q_{\bar{j}}(x_{\bar{j}})}\left[\log p(x,y)\right]\right)$$

as we know that $q(x_{\bar{j}})$ has to be a distribution we simply work in log-space

$$\Rightarrow \log(q(x_{\bar{j}})) = \mathbb{E}_{q_{\bar{j}}(x_{\bar{j}})}\left[\log p(x,y)\right] + \text{const.}$$

Example: Markow Random Field

Binary Image



$y \in [0,1]$

$x \in \{-1, 1\}$

x - is the binary image

y - is the noisy observations

Joint distribution: $p(x,y) = p(y|x)p(x)$

prior: $p(x) = \frac{1}{Z_0} \cdot e^{-E_0(x)}$

$$E_0(x) = -\sum_{i}^{D} \sum_{j \in N(i)} w_{ij} x_i x_j$$

likelihood: $p(y|x) = \prod_i p(y_i|x_i) = \sum_i e^{-L_i(x_i)}$

- Find approximate fully factorised posterior over the latent representation.

$q(x) = \prod_i q(x_i, \mu_i)$

variational parameter

Write down joint distribution

$$\log p(x,y) = \log p(y|x)p(x) =$$

$$= \log\left(\prod_i e^{-L_i(x_i)} \cdot \frac{1}{Z_0} e^{\sum_i \sum_{j\in N(i)} w_{ij} x_i x_j}\right) =$$

$$= \sum_i L_i(x_i) + \sum_i \sum_{j\in N(i)} w_{ij} x_i x_j + \text{const.} =$$

$$= \left\{\begin{array}{l}\text{If we are only seeing each term} \\ \text{in turn we can make the non } x_i \\ \text{terms into the const}\end{array}\right\} =$$

$$= L_i(x_i) + x_i \sum_{j\in N(i)} w_{ij} x_j + \text{const.}$$

Lets compute the expectation to get the update.

$$\log q_i(x_i) = \log f_i(x_i) = \int_{\neg i} \prod_{j\neq i} q_i(x_j) \cdot \log p(x, D)\, dx_{\neg i} =$$

$$= \left\{\mathbb{E}_{q_j}[x_j] = \mu_j\right\} =$$

$$= \int_{j\neq i} q_j(x_j)\left(L_i(x_i) + x_i \sum_{j\in N(i)} w_{ij} x_j + \text{const.}\right) dx_{\neg j} =$$

$$= \left\{\mathbb{E}[ax] = a\cdot\mathbb{E}[x]\right\} = L_i(x_i) + x_i \cdot \underbrace{\sum_{j\in N(i)} w_{ij}\mu_j}_{m_i} + \text{const.}$$

$$\Rightarrow q_i(x_i) \propto e^{x_i \cdot m_i + L_i(x_i)}$$

$q_i(x_i)$ is a distribution which means it needs to sum to 1. However it is a really simple task as x_i can only take two values -1 or 1.

$$q_i(x_i=1) = \frac{1}{q(x_i=1)+q(x_i=-1)} \cdot q(x_i=1) =$$

$$= \frac{e^{m_i + L_i(1)}}{e^{m_i+L_i(1)} + e^{-m_i+L_i(-1)}} =$$

$$= \left\{ \frac{e^a}{e^a+e^b} = \frac{1}{e^{-a}(e^a+e^b)} = \frac{1}{1+e^{b-a}} \right\} =$$

$$= \frac{1}{1+e^{-2m_i - L_i(1)+L_i(-1)}} = \frac{1}{1+e^{-2\underbrace{(m_i+\frac{1}{2}(L_i(1)-L_i(-1)))}_{a_i}}} =$$

$$= \frac{1}{1+e^{-2a_i}} = \mathrm{sigm}(2a_i)$$

$$\Rightarrow q_i(x_i=-1) = \mathrm{sigm}(-2a_i)$$

A sigmoid posterior makes perfect sense for a classification task.

We want to update the variational parameter of each proposal distribution.

$$\mu_i = \mathbb{E}_{q_i(x_i)}[x_i] = (+1)\cdot q_i(x_i=1) + (-1)\cdot q_i(x_i=1) =$$

$$= \frac{1}{1+e^{-2a_i}} - \frac{1}{1+e^{2a_i}} = \frac{e^{a_i}}{e^{a_i}+e^{-a_i}} - \frac{e^{-a_i}}{e^{-a_i}+e^{a_i}} = \frac{e^{a_i}-e^{-a_i}}{e^{a_i}+e^{-a_i}}$$

$$= \tanh(a_i) = \tanh\left(m_i + \frac{1}{2}(L_i(1) - L_i(-1))\right) =$$

$$= \tanh\left(\sum_{j\in N(i)} W_{ij}\mu_j + \frac{1}{2}(L_i(1) - L_i(-1))\right)$$

Iteratively update these in turn

$$\mu_i^t = \tanh\left(\sum_{j\in N(i)} W_{ij}\mu_j^{t-1} + \frac{1}{2}(L_i(1) - L_i(-1))\right)$$

Or a damped update

$$\mu_i^t = (1-\lambda)\mu_i^{t-1} + \lambda\cdot\tanh\left(\sum_{j\in N(i)} W_{ij}\mu_j^{t-1} + \frac{1}{2}(L_i(1) - L_i(-1))\right)$$