1 Binary sparse coding model definition

If h_i is 1, the edge is included in the image.

Choose the edges independently of each other:

$p(h_i = 1) = \sigma(b_i)$

We can add up the edges with a matrix multiply:

$$Wh$$

To get a smooth distribution over images v, we add some Gaussian noise:

$$p(v \mid h) = \mathcal{N}(v \mid Wh, I)$$

By multiplying all of the $p(h_i)$ and $p(v \mid h)$ together we get:

$$p(h, v) = \Pi_i \frac{\exp(b_i h_i)}{1 + \exp(b_i)} \Pi_j \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2}\|v - Wh\|_2^2\right)$$

This is an *energy-based model* $p(h, v) = \frac{1}{Z}\exp(-E(h, v))$ with

$$E(v, h) = -b^T h + \frac{1}{2}\|v - Wh\|_2^2$$

$$Z = \Pi_i \sigma(-b_i) \Pi_j \sqrt{2\pi}$$

2 Maxmimum likelihood in the binary sparse coding model

To train the model, we want to maximize the log likelihood.

We do this by following the derivatives of $\log p(v)$:

$$\frac{d}{d\theta} \log p(v) = \frac{d}{d\theta} \log \sum_h p(h, v)$$

$$= \frac{d}{d\theta} \log \sum_h \frac{1}{Z} \exp\left(-E(h, v)\right)$$

$$= \frac{d}{d\theta} \log \frac{1}{Z} \sum_h \exp\left(-E(h, v)\right)$$

$$= \frac{d}{d\theta} \left[\log \sum_h \exp\left(-E(h, v)\right) - \log Z \right]$$

This is exactly the same as training an RBM, but with a new $E(h, v)$ and Z.

$$Z = \Pi_i \sigma(-b_i) \Pi_j \sqrt{2\pi}$$

so

$$\log Z = \sum_i -\log(1 + \exp(b_i)) - \frac{1}{2} \sum_j \log 2\pi$$

$$\frac{d}{dW} \log Z = 0$$

$$\frac{d}{db} \log Z = -\sigma(b)$$

The positive phase is expensive:

$$\frac{d}{d\theta} \log \sum_h \exp\left(-E(h, v)\right)$$

$$= -\mathbb{E}_{h \sim p(h|v)} \frac{d}{d\theta} E(h, v)$$

The expectation requires computing $p(h \mid v)$.

For RBMs, this is easy. But for binary sparse coding, even drawing samples is hard!

$$p(h \mid v) \propto \exp\left(b^T h - \frac{1}{2}\|v - Wh\|_2^2\right)$$

$$= \exp\left(b^T h - \frac{1}{2}v^2 v + v^T W h - \frac{1}{2}h^T W^T W h\right)$$

$$\propto \exp\left(b^T h + v^T W h - \frac{1}{2}h^T W^T W h\right)$$

$$= \Pi_i \frac{\exp(b_i h_i)\exp(v^T W_{:i} h_i)}{\Pi_j \exp(\frac{1}{2}W_{:i}^T W_{:j} h_i h_j)}$$

Every h_i interacts with every h_j!

This means we can't normalize the distribution

over each h_i separately from the others.

3 Variational inference

$$q(h) = \Pi_i q(h_i)$$

$$\mathbb{E}_{h \sim q(h)} \log p(h, v)$$

$$q(h) = \text{argmin}_q KL\left(q(h) \| p(h \mid v)\right)$$

$$\text{subject to } q(h) = \Pi_i q(h_i)$$

$$q(h) = \Pi_i q(h_i)$$

where $q(h_i = 1) = \hat{h}_i$

and $\hat{h_i} \in [0, 1]$ is an optimization parameter

We can solve the minimization problem

$$\min_{\hat{h}} KL\left(q(h) \| p(h \mid v)\right)$$

just by algebra, by solving

$$\nabla_{\hat{h}} KL(q(h) \| p(h) \mid v) = 0$$

for $\hat{h}$.

$$\frac{\partial}{\partial \hat{h}_i} KL\left(q(h) \| p(h \mid v)\right) = 0$$

4 Fixed point equations for binary sparse coding

$$\frac{\partial}{\partial \hat{h}_i} KL\left(q(h) \| p(h \mid v)\right) = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_h \left[q(h) \log q(h) - q(h) \log p(h \mid v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_h \left[q(h) \sum_i \log q_i(h) - q(h) \log p(h, v) + q(h) \log p(v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_h \left[q(h) \sum_i \log q_i(h) - q(h) \log p(h, v)\right] + \frac{\partial}{\partial \hat{h}_i} \log p(v) = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_h \left[q(h) \sum_i \log q_i(h) - q(h) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \left[\sum_i \sum_h q(h) \log q_i(h) - \sum_h q(h) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \left[\sum_i \left(\sum_{h_i} q(h_i) \log q_i(h)\right) - \sum_h q(h) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_{h_i} \left[q(h_i) \log q_i(h) - \sum_{h_{-i}} q(h) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_{h_i} \left[q(h_i) \log q_i(h) - \sum_{h_{-i}} q(h_i) q(h_{-i}) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \sum_{h_i} \left[q(h_i) \log q_i(h) - q(h_i) \sum_{h_{-i}} q(h_{-i}) \log p(h, v)\right] = 0$$

$$\frac{\partial}{\partial \hat{h}_i} \left[\hat{h_i} \log \hat{h}_i + (1 - \hat{h_i}) \log(1 - \hat{h}_i) - \sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \log p(h_i) p(h_{-i}) p(v \mid h)\right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \log p(h_i) p(h_{-i}) p(v \mid h) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \left(\log p(h_i) + \log p(h_{-i}) + \log p(v \mid h) \right) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \left(\log p(h_i) + \sum_{h_{-i}} q(h_{-i}) \log p(v \mid h) \right) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\hat{h}_i \log \sigma(b_i) + (1 - \hat{h}_i) \log \sigma(-b_i) + \sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \log p(v \mid h) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\hat{h}_i \log \sigma(b_i) + (1 - \hat{h}_i) \log \sigma(-b_i) + \sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \sum_j \log p(v_j \mid h) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \frac{\partial}{\partial \hat{h}_i} \left[\hat{h}_i \log \sigma(b_i) - \hat{h}_i \log \sigma(-b_i) + \sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \sum_j \log \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2} (v_j - W_{j:} h)^2 \right) \right]$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \log \sigma(b_i) + \log \sigma(-b_i) - \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \sum_j \log \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2} (v_j - W_{j:} h)^2 \right) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \log \sigma(b_i) + \log \sigma(-b_i) + \frac{1}{2} \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \sum_j (v_j - W_{j:} h)^2 \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \log \sigma(b_i) + \log \sigma(-b_i) + \frac{1}{2} \frac{\partial}{\partial \hat{h}_i} \left[\sum_{h_i} q(h_i) \sum_{h_{-i}} q(h_{-i}) \sum_j (v_j^2 - 2 v_j W_{j:} h + h^T W_{j:}^T W_{j:} h) \right] = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - \log \sigma(b_i) + \log \sigma(-b_i) - v^T W_{:i} + \frac{1}{2} W_{:i}^T W_{:i} + \frac{1}{2} \sum_{j \neq i} W_{j:}^T W_{i:} \hat{h}_j = 0$$

$$\log \hat{h}_i - \log(1 - \hat{h}_i) - b_i - v^T W_{:i} + \frac{1}{2} W_{:i}^T W_{:i} + \frac{1}{2} \sum_{j \neq i} W_{j:}^T W_{i:} \hat{h}_j = 0$$

$$\hat{h}_i = \sigma \left(v^T W_{:i} + b_i - \frac{1}{2} W_i^T W_i - \frac{1}{2} \sum_j W_j^T W_i \hat{h}_j \right)$$

5 Variational inference with continuous values

The Euler-Lagrange equations state that if

$$F[f] = \int G\left(f(x), f'(x), x\right) dx$$

then Fmay be minimized by solving

$$\frac{\partial G}{\partial f} = \frac{d}{dx}\left(\frac{\partial G}{\partial f'}\right)$$

If we split h into disjoint groups, then the group

with indices in set S has

$$q(h_S) \propto \exp\left(\mathbb{E}_{h_{-S} \sim q} \log p(h, v)\right)$$

where h_{-S} is all of the variables not in this group.