Neural networks

Restricted Boltzmann machine - contrastive divergence (parameter update)

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial\theta} = \underbrace{\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\,\mathbf{x}^{(t)}\right]}_{\text{positive phase}} - \underbrace{\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]}_{\text{negative phase}}$$

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

hard to compute

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial \theta}\,\middle|\,\mathbf{x}^{(t)}\right]}_{\text{positive phase}} - \underbrace{\boxed{\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\right]}}_{\text{negative phase}}$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}$ for $\theta = W_{jk}$

$$\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}} = \frac{\partial}{\partial W_{jk}} \left(-\sum_{jk} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \right)$$

$$= -\frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k$$

$$= -h_j x_k$$

$$\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h}) = -\mathbf{h}\,\mathbf{x}^\top$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\Big|\mathbf{x}\right]$ for $\theta = W_{jk}$

$$\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}}\Big|\mathbf{x}\right] = \mathbb{E}_{\mathbf{h}}\left[-h_j x_k \Big|\mathbf{x}\right] = \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j|\mathbf{x})$$

$$= -x_k p(h_j = 1|\mathbf{x})$$

$$\mathbf{h}(\mathbf{x}) \stackrel{\text{def}}{=} \begin{pmatrix} p(h_1 = 1|\mathbf{x}) \\ \dots \\ p(h_H = 1|\mathbf{x}) \end{pmatrix}$$

$$= \operatorname{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x})$$

$$\mathrm{E}_{\mathbf{h}}\left[\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h})\,|\mathbf{x}\right] = -\mathbf{h}(\mathbf{x})\,\mathbf{x}^\top$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Given $\mathbf{x}^{(t)}$ and $\tilde{\mathbf{x}}$ the learning rule for $\theta = \mathbf{W}$ becomes

$$
\begin{aligned}
\mathbf{W} \quad & \Longleftarrow \quad \mathbf{W} - \alpha \left(\nabla_{\mathbf{W}} - \log p(\mathbf{x}^{(t)}) \right) \\
& \Longleftarrow \quad \mathbf{W} - \alpha \left(\mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \,\middle|\, \mathbf{x}^{(t)} \right] - \mathrm{E}_{\mathbf{x},\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h}) \right] \right) \\
& \Longleftarrow \quad \mathbf{W} - \alpha \left(\mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \,\middle|\, \mathbf{x}^{(t)} \right] - \mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}}, \mathbf{h}) \,|\tilde{\mathbf{x}} \right] \right) \\
& \Longleftarrow \quad \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)})\, {\mathbf{x}^{(t)}}^{\top} - \mathbf{h}(\tilde{\mathbf{x}})\, \tilde{\mathbf{x}}^{\top} \right)
\end{aligned}
$$

CD-K: PSEUDOCODE

Topics: contrastive divergence

1. For each training example $\mathbf{x}^{(t)}$
   i. generate a negative sample $\tilde{\mathbf{x}}$ using
      k steps of Gibbs sampling, starting at $\mathbf{x}^{(t)}$
   ii. update parameters

$$\mathbf{W} \Longleftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)})\, {\mathbf{x}^{(t)}}^\top - \mathbf{h}(\tilde{\mathbf{x}})\, \tilde{\mathbf{x}}^\top \right)$$

$$\mathbf{b} \Longleftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}}) \right)$$

$$\mathbf{c} \Longleftarrow \mathbf{c} + \alpha \left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}} \right)$$

2. Go back to 1 until stopping criteria

CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence

- CD-k: contrastive divergence with k
  iterations of Gibbs sampling
- In general, the bigger k is, the less **biased** the estimate of the gradient will be
- In practice, k=1 works well for pre-training