GRAPHICAL MODELS

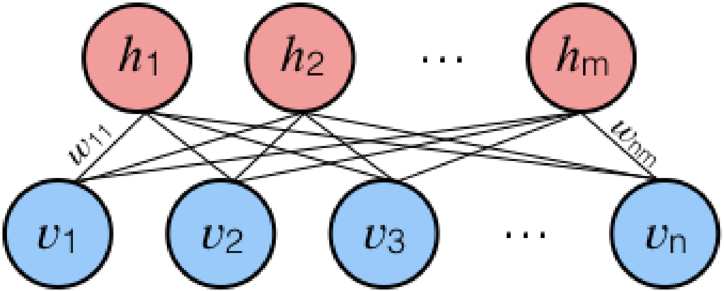
Classification Restricted Boltzmann Machines

Present by:
Son Tung LE
Thi Ha Giang NGUYEN

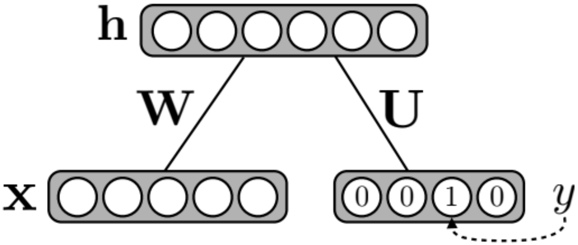
19 avril 2021

Plan

Restricted Boltzmann Machines (RBM)



Classification Restricted Boltzmann Machines (ClassRBM)



Energy function

$$E(y, \mathbf{x}, \mathbf{h}) = -\mathbf{h}^T\mathbf{W}\mathbf{x} - \mathbf{b}^T\mathbf{x} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{e}_y - \mathbf{h}^T\mathbf{U}\mathbf{e}_y$$

Classification Restricted Boltzmann Machines (ClassRBM)

- Probability of every possible set of an input, a label and a hidden vector :

$$p(y, \mathbf{x}, \mathbf{h}) = \frac{\exp(-E(y, \mathbf{x}, \mathbf{h}))}{\sum_{y,\mathbf{x},\mathbf{h}} \exp(-E(y, \mathbf{x}, \mathbf{h}))} = \frac{\exp(-E(y, \mathbf{x}, \mathbf{h}))}{Z}.$$

- Conditional probability :

$$p(y|\mathbf{x}) = \frac{\exp(-F(y, \mathbf{x}))}{\sum_{y^* \in \{1,\dots,C\}} \exp(-F(y^*, \mathbf{x}))}$$

where $F(y, \mathbf{x}) = d_y + \sum_j \text{softplus}(c_j + U_{jy} \sum_i W_{ij} x_i)$ and $\text{softplus}(a) = \log(1 + \exp(a))$.

Generative training

Generative objective function

$$\mathcal{L}_{gen}(\mathcal{D}_{train}) = -\sum_{t=1}^{|\mathcal{D}_{train}|} \log p(y_t, \mathbf{x}_t).$$

Its gradient :

$$\frac{\partial \log p(y_t, \mathbf{x}_t)}{\partial \theta} = -\mathbb{E}_{h|y_t,\mathbf{x}_t}\left[\frac{\partial}{\partial \theta} E(y_t, \mathbf{x}_t, \mathbf{h})\right] + \mathbb{E}_{y,\mathbf{x},\mathbf{h}}\left[\frac{\partial}{\partial \theta} E(y, \mathbf{x}, \mathbf{h})\right]$$

Generative training

Gibbs sampling

For a sample $(y_t, \mathbf{x}_t)$:

1. Compute the conditional probability

$$\hat{\mathbf{h}}_t = p(\mathbf{h}|y_t, \mathbf{x}_t) = sigm(-\mathbf{W}\mathbf{x_t} - \mathbf{U}\mathbf{e}_{y_t} - \mathbf{c})$$

2. Take sample $\mathbf{h}$ ~p(h|y$_t$, $\mathbf{x}_t$)
3. From $\mathbf{h}$, sample a reconstruction $(y_t^1, \mathbf{x}_t^1)$
4. Compute the reconstructed conditional probability

$$\hat{\mathbf{h}}_t^1 = p(\mathbf{h}|y_t^1, \mathbf{x}_t^1)$$

Generative training

Generative gradient :

$$\nabla_{\mathbf{W}} \log p(y_t|\mathbf{x}_t) = -\hat{h}_t \otimes \mathbf{x}_t + \hat{h}_t^1 \otimes \mathbf{x}_t^1$$

$$\nabla_{\mathbf{U}} \log p(y_t|\mathbf{x}_t) = -\hat{h}_t \otimes y_t + \hat{h}_t^1 \otimes y_t^1$$

$$\nabla_{\mathbf{c}} \log p(y_t|\mathbf{x}_t) = -\hat{h}_t + \hat{h}_t^1$$

$$\nabla_{\mathbf{d}} \log p(y_t|\mathbf{x}_t) = -y_t + y_t^1$$

$$\nabla_{\mathbf{b}} \log p(y_t|\mathbf{x}_t) = -\mathbf{x}_t + \mathbf{x}_t^1$$

Discriminative training

Discriminative objective function

$$\mathcal{L}_{disc}(\mathcal{D}_{train}) = -\sum_{t=1}^{|\mathcal{D}_{train}|} \log p(y_t|\mathbf{x}_t).$$

Its gradient :

$$\frac{\partial \log p(y_t|\mathbf{x}_t)}{\partial \theta} = -\mathbb{E}_{\mathbf{h}|y_t,\mathbf{x}_t}\left[\frac{\partial}{\partial \theta}E(y_t,\mathbf{x}_t,\mathbf{h})\right] + \mathbb{E}_{y,h|\mathbf{x}_t}\left[\frac{\partial}{\partial \theta}E(y,\mathbf{x}_t,\mathbf{h})\right].$$

Consider the second term of this gradient :

$$\mathbb{E}_{y,\mathbf{h}|\mathbf{x}_t}\left[\frac{\partial}{\partial \theta}E(y,\mathbf{h}|\mathbf{x}_t)\right] = \mathbb{E}_{y|\mathbf{x}_t}\left[\mathbb{E}_{\mathbf{h}|y,\mathbf{x}_t}\left[\frac{\partial}{\partial \theta}E(y,\mathbf{x}_t,\mathbf{h})\right]\right].$$

Discriminative gradients

$$\nabla_{\mathbf{W}} \log p(y_t|\mathbf{x}_t) = \vec{p}(\mathbf{h}|y_t, \mathbf{x}_t) \otimes \mathbf{x}_t - \sum_{y^* \in \{1,\ldots C\}} \left[\vec{p}(\mathbf{h}|y^*, \mathbf{x}_t) \otimes \mathbf{x}_t\right] p(y^*|\mathbf{x}_t)$$

$$\nabla_{\mathbf{U}} \log p(y_t|\mathbf{x}_t) = \vec{p}(\mathbf{h}|y_t, \mathbf{x}_t) \otimes \mathbf{e}_{y_t} - \sum_{y^* \in \{1,\ldots C\}} \left[\vec{p}(\mathbf{h}|y^*, \mathbf{x}_t) \otimes \mathbf{e}_{y^*}\right] p(y^*|\mathbf{x}_t)$$

$$\nabla_{\mathbf{c}} \log p(y_t|\mathbf{x}_t) = \vec{p}(\mathbf{h}|y_t, \mathbf{x}_t) - \sum_{y^* \in \{1,\ldots C\}} \vec{p}(\mathbf{h}|y^*, \mathbf{x}_t) p(y^*|\mathbf{x}_t)$$

$$\nabla_{\mathbf{d}} \log p(y_t|\mathbf{x}_t) = \mathbf{e}_{y_t} - \sum_{y^* \in \{1,\ldots C\}} \mathbf{e}_{y^*} \; p(y^*|\mathbf{x}_t)$$

$$\nabla_{\mathbf{b}} \log p(y_t|\mathbf{x}_t) = 0$$

Hybrid training

Hybrid objective function

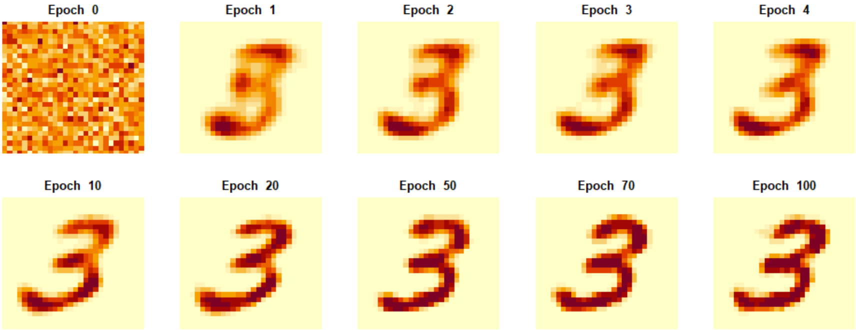
$$\mathcal{L}_{hybrid}(\mathcal{D}_{train}) = \mathcal{L}_{disc}(\mathcal{D}_{train}) + \alpha\mathcal{L}_{gen}(\mathcal{D}_{train}).$$

Implementation

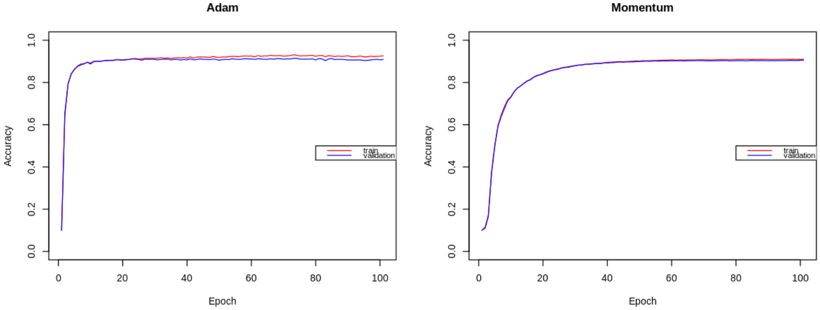
Settings :

- MNIST dataset :
  - 10,000 samples for training
  - 10,000 samples for validation
- 100 hidden units
- minibatch size = 100

Generative training



Generative training



Classification performances

Classification performances : accuracy train - validation

- Generative : 92.62% - 90.95%
- Discriminative : 100% - 94.61%
- Hybrid : 99.96% - 92.03%

Semi-supervised learning

Semi-supervised objective function

$$\mathcal{L}_{semi}(\mathcal{D}_{train}, \mathcal{D}_{unlabel}) = \mathcal{L}_{sup}(\mathcal{D}_{train}) + \beta \mathcal{L}_{unsup}(\mathcal{D}_{unlabel})$$

where $\mathcal{L}_{sup}$ is one of the above objective functions and

$$\mathcal{L}_{unsup}(\mathcal{D}_{unlabel}) = - \sum_{t=1}^{|\mathcal{D}_{unlabel}|} \log p(\mathbf{x}_t)$$

Gradient of unsupervised part :

$$\frac{\partial \log p(\mathbf{x}_t)}{\partial \theta} = -\mathbb{E}_{y,\mathbf{h}|\mathbf{x}_t}\left[\frac{\partial}{\partial \theta} E(y, \mathbf{x}_t, \mathbf{h})\right] + \mathbb{E}_{y,\mathbf{h},\mathbf{x}}\left[\frac{\partial}{\partial \theta} E(y, \mathbf{x}, \mathbf{h})\right].$$

Initialize $y_t^0 \sim p(y|\mathbf{x}_t)$ and use Gibbs sampling.