Classification using Discriminative Restricted Boltzmann Machines
Classification using Discriminative Restricted Boltzmann Machines

Hugo Larochelle
Yoshua Bengio


Dept. IRO, Université de Montréal C.P. 6128, Montreal, Qc, H3C 3J7, Canada
• Introduction
• Hybrid Discriminative RBM
• Semi-supervised Learning
• Experiments and results
• Conclusion
Introduction

Types of learning:
- Supervised
  - Continuous
    - Regression
  - Categorical
    - Classification
- Unsupervised
  - No label
    - Clustering
- Semi-supervised
  - A lot of unlabeled observations and a few labeled
    - Classification and clustering
Restricted Boltzmann Machines (RBM)

- Using hidden layer of variable to model a distribution over visible layers.
- Mostly they are trained to model the input of classification tasks.
- They can also model the joint distribution of the inputs and associated target classes.

A training set

\[ \mathcal{D}_{\text{train}} = \{(x_i, y_i)\} \]

Target class

\[ y_i \in \{1, \ldots, C\} \]

Minimization of negative log-likelihood

\[ \mathcal{L}_{\text{gen}}(\mathcal{D}_{\text{train}}) = -\sum_{i=1}^{\mathcal{D}_{\text{train}}} \log p(y_i, x_i) \]
Restricted Boltzmann Machines (RBM)

Joint distribution between a layer of hidden variables and the observed variables made of $x$ and $y$.

$$E(y, x, h) = -h^T W x - b^T x - c^T h - d^T y - h^T U y$$

$$p(y, x, h) \propto e^{-E(y, x, h)}$$

$$\theta \in (W, b, c, d, U)$$

Diagram:
- Hidden units
- Target units (1 out of C)
- Input units

Diagram shows the connections between hidden units, target units, and input units with variables $y$, $x$, and $h$.
Restricted Boltzmann Machines (RBM)

\[ p(x_i = 1|h) = \operatorname{sigm}(b_i + \sum_j W_{ji}h_j) \]

\[ p(y|h) = \frac{e^{d_y + \sum_j U_{yj}h_j}}{\sum_{y^*} e^{d_{y^*} + \sum_j U_{y^*j}h_j}} \]

\[ p(h_j = 1|y, x) = \operatorname{sigm}(c_j + U_{yj} + \sum_i W_{ji}x_i). \]

In order to minimize the negative log-likelihood:

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

✓ Contrastive Divergence
Restricted Boltzmann Machines (RBM)

It is possible to compute $p(y|x)$, sample from it, or choose the most probable class under this model.

$$p(y|x) = \frac{e^{d_y} \prod_{j=1}^{n} (1 + e^{c_j + U_{jy} + \sum_i W_{ji} x_i})}{\sum_{y} e^{d_y} \prod_{j=1}^{n} (1 + e^{c_j + U_{jy} + \sum_i W_{ji} x_i})}.$$ 

Therefore classification can be performed.

This conditional distributions can be computed in:

$O(nd + nC)$

n: number of hidden units
D: input dimensionality
C number of classes

The features representations learned by the hidden layer is not guaranteed to be useful for classification task.
Discriminative RBM

We directly optimize $p(y|x)$ instead of the joint distributions ($p(y,x)$)

$$
\mathcal{L}_{disc}(\mathcal{D}_{train}) = - \sum_{i=1}^{\mid \mathcal{D}_{train} \mid} \log p(y_i \mid x_i)
$$

Training:

$$
\frac{\partial \log p(y_i \mid x_i)}{\partial \theta} = \sum_j \text{sigm}(o_{yj}(x_i)) \frac{\partial o_{yj}(x_i)}{\partial \theta} - \sum_{j,y^*} \text{sigm}(o_{y^*j}(x_i))p(y^* \mid x_i) \frac{\partial o_{y^*j}(x_i)}{\partial \theta}
$$

$$
o_{yj}(x) = c_j + \sum_k W_{jk} x_k + U_{jy}
$$
Hybrid Discriminative RBM

Size of training set

- Smaller training sets tend to favor generative learning.
- Bigger training sets tend to favor discriminative learning.

By combining respective training criteria:

\[ \mathcal{L}_{\text{hybrid}}(\mathcal{D}_{\text{train}}) = \mathcal{L}_{\text{disc}}(\mathcal{D}_{\text{train}}) + \alpha \mathcal{L}_{\text{gen}}(\mathcal{D}_{\text{train}}) \]

\( \alpha \) is a parameter for controlling generative criterion and can be optimized (based on the validation set classification error).

Generative criterion acts as a data-dependent regularizer.
Semi-supervised learning

✓ Few labeled training data but many unlabeled examples of inputs

\[
\mathcal{L}_{semi-sup}(\mathcal{D}_{train}, \mathcal{D}_{unlab}) = \mathcal{L}_{TYPE}(\mathcal{D}_{train}) + \beta \mathcal{L}_{unsup}(\mathcal{D}_{unlab})
\]

\[\text{TYPE } \in \{\text{gen, disc, hybrid}\}\]

For unlabeled data:

\[
\mathcal{L}_{unsup}(\mathcal{D}_{unlab}) = -\sum_{i=1}^{\lvert \mathcal{D}_{unlab} \rvert} \log p(x_i)
\]

\[\mathcal{D}_{unlab} = \{(x_i)\}_{i=1}^{\lvert \mathcal{D}_{unlab} \rvert} \quad \text{Set of unlabeled data}\]
Semi-supervised learning

Using a contrastive divergence approximation:

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

\[
\mathbb{E}_{y|x_i} \left[ \mathbb{E}_{h|y,x_i} \left[ \frac{\partial}{\partial \theta} E(y_i, x_i, h) \right] \right]
\]

\[
P(A, B|C) = \frac{P(A, B, C)}{P(C)} = \frac{P(A, C)}{P(C)} \times \frac{P(A, B, C)}{P(A, C)} = P(A|C) \times P(B|A, C)
\]
Experiments

• Character recognition

MNIST dataset
Training set: 50000 examples
Validation set: 10000 examples
Test set: 10000 examples

n: number of training iterations
λ: learning rate

<table>
<thead>
<tr>
<th>Model</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBM (λ = 0.005, n = 6000)</td>
<td>3.39%</td>
</tr>
<tr>
<td>DRBM (λ = 0.05, n = 500)</td>
<td>1.81%</td>
</tr>
<tr>
<td>RBM+NNet</td>
<td>1.41%</td>
</tr>
<tr>
<td>HDRBM (α = 0.01, λ = 0.05, n = 1500 )</td>
<td>1.28%</td>
</tr>
<tr>
<td>Sparse HDRBM (idem + n = 3000, δ = 10⁻⁴)</td>
<td>1.16%</td>
</tr>
<tr>
<td>SVM</td>
<td>1.40%</td>
</tr>
<tr>
<td>NNet</td>
<td>1.93%</td>
</tr>
</tbody>
</table>

\[ \mathcal{L}_{\text{hybrid}}(\mathcal{D}_{\text{train}}) = \mathcal{L}_{\text{disc}}(\mathcal{D}_{\text{train}}) + 0.01\mathcal{L}_{\text{gen}}(\mathcal{D}_{\text{train}}) \]

Sparse HDRBM: Subtracting a small value from the biases c in the hidden layer after each update.
Experiments

- Document classification

20-newsgroup Dataset

20 news-group topic (Targets)

Training set: 9578 examples

Validation set: 1691 examples

Test set: 5000 examples

<table>
<thead>
<tr>
<th>Model</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBM ($\lambda = 0.0005, n = 1000$)</td>
<td>24.9%</td>
</tr>
<tr>
<td>DRBM ($\lambda = 0.0005, n = 50$)</td>
<td>27.6%</td>
</tr>
<tr>
<td>RBM+NNet</td>
<td>26.8%</td>
</tr>
<tr>
<td>HDRBM ($\alpha = 0.005, \lambda = 0.1, n = 1000$)</td>
<td>23.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>32.8%</td>
</tr>
<tr>
<td>NNet</td>
<td>28.2%</td>
</tr>
</tbody>
</table>
Experiments

- Document classification

20-newsgroup Dataset

Similarity matrix of newsgroup weight vectors $U_{xy}$. 
Conclusion

- RBMs can be employed as stand-alone non-linear classifiers, and not only simple feature extractors.
- Discriminative RBM discover features of inputs with their use in classification.
- On smaller training set generative model performs better.
- Hybrid Discriminative RBM has been introduced.
- Novel semi-supervised algorithm for RBMs has been presented.

Future Work

- Use of DRBMs in more challenging settings such as in multi-task or structured-out problems.