Classification using Discriminative Restricted Boltzmann Machines

Motivation
• Recently, many applications for Restricted Boltzmann Machines (RBMs) have been developed for a large variety of learning problems
• They are usually used to extract features or to initialize deep neural networks
• We argue that RBMs provide a self-contained framework for deriving competitive non-linear classifiers
• We present algorithms that introduce a discriminative component to RBM training
• We demonstrate how discriminative RBMs can also be successfully employed in a semi-supervised setting

Discriminative Restricted Boltzmann Machines (DRBM)
• In a classification setting, we are not interested in obtaining a good model of the input distribution p(x). It can be advantageous to maximize the conditional likelihood of the targets, or equivalently minimize:
  \[ \sum_{i \in \text{train}} \log p(y_i | x_i) \]
• DRBM could be trained by Contrastive Divergence too, however exact gradient can be computed:
  \[ \frac{\partial \log p(y_i | x_i)}{\partial \theta} = \sum_{n \in \theta} \left( \sum_j s_j(y_i, n_j) \right) \]
  \[ \frac{\partial \log p(x_i | y_i)}{\partial \theta} = \sum_{n \in \theta} \left( \sum_j s_j(x_i, n_j) \right) \]
  where \( s_j(x) = c_j + \sum_{i \in \text{train}} W_{ji} x_i \)
  • Using this gradient, we can perform stochastic gradient descent

Contrastive Divergence
Algorithm for Contrastive Divergence parameter update:
Input: training pair \((x, y)\) and learning rate \(\alpha\)
• Notation: \( a \) — means \( a \) is set to value \( b \)
• Positive phase
  \[ p(y^* | x, y) = \frac{1}{1 + \exp(-\alpha)} \]
• Negative phase
  \[ q_i(y^* | x, y) = \frac{1}{1 + \exp(-\alpha)} \]
• Update for \( \theta = \theta - \alpha \frac{\partial \log p(x | y)}{\partial \theta} \)

Semi-supervised Learning in RBMs
• What about a classification setting where there are few labeled training data but many unlabeled examples of inputs?
• Semi-supervised learning algorithms address this situation by using the unlabeled data to introduce constraints on the trained model
• Trained as generative models, i.e. maximize the joint likelihood of the targets and inputs, or equivalently minimize:
  \[ \mathcal{L}_{unlabeled}(\theta) = \sum_i \log p(x_i | y_i) \]
• Training algorithm based on stochastic descent, where gradient for parameters \( \theta \) is:
  \[ \frac{\partial \log p(x_i | y_i)}{\partial \theta} = \mathbb{E}_{x \sim \text{RBMs}} \left[ \frac{\partial \log p(x_i | y_i)}{\partial \theta} \right] \]
  is estimated using Contrastive Divergence

Future Work
• Investigate the use of discriminative versions of RBMs in more challenging settings such as in multi-task or structured output problems
• Explore ways to introduce generative learning in RBMs and HDRBMs which would be less computationally expensive when the input vectors are large but sparse

References

Two-dimensional PCA embedding of the newsgroups weights

Character Recognition
• Experiment on the MNIST dataset (5000, 10000 and 10000 example in training, validation and test sets)
• Sparse version of HDRBM: push biases of hidden units down by subtracting \( \lambda \) after every parameter update

Document Classification
• Experiment on the 20newsgroup dataset (5000 most frequent words, with 9578, 1691 and 7953 examples in training, validation and test sets)