CS786
Lecture 22: July 17, 2012

Deep Learning

Deep Learning

- Definition: training of graphical models (Bayes net, Markov net or neural net) with several layers of hidden nodes
  - Can learn expressive models
- Problems:
  - Non-convex optimization
    - Local optima
  - Vanishing gradient
    - Only top layers tend to be learned

Deep Belief Networks

- Picture

- Nodes: sigmoid units
- Network: mix of directed and undirected edges

Greedy Layer-wise Training

- Algorithm

Construct RBM with input layer $\boldsymbol{v}$ and hidden layer $\boldsymbol{h}_1$
$\boldsymbol{W}^{(1)} \leftarrow$ Train the RBM for $\boldsymbol{h}_1$ with data for $\boldsymbol{v}$
For $i = 2$ to n
    Add a layer of hidden units $\boldsymbol{h}_i$ with RBM architecture
    Fix $\boldsymbol{W}^{(i-1)}$ and sample $\boldsymbol{h}_{i-1}$ from $\Pr(\boldsymbol{h}_{i-1}|v)$
    $\boldsymbol{W}^{(i)} \leftarrow$ Train the RBM for $\boldsymbol{h}_i$ with data sampled for $\boldsymbol{h}_{i-1}$

Supervised learning: adjust all weights by gradient descent

Greedy Layer-wise Training

- Picture

Greedy Layer-wise Training

- Does not suffer from vanishing gradients
  - Greedy training
- Circumvents local optima
  - Each additional layer creates new dimensions to escape local optima
- Guarantee
  - Each layer improves log likelihood

$$\log P(v) \geq H_{P(h|v)} + \sum_h P(h|v)(\underbrace{\log P(h)}_{\text{Trained by the second layer RBM}} + \log P(v|h))$$