CS786
Lecture 21: July 10th, 2012

Probabilistic graphical models as neural networks

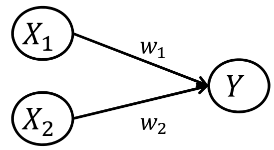
Neural Nets with Sigmoid Activation Functions

- Recall the shape of the sigmoid function

- Output is between 0 and 1
  → **Interpret output as distribution over binary variable**

Sigmoid Unit

- Sigmoid unit: $Y = \frac{1}{1+e^{-(w_1 X_1 + w_2 X_2)}}$



- Consider the following conditional random field

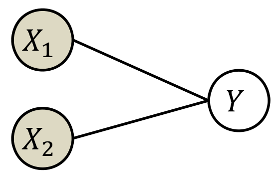
  $\Pr(Y|\boldsymbol{X}) = \frac{e^{\lambda_1 \phi_1(X_1,Y)+\lambda_2\phi_2(X_2,Y)}}{\sum_Y e^{\lambda_1 \phi_1(X_1,Y)+\lambda_2\phi_2(X_2,Y)}}$

  Let $\lambda_i = w_i$ and $\phi_i(X_i, Y) = X_i Y$

  Then $\Pr(Y|\boldsymbol{X}) = \frac{e^{w_1 X_1 Y + w_2 X_2 Y}}{\sum_Y e^{w_1 X_1 Y + w_2 X_2 Y}}$

  For $Y = 1$: $\Pr(Y = 1|\boldsymbol{X}) = \frac{e^{w_1 X_1 + w_2 X_2}}{e^{w_1 X_1 + w_2 X_2 + 1}}$

  $= \frac{1}{1+e^{-(w_1 X_1 + w_2 X_2)}}$



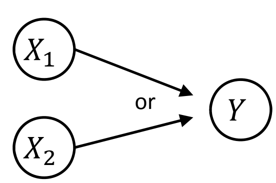
Noisy-or conditional distribution

- Sigmoid unit can also represent a Bayes net conditional distribution with a noisy-or structure
- Recall noisy-or formula:

  $\Pr(Y = 0|\boldsymbol{X}) = \prod_i q_i^{X_i}$ where $q_i = \Pr(Y = 0|X_i = 1, \boldsymbol{X}_{\sim i} = 0)$

- Let $\bar{Y} = 1 - Y$ and $q_i = \Pr(\bar{Y} = 0|X_i = 1, \boldsymbol{X}_{\sim i} = 0) = \frac{e^{w_i X_i}}{1+e^{\sum_i w_i X_i}}$
- Then $\Pr(\bar{Y} = 0|\boldsymbol{X}) = \frac{e^{w_1 X_1 + w_2 X_2}}{e^{w_1 X_1 + w_2 X_2 + 1}}$

  $= \frac{1}{1+e^{-(w_1 X_1 + w_2 X_2)}}$



Equivalences

Sigmoid unit

$\Longleftrightarrow$

Single output conditional random field

$\Longleftrightarrow$

Bayes net noisy-or conditional distribution

Acyclic Directed Graphs

Feed-forward sigmoid neural net

- Sigmoid units

- Output: composition of sigmoid functions

$$y_k = \sigma\left(\sum_j w_{kj}^{(2)} \sigma\left(\sum_i w_{ji}^{(1)} x_i\right)\right)$$

Sigmoid belief net (Bayes net)

- Sigmoid noisy-or conditional distributions

- Joint dist: product of sigmoid

$$\Pr(X_i = 1\ \forall i)$$
$$= \prod_i \sigma\left(\sum_{\{j | X_j \in pa(X_i)\}} w_{ji} X_j\right)$$

Hopfield Network

- Special type of recurrent neural network
  - Binary units
    - i.e., $X_i \in \{0,1\}\ \ \forall i$
  - Symmetric connections
    - i.e., $w_{ij} = w_{ji}$ and $w_{ii} = 0$
  - Threshold activation function
    - i.e., $h(a_j) = \begin{cases} 1 \text{ if } a_j > 0 \\ 0\ otherwise \end{cases}$ where $a_j = \sum_i w_{ji} X_i - \theta_j$

Hopfield Network

- Properties:
  - Symmetric connections ensure convergence of the network to a stationary configuration
  - Stationary configuration is a local minimum of an energy function

$$E(\boldsymbol{X}) = -\frac{1}{2}\sum_{ij} w_{ij} X_i X_j + \sum_i \theta_i X_i$$

- Usage:
  - Associative memory
  - combinatorial optimization
    - e.g., traveling salesman, Sudoku, n-queens

Boltzmann Machine

- It is a stochastic Hopfield network
  - Binary units
    - i.e., $X_i \in \{0,1\}\ \forall i$ (sampled from distribution implied by sigmoid)
  - Symmetric connections
    - i.e., $w_{ij} = w_{ji}$ and $w_{ii} = 0$
  - Sigmoid activation function
    - i.e., $h(a_j) = \frac{1}{1+e^{-a_j}}$ where $a_j = \sum_i w_{ji} X_i - \theta_j$

Boltzmann Machine

- Properties:
  - Symmetric connections ensure convergence of the network to a stationary distribution $\Pr(\boldsymbol{X})$
  - Stationary distribution proportional to exponential of negative energy (same energy as for Hopfield nets)

$$\Pr(\boldsymbol{X}) \propto e^{-E(\boldsymbol{X})}$$

where $E(\boldsymbol{X}) = -\frac{1}{2}\sum_{ij} w_{ij} X_i X_j + \sum_i \theta_i X_i$

- Usage:
  - Stochastic combinatorial optimization
  - Feature extraction

Relation to Markov Networks

- A Boltzmann machine defines a pairwise Markov network

$$\Pr(\boldsymbol{X}) \propto e^{-E(\boldsymbol{X})} = \mathrm{e}^{\sum_{i<j} w_{ij} X_i X_j - \sum_i \theta_i X_i}$$

  – Singleton features: $\phi_i(X_i) = X_i$
  – Pairwise features $\phi_{ij}(X_i, X_j) = X_i X_j$

Relation to Markov Networks

- Stochastic updates of the units of a Boltzmann machine correspond to Gibbs sampling
- Unit update

$$\Pr(X_j = 1 | \boldsymbol{X}_{\sim j}) = \frac{1}{1 + e^{-\sum_i w_{ji} X_j + \theta_j}}$$

- Gibbs sampling

$$\Pr(X_j | \boldsymbol{X}_{\sim j}) = \frac{e^{\sum_i w_{ji} X_i X_j - \theta_j X_j}}{\sum_{X_j} e^{\sum_i w_{ji} X_i X_j - \theta_j X_j}}$$

$$\Pr(X_j = 1 | \boldsymbol{X}_{\sim j}) = \frac{e^{\sum_i w_{ji} X_i - \theta_j}}{e^{\sum_i w_{ji} X_i - \theta_j} + 1}$$

$$= \frac{1}{1 + e^{-\sum_i w_{ji} X_i + \theta_j}}$$

Inference

- Inference by Gibbs sampling is too slow in Boltzmann machines
  - Linear convergence of Gibbs leads to a large # of iterations
- Alternative: "restricted" Boltzmann machine
  - Bipartite Boltzmann machine

Restricted Boltzmann Machine (RBM)

- Two sets of nodes
  - Visible nodes: data
  - Hidden nodes: features
- Connections only between visible and hidden nodes
  - i.e., no visible-visible or hidden-hidden connections

Inference

- Inference in RBMs: $\Pr(hidden|visible)$
  - Gibbs sampling: convergence in one step