Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program

605-447/625-438

Dr. Mark Fleischer

Module 10.1: The Boltzmann Machine

What We've Covered So Far that is Relevant to Boltzmann Machines

Simulated Annealing

- The thermodynamic basis of SA
- The stationary and transition probabilities associated with SA
- Examples of combinatorial optimization problems

Hopfield Recurrent Neural Networks

- How to define the weight matrix.
- Examined their capability to recall/remember exemplar patterns.
- Examined their memory capacity.

In This Module We Will Cover

The Boltzmann Machine

- A stochastic version of the Hopfield Network
- Consensus/Energy function
- Using the sigmoid function as the activation function
- Briefly discussed the training formulae
- Look at calculating the respective stationary probabilities of various configurations (sets of states)
- Look at how to modify the weights so as to obtain desired stationary probabilities
- Examples

Recall the Hopfield Network

- Memory capacity about 11% of the length of exemplars. E.g., an exemplar vector with 100 elements could store approximately 11 exemplars with very high accuracy.
- Accuracy based on statistical independence of the exemplars, and a 3σ standard deviation.
  - This means a near certainty (≥ 99%) for accurately determining the most likely exemplar that an input vector represents.
  - Remember, the input vector is an exemplar perturbed by some noise.
  - Variance of the ‘noise’ associated with inputs is approximately $(n - 1)(P - 1)$

A Tradeoff!

- Can we sacrifice some certainty associated with memory recall/completion for greater memory capacity?
- Can't be based on issue of 'noise' --- once an input is provided, it is known/certain.
- Don't want it to be based on statistical independence --- relationship among exemplars is not the issue insofar as 'certainty' is concerned.

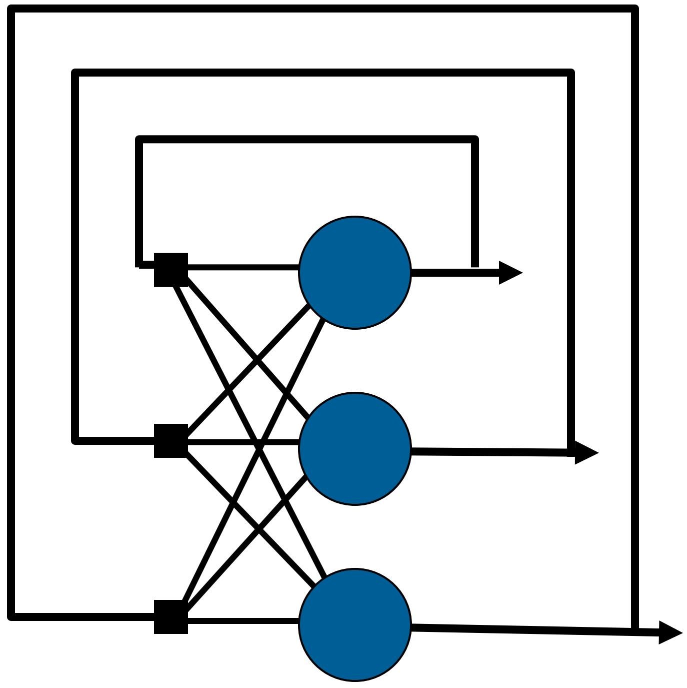
The ‘Uncertainty’

- Want to let the network ‘run’ and hopefully arrive at the correct exemplar.
- Could allow some probability the network will ‘arrive’ at the wrong exemplar.
- Allow the network to make some mistakes.
- How?

Let the nodes take on random states!

Recurrent Network Topology Reprise

Let’s view the network a bit differently.



Exemplar:

(1, -1, -1)

What is the weight from Node 1 to Node 2?

Boltzmann/Hopfield Comparisons

- Very similar in architecture.
- Boltzmann uses stochastic methods for updating node states.
- Hopfield uses bipolar state values.
- Boltzmann typically uses binary state values.

Activity Functions and Energy

- Asynchronous update of neuron activations.
- Cell activity S_i is computed (here without a bias term):

$$S_i = \sum_j w_{ij} x_j$$

The Hecht-Nielsen Function

Define the energy function *E*

$$E = -\sum_{i<j} w_{ij} x_i x_j + \sum_i \theta_i x_i$$

$$E = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j + \sum_i \theta_i x_i$$

Energy → Consensus

Minimize Energy

$$E = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j + \sum_i \theta_i x_i$$

Maximize Consensus

$$C = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j - \sum_i \theta_i x_i$$

Energy and Consensus values are
additive inverses of one another!

An Optimization Problem ala Simulated Annealing

- Minimize energy or Maximize consensus.
- Change the state of a node to modify a candidate energy/consensus function value.
  - Means changing it from 0 → 1 or 1 → 0.
- Accept new configuration probabilistically as in SA.

Calculating ΔE

Recall:

$$E = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j + \sum_i \theta_i x_i$$

WLG:

$$E = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n} \theta_i x_i - \frac{1}{2}\sum_{j=1}^{n} w_{kj}x_k x_j - \frac{1}{2}\sum_{i=1}^{n} w_{ik}x_i x_k + \theta_k x_k$$

Calculating ΔE

$$E_{\text{cand}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - \frac{1}{2}\sum_{j=1}^{n} w_{kj}x'_k x_j - \frac{1}{2}\sum_{i=1}^{n} w_{ik}x_i x'_k + \theta_k x'_k$$

$$E_{\text{cur}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - \frac{1}{2}\sum_{j=1}^{n} w_{kj}x_k x_j - \frac{1}{2}\sum_{i=1}^{n} w_{ik}x_i x_k + \theta_k x_k$$

$$E_{\text{cand}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - x'_k\sum_{i=1}^{n} w_{ik}x_i + \theta_k x'_k$$

$$E_{\text{cur}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - x_k\sum_{i=1}^{n} w_{ik}x_i + \theta_k x_k$$

Calculating ΔE

$$E_{\text{cand}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - x'_k\sum_{i=1}^{n} w_{ik}x_i + \theta_k x'_k$$

$$E_{\text{cur}} = -\frac{1}{2}\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} w_{ij}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n}\theta_i x_i - x_k\sum_{i=1}^{n} w_{ik}x_i + \theta_k x_k$$

$$\Delta E = E_{\text{cand}} - E_{\text{cur}}$$

Calculating ΔE

$$E_{\text{cand}} - E_{\text{cur}} = -x_k' \sum_{i=1}^{n} w_{ik} x_i + \theta_k x_k' - -x_k \sum_{i=1}^{n} w_{ik} x_i - \theta_k x_k$$

$$\Delta E = -x_k' \sum_{i=1}^{n} w_{ik} x_i + \theta_k x_k' + x_k \sum_{i=1}^{n} w_{ik} x_i - \theta_k x_k$$

$$= x_k' \left[-\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right] + x_k \left[\sum_{i=1}^{n} w_{ik} x_i - \theta_k \right]$$

$$= x_k' \left[-\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right] - x_k \left[-\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right]$$

Change in state of Node *k*

$$= \overbrace{\left(x_k' - x_k \right)}^{\text{Change in state of Node } k} \left[-\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right]$$

Calculating ΔC

Similarly for the consensus value. Thus,

$$C_{\text{cand}} - C_{\text{cur}} = x_k' \sum_{i=1}^{n} w_{ik} x_i - \theta_k x_k' - x_k \sum_{i=1}^{n} w_{ik} x_i + \theta_k x_k$$

$$\Delta C = x_k' \sum_{i=1}^{n} w_{ik} x_i + \theta_k x_k' - x_k \sum_{i=1}^{n} w_{ik} x_i + \theta_k x_k$$

$$= x_k' \left[\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right] - x_k \left[\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right]$$

$$= x_k' \left[\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right] - x_k \left[\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right]$$

$$= \left(x_k' - x_k \right) \left[\sum_{i=1}^{n} w_{ik} x_i + \theta_k \right]$$

Relating the Change of State to Probability

$$\pi_i(t) = \frac{e^{-E_i/t}}{\sum_i e^{-E_i/t}}$$

$$\frac{\pi_i(t)}{\pi_{i'}(t)} = \frac{\dfrac{e^{-E_i/t}}{\sum_i e^{-E_i/t}}}{\dfrac{e^{-E_{i'}/t}}{\sum_i e^{-E_i/t}}}$$

$$= \frac{e^{-E_i/t}}{e^{-E_{i'}/t}} = e^{(E_{i'}-E_i)/t} = e^{\Delta E/t}$$

Now, taking the logarithm of both sides we get …

Relating the Change of State to Probability

$$\ln\left(\frac{\pi_i(t)}{\pi_{i'}(t)}\right) = \ln e^{\Delta E/t} = \frac{\Delta E}{t}$$

$$\ln \pi_i(t) - \ln \pi_{i'}(t) = \frac{\Delta E}{t}$$

$$\ln\left(\Pr\{x_k = 1\}\right) - \ln\left(\Pr\{x_k = 0\}\right) = \frac{\Delta E}{t}$$

$$\ln\left(\Pr\{x_k = 1\}\right) - \ln\left(1 - \Pr\{x_k = 1\}\right) = \frac{\Delta E}{t}$$

Relating the Change of State to Probability

$$\ln\left(\Pr\{x_k = 1\}\right) - \ln\left(1 - \Pr\{x_k = 1\}\right) = \frac{\Delta E}{t}$$

$$\ln\left(\frac{\Pr\{x_k = 1\}}{1 - \Pr\{x_k = 1\}}\right) = \frac{\Delta E}{t}$$

$$\ln\left(\frac{1 - \Pr\{x_k = 1\}}{\Pr\{x_k = 1\}}\right) = \frac{-\Delta E}{t}$$

Relating the Change of State to Probability

$$\ln\left(\frac{1-\Pr\{x_k=1\}}{\Pr\{x_k=1\}}\right)=\frac{-\Delta E}{t}$$

$$\frac{1-\Pr\{x_k=1\}}{\Pr\{x_k=1\}}=e^{-\Delta E/t}$$

$$\frac{1}{\Pr\{x_k=1\}}-1=e^{-\Delta E/t}$$

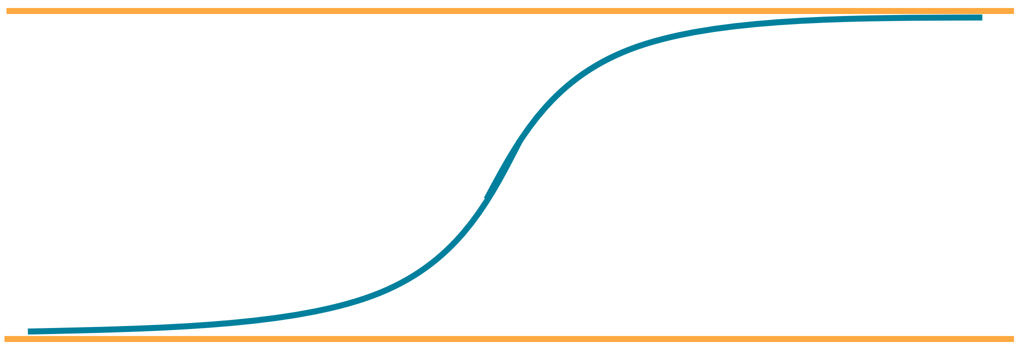
$$\frac{1}{\Pr\{x_k=1\}}=1+e^{-\Delta E/t}$$

$$\boxed{\Pr\{x_k=1\}=\frac{1}{1+e^{-\Delta E/t}}}$$

Dynamics

Recall the Sigmoid activation function

$$\frac{1}{1+e^{-S_i/T}}$$



We're dealing with stochastic neurons.
What does this curve remind you of?

Dynamics

- Yes, a probability distribution function (monotonically increasing to 1).
- Set cell activation (state) according to:

$$x_i = \begin{cases} 1 & \text{w/prob} \quad p_i = \dfrac{1}{1+e^{-S_i/T}} \\ 0 & \text{w/prob} \quad 1 - p_i \end{cases}$$

At high temperatures, what is the probability of $x_i = 1$?

Summary

- Each node is update asynchronously and probabilistically.
- The temperature is lowered to minimize the energy value of the network or maximize the consensus value of the network as the case may be.
- Since the node states are probabilistic, all information is encoded in the weights!