Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program

605-447/625-438

Dr. Mark Fleischer

Module 11.2: RBM Mathematics

What We've Covered So Far

- Probabilistic foundations of RBMs.
  - *Energy/Consensus* associated with a visible and hidden pair of vectors.
  - Probability of a node's states

Goal

- Raise the probability that a visible vector with be faithfully reconstructed when a 'hidden' vector is presented to the visible layer.

Question:

How do we train an RBM so that 'reconstructions' are likely to create a reasonable facsimile of the original data?

Weights → Energies → Probabilities

- Each possible joint configuration of the visible and hidden units has an energy
  - The energy is determined by the weights and biases (as in a Hopfield net).
- The energy of a joint configuration of the visible and hidden units determines its probability:

$$p(\mathbf{v},\mathbf{h}) \propto e^{-E(\mathbf{v},\mathbf{h})}$$

- The probability of a configuration over the visible units is found by summing the probabilities of all the joint configurations that contain it.

From Hinton 2007 (modified)

Using energies to define probabilities

- The probability of a joint configuration over both visible and hidden units depends on the energy of that joint configuration compared with the energy of all other joint configurations.
- The probability of a configuration of the visible units is the sum of the probabilities of all the joint configurations that contain it.

$$p(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\underbrace{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}_{\text{partition function}}}$$

$$p(\mathbf{v}) = \frac{\sum_{g} e^{-E(\mathbf{v},\mathbf{h}^g)}}{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}$$

From Hinton 2007 (modified)

How Do We Train an RBM?

$$p(\mathbf{v}) = \frac{\sum_{g} e^{-E(\mathbf{v},\mathbf{h}^g)}}{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}} \rightarrow \ln p(\mathbf{v}) = \ln\left(\frac{\sum_{g} e^{-E(\mathbf{v},\mathbf{h}^g)}}{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}\right)$$

$$= \ln \sum_{g} e^{-E(\mathbf{v},\mathbf{h}^g)} - \ln \sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}$$

$$\frac{\partial \ln p(\mathbf{v})}{\partial w_{ij}} = \underbrace{\frac{\partial}{\partial w_{ij}} \ln \sum_{g} e^{-E(\mathbf{v},\mathbf{h}^g)}}_{A} - \underbrace{\frac{\partial}{\partial w_{ij}} \ln \sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}_{B}$$

Looking at Term A:

$$\frac{\partial}{\partial w_{ij}} \ln \sum_g e^{-E(\mathbf{v},\mathbf{h}^g)} = \frac{1}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} \cdot \frac{\partial}{\partial w_{ij}} \sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}$$

$$= \frac{1}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} \cdot \sum_g \frac{\partial e^{-E(\mathbf{v},\mathbf{h}^g)}}{\partial w_{ij}}$$

$$= \frac{1}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} \cdot \sum_g e^{-E(\mathbf{v},\mathbf{h}^g)} \cdot \frac{\partial\left(-E(\mathbf{v},\mathbf{h}^g)\right)}{\partial w_{ij}}$$

Looking at Term A:

Recall that $$E(\mathbf{v},\mathbf{h}) = -\sum_{i,j} v_i h_j w_{ij}$$

$$\frac{\partial}{\partial w_{ij}} \ln \sum_g e^{-E(\mathbf{v},\mathbf{h}^g)} = \frac{1}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} \cdot \sum_g e^{-E(\mathbf{v},\mathbf{h}^g)} \cdot \frac{\partial\left(-E(\mathbf{v},\mathbf{h}^g)\right)}{\partial w_{ij}}$$

$$= \frac{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)} v_i h_j^g}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} = \sum_g p(\mathbf{h}^g \mid \mathbf{v}) v_i h_j^g = \langle v_i \cdot h_j \rangle_{\mathbf{v}}$$

Where does that conditional probability come from?

$$\frac{e^{-E(\mathbf{v},\mathbf{h}^g)}}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} = p(\mathbf{h}^g|\mathbf{v})$$

$$p(\mathbf{h}^g|\mathbf{v}) = \frac{p(\mathbf{v},\mathbf{h}^g)}{p(\mathbf{v})} = \frac{\frac{1}{Z}e^{-E(\mathbf{v},\mathbf{h}^g)}}{\frac{1}{Z}\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}} = \frac{e^{-E(\mathbf{v},\mathbf{h}^g)}}{\sum_g e^{-E(\mathbf{v},\mathbf{h}^g)}}$$

Looking at Term B:

$$\frac{\partial}{\partial w_{ij}} \ln \sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)} = \frac{1}{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}} \cdot \frac{\partial}{\partial w_{ij}} \sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}$$

$$= \frac{1}{\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}} \cdot \sum_{u,g} \frac{\partial e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}{\partial w_{ij}}$$

Looking at Term B:

$$\frac{\partial}{\partial w_{ij}} \ln \sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)} = \frac{\displaystyle\sum_{u,g} \frac{\partial e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}{\partial w_{ij}}}{\displaystyle\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}$$

$$= \frac{\displaystyle\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)} \frac{\partial\left(-E(\mathbf{v}^u,\mathbf{h}^g)\right)}{\partial w_{ij}}}{\displaystyle\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}}$$

$$= \frac{\displaystyle\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)} v_i^u h_j^g}{\displaystyle\sum_{u,g} e^{-E(\mathbf{v}^u,\mathbf{h}^g)}} = \sum_{u,g} p(\mathbf{v}^u,\mathbf{h}^g) v_i^u h_j^g = \left\langle v_i \cdot h_j \right\rangle_{\mathbf{vh}}$$

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Basis of *Contrastive Divergence*

$$\frac{\partial \ln p(\mathbf{v})}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \ln \sum_{g} e^{-E(\mathbf{v},\mathbf{h}^{g})} - \frac{\partial}{\partial w_{ij}} \ln \sum_{u,g} e^{-E(\mathbf{v}^{u},\mathbf{h}^{g})}$$

$$= \left\langle v_i \cdot h_j \right\rangle_{\mathbf{v}} - \left\langle v_i \cdot h_j \right\rangle_{\mathbf{vh}}$$

Summary

- Showed the derivative of the log probability with respect to weights
- This can serve as the basis of a gradient ascent method for increasing the probability of the reconstructed vector v.