CPSC 550: Machine Learning II **2008/9 Term 2**

Lecture 10 — Feb. 24

Lecturer: Nando de Freitas *Scribe: Gustavo Lacerda*

Boltzmann Machines (Random / Gibbs Fields)

Notation

$$\begin{aligned}\mathbf{h} &\equiv \text{latent/hidden random variables}\\ \mathbf{v} &\equiv \text{visible random variables}\\ \mathbf{w} &\equiv \text{parameters}\\ \sigma(t) = \text{logit}(t) &= \frac{1}{1+e^{-t}} \quad \text{(the logistic function)}\end{aligned}$$

General random fields

A random field is described by the joint probability equation:

$$p_{\mathbf{w}}(\mathbf{v},\mathbf{h}) = \frac{1}{Z(\mathbf{w})} e^{-E_{\mathbf{w}}(\mathbf{h},\mathbf{v})}$$

Here, E is called the *energy function*, and Z is the *partition function*. For continuous-valued nodes, we have:

$$Z(\mathbf{w}) = \iint e^{-E_{\mathbf{w}}(\mathbf{h},\mathbf{v})} dp_{\mathbf{w}}(\mathbf{v},\mathbf{h})$$

In general, inference and learning in Boltzmann machines is difficult due to the necessity of computing Z.

The Restricted Boltzmann Machine (RBM)

An RBM is a type of pairwise random field with missing edges. In particular, the field can be represented as a bipartite graph, where nodes in different partitions (layers) are fully connected but there are no edges between nodes in the same layer; an example is illustrated in Figure 10.1. The joint probability distribution of an RBM is given by:

$$p_{\mathbf{w}}(\mathbf{h},\mathbf{v}) = \frac{1}{Z(\mathbf{w})} \exp\left(\sum_{i=1}^{n_{\mathbf{v}}} b_i \mathbf{v}_i + \sum_{j=1}^{n_{\mathbf{h}}} b_j \mathbf{h}_j + \sum_{i=1}^{n_{\mathbf{v}}}\sum_{j=1}^{n_{\mathbf{h}}} \mathbf{v}_i \mathbf{w}_{ij} \mathbf{h}_j\right) \tag{10.1}$$

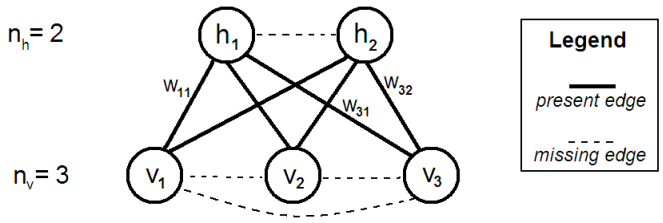

Figure 10.1. An RBM with 3 visible and 2 hidden nodes.

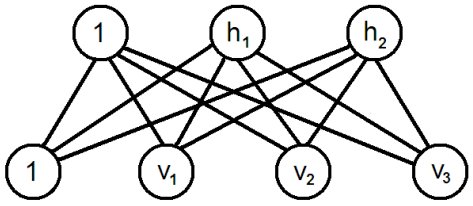

Figure 10.2. The RBM of Figure 10.1 with bias nodes shown.

Bias nodes

As illustrated in Figure 10.2, it is common to include a permanently activated *bias node* in each layer, whose value is always 1. This allows the first two terms of Equation 10.1 to be assimilated into the third:[1]

$$p_{\mathbf{w}}(\mathbf{h}, \mathbf{v}) = \frac{1}{Z(\mathbf{w})} \exp\left(\sum_{i=1}^{n_{\mathbf{v}}} \sum_{j=1}^{n_{\mathbf{h}}} \mathbf{v}_i \mathbf{w}_{ij} \mathbf{h}_j\right) \tag{10.2}$$

$$= \frac{1}{Z(\mathbf{w})} \prod_{i=1}^{n_{\mathbf{v}}} \prod_{j=1}^{n_{\mathbf{h}}} \exp\left(\mathbf{v}_i \mathbf{w}_{ij} \mathbf{h}_j\right) \tag{10.3}$$

$$= \frac{1}{Z(\mathbf{w})} \prod_{i=1}^{n_{\mathbf{v}}} \prod_{j=1}^{n_{\mathbf{h}}} \phi(\mathbf{v}_i, \mathbf{h}_j, \mathbf{w}_{ij}) \tag{10.4}$$

where $\phi(\cdot)$ in Equation 10.4 is known as a *potential function*.

Aside: connection between potentials and factor graphs

As illustrated in Figure 10.3, any distribution represented by a directed acyclic graph can be "moralized" to form an undirected graphical model, and then further factored to

[1] An analogous technique is used in linear regression, where it is common to add a column of 1s to the design matrix in order to avoid explicitly dealing with offset terms.

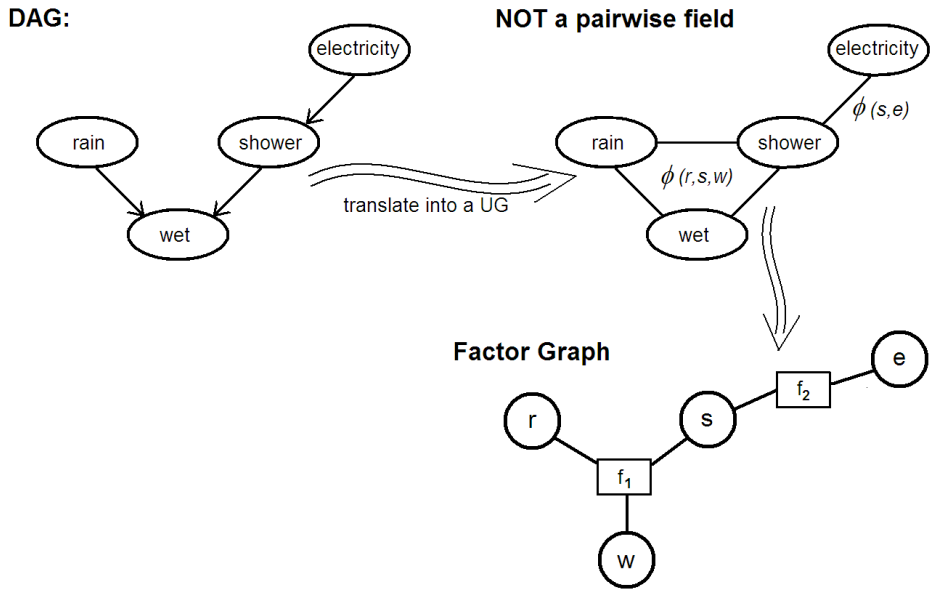

Figure 10.3. A probabilistic model respresented as a directed acyclic graph, an undirected graphical model, and a factor graph. One can see that this is not a pairwise field by the fact that f_1 connects to 3 variables.

create a factor graph. Potential functions correspond to these factors; for this example:

$$\begin{aligned} p(w,s,r,e) &= p(w|s,r)p(s|e)p(r)p(e) \\ &= \phi(s,e)\phi(r,s,w) \\ &= f_1 f_2 \end{aligned}$$

Pairwise fields and the exponential family

We can show that pairwise fields are in the exponential family by first rewriting Equation 10.3 as:

$$p_{\mathbf{w}}(\mathbf{h},\mathbf{v}) = \frac{1}{Z(\mathbf{w})} \prod_i \prod_j \exp\left(\langle \mathbf{w}, f(\mathbf{h},\mathbf{v})\rangle\right) = \frac{1}{Z(\mathbf{w})} \exp\left(f'(\mathbf{v})^T \mathbf{w} f'(\mathbf{h})\right)$$

where $f'(\cdot)$ are *sufficient statistics* (features), and $\mathbf{w}$ are the *natural parameters*. It is left as an exercise for the reader to complete showing that RBMs are in the exponential family.

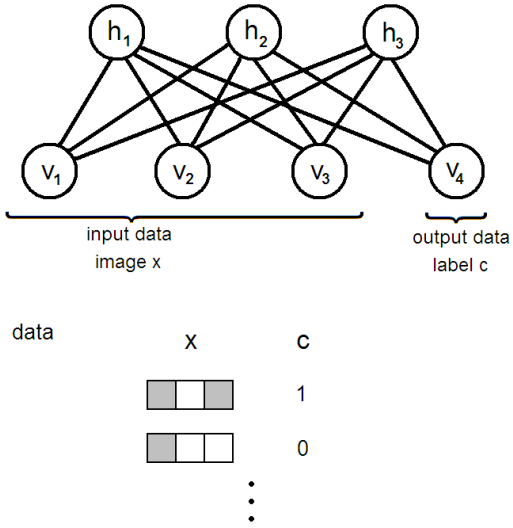

Figure 10.4. Binary-valued RBM for pattern completion.

Application: Pattern completion

Reference: (Hopfield; Duda & Hart)

Consider the RBM of Figure 10.4, which is binary-valued; i.e.:

$$\mathbf{h}_i \in \{0,1\}, \mathbf{v}_i \in \{0,1\}$$

By first computing values for the hidden nodes $\mathbf{h}$ from the data $\mathbf{v} = \mathbf{x}$ and initial parameters $\mathbf{w}$, simulated points $\tilde{\mathbf{v}}$ can be generated. By minimizing the difference between these "hallucinations" and the true data, the model can be trained:

$$\mathbf{v} \rightarrow \mathbf{h}$$
$$\mathbf{h} \rightarrow \tilde{\mathbf{v}}$$
$$\min_{\mathbf{w},\mathbf{h}} ||\mathbf{v} - \tilde{\mathbf{v}}||$$

At training time, some subset of the input is given, and the model is used to simulate the missing input bits.

Update equations

From Equation 10.2 we can express the conditional probabilities for this simple network:

$$p_{\mathbf{w}}(\mathbf{h}_j = 1|\mathbf{v}) \propto \exp\left(\sum_{i=1}^{n_{\mathbf{v}}} \mathbf{v}_i \mathbf{w}_{ij}(1)\right) = \exp\left(\mathbf{v}^T \mathbf{w}_j\right) \tag{10.5}$$

$$p_{\mathbf{w}}(\mathbf{h}_j = 0|\mathbf{v}) \propto \exp\left(\sum_{i=1}^{n_{\mathbf{v}}} \mathbf{v}_i \mathbf{w}_{ij}(0)\right) = 1 \tag{10.6}$$

Normalizing:

$$\begin{aligned} p_{\mathbf{w}}(\mathbf{h}_j = 1|\mathbf{v}) &= \frac{\exp\left(\mathbf{v}^T \mathbf{w}_j\right)}{1 + \exp\left(\mathbf{v}^T \mathbf{w}_j\right)} = \frac{1}{1 + \exp\left(-\mathbf{v}^T \mathbf{w}_j\right)} \\ &= \sigma\left(\mathbf{v}^T \mathbf{w}_j\right) \qquad (10.7) \\ p_{\mathbf{w}}(\mathbf{h}_j = 0|\mathbf{v}) &= 1 - \sigma\left(\mathbf{v}^T \mathbf{w}_j\right) \qquad (10.8) \end{aligned}$$

So we can write the probability of $\mathbf{h}$ given $\mathbf{v}$ as:

$$\begin{aligned} p_{\mathbf{w}}(\mathbf{h}|\mathbf{v}) &= \prod_{j=1}^{n_{\mathbf{h}}} p_{\mathbf{w}}(\mathbf{h}_j|\mathbf{v}) \\ &= \prod_{j=1}^{n_h} \sigma(\mathbf{v}^T \mathbf{w}_j)^{\mathbf{h}_j} (1 - \sigma(\mathbf{v}^T \mathbf{w}_j))^{1-\mathbf{h}_j} \end{aligned} \tag{10.9}$$

Similarly, for $\mathbf{v}$ given $\mathbf{h}$:

$$p_{\mathbf{w}}(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n_{\mathbf{v}}} \sigma(\mathbf{w}_i \mathbf{h})^{\mathbf{v}_i} (1 - \sigma(\mathbf{w}_i \mathbf{h}))^{1-\mathbf{v}_i} \tag{10.10}$$

This second derivation is left as an exercise for the reader.

Graphical interpretation of weights

If the weight vector $\mathbf{w}_j$ corresponding to a single hidden unit $\mathbf{h}_j$ is plotted as an image, the response of that unit to input features can be visualized. This is especially revealing when the input bits correspond to image pixels, where such plots show that hidden units often act as "edge detectors" – highly sensitive to specific regions of the input image. An example of this is taken from "Classification using Discriminative Restricted Boltzmann Machines" (Larochelle and Bengio, ICML 2008):