Gaussian Restricted Boltzmann Classifier

G.A. Jarrad

August 19, 2018

1 Definition

Consider a restricted Boltzmann machine (RBM) with a real-valued input layer, a binary-valued hidden layer, and a binary-valued output layer, as shown in Figure **??**. A suitable energy function is given by

$$E(\mathbf{x},\mathbf{h},\mathbf{y};\Theta) \;=\; \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T W\mathbf{x} - \mathbf{c}^T\mathbf{y} - \mathbf{h}^T U\mathbf{y}\,, \tag{1.1}$$

with input feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_F) \in \mathbb{X} \subseteq \mathbb{R}^F$, hidden binary vector $\mathbf{h} = (h_1, h_2, \ldots, h_H) \in \mathbb{H} = \{0,1\}^H$, and output binary vector $\mathbf{y} = (y_1, y_2, \ldots, y_C) \in \mathbb{Y} = \{0,1\}^C$. The model parameters are $\Theta = (\mathbf{a}, \mathbf{b}, \mathbf{c}, W, U)$. The joint probability of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{h}$ is then

$$p(\mathbf{x},\mathbf{h},\mathbf{y} \mid \Theta) \;=\; \frac{e^{-E(\mathbf{x},\mathbf{h},\mathbf{y};\Theta)}}{\int_{\mathbb{X}} \sum_{\mathbf{h}'\in\mathbb{H}} \sum_{\mathbf{y}'\in\mathbb{Y}} e^{-E(\mathbf{x}',\mathbf{h}',\mathbf{y}';\Theta)}\, d|\mathbf{x}'|}\,, \tag{1.2}$$

which is intractible to compute in general.

In order to turn the RBM into a restricted Boltzmann classifier (RBC), let us now suppose that the binary vector $\mathbf{y}$ is really a one-in-C vector of $C-1$ zeros and a single one, restricted to the set $\mathbb{Y}' = \{\mathbf{y} \in \mathbb{Y} \mid \sum_{k=1}^{C} y_k = 1\}$. Then there is a one-to-one correspondence between each vector $\mathbf{y} \in \mathbb{Y}'$ and some scalar $y \in \{1, 2, \ldots, C\}$, such that, for example, the term $U\mathbf{y}$ selects the y-th column of U, denoted by $\mathbf{u}_y$. Hence we obtain a final mapping to a multinomial output, suitable for a classifier. The joint probability then becomes

$$\begin{aligned} p(\mathbf{x},\mathbf{h},y \mid \Theta) \;&=\; \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{x} + c_y + \mathbf{h}^T\mathbf{u}_y}}{\int_{\mathbb{X}} \sum_{y'=1}^{C} \sum_{\mathbf{h}'\in\mathbb{H}} e^{-\frac{1}{2}\|\mathbf{x}'-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h}' + \mathbf{h}'^T W\mathbf{x}' + c_{y'} + \mathbf{h}^T\mathbf{u}_{y'}}\, d|\mathbf{x}'|} \\ &=\; \frac{e^{c_y - \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2} \prod_{i=1}^{H} e^{h_i(b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy})}}{\int_{\mathbb{X}} \sum_{y'=1}^{C} e^{c_{y'} - \frac{1}{2}\|\mathbf{x}'-\mathbf{a}\|^2} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}\right] d|\mathbf{x}'|}\,, \end{aligned} \tag{1.3}$$

where $\mathbf{w}_i^T$ is the i-th row of W. The discriminative form of the RBC can then be specified as

$$\begin{aligned} p(y \mid \mathbf{x}, \Theta) \;&=\; \frac{\sum_{\mathbf{h}'\in\mathbb{H}} p(\mathbf{x},\mathbf{h}',y \mid \Theta)}{\sum_{y'=1}^{C} \sum_{\mathbf{h}'\in\mathbb{H}} p(\mathbf{x},\mathbf{h}',y' \mid \Theta)} \\ &=\; \frac{e^{c_y - \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}\right]}{\sum_{y'=1}^{C} e^{c_{y'} - \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy'}}\right]} \\ &=\; \frac{e^{c_y} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}\right]}{\sum_{y'=1}^{C} e^{c_{y'}} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy'}}\right]}\,, \end{aligned} \tag{1.4}$$

which is a nonlinear form of logistic classifier.

Now, the bipartite restriction depicted in Figure **??** ensures that $\mathbf{x}$ and y are conditionally independent

given $\mathbf{h}$. Observe, for instance, that

$$\begin{aligned}
p(\mathbf{x} \mid \mathbf{h}, y, \Theta) &= \frac{p(\mathbf{x}, \mathbf{h}, y \mid \Theta)}{\int_{\mathbb{X}} p(\mathbf{x}', \mathbf{h}, y \mid \Theta)\, d|\mathbf{x}'|} \\
&= \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{x} + c_y + \mathbf{h}^T\mathbf{u}_y}}{\int_{\mathbb{X}} e^{-\frac{1}{2}\|\mathbf{x}'-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{x}' + c_y + \mathbf{h}^T\mathbf{u}_y}\, d|\mathbf{x}'|} \\
&= \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \mathbf{h}^T W\mathbf{x}}}{\int_{\mathbb{X}} e^{-\frac{1}{2}\|\mathbf{x}'-\mathbf{a}\|^2 + \mathbf{h}^T W\mathbf{x}'}\, d|\mathbf{x}'|} \\
&= \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}-W^T\mathbf{h}\|^2 + \mathbf{h}^T W\mathbf{a} + \frac{1}{2}\mathbf{h}^T WW^T\mathbf{h}}}{\int_{\mathbb{X}} e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}-W^T\mathbf{h}\|^2 + \mathbf{h}^T W\mathbf{a} + \frac{1}{2}\mathbf{h}^T WW^T\mathbf{h}}\, d|\mathbf{x}'|} \\
&= \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}-W^T\mathbf{h}\|^2}}{\int_{\mathbb{X}} e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}-W^T\mathbf{h}\|^2}\, d|\mathbf{x}'|} \\
&= N(\mathbf{x} \mid \mathbf{a} + W^T\mathbf{h}, I)\,. \qquad (1.5)
\end{aligned}$$

Hence, $\mathbf{x}$ is conditionally normally distributed with mean $\mathbf{a} + W^T\mathbf{h}$ and unit spherical variance I (the identity matrix).

Similarly, observe that

$$\begin{aligned}
p(y \mid \mathbf{x}, \mathbf{h}, \Theta) &= \frac{p(\mathbf{x}, \mathbf{h}, y \mid \Theta)}{\sum_{y'=1}^{C} p(\mathbf{x}, \mathbf{h}, y' \mid \Theta)} \\
&= \frac{e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{x} + c_y + \mathbf{h}^T\mathbf{u}_y}}{\sum_{y'=1}^{C} e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{x} + c_{y'} + \mathbf{h}^T\mathbf{u}_{y'}}} \\
&= \frac{e^{c_y + \mathbf{h}^T\mathbf{u}_y}}{\sum_{y'=1}^{C} e^{c_{y'} + \mathbf{h}^T\mathbf{u}_{y'}}}\,. \qquad (1.6)
\end{aligned}$$

This result is just the *soft-max* function, or standard logistic classifier.

Conversely, $\mathbf{h}$ depends upon both $\mathbf{x}$ and $\mathbf{y}$ via

$$\begin{aligned}
p(\mathbf{h} \mid \mathbf{x}, y, \Theta) &= \frac{p(\mathbf{x}, \mathbf{h}, y \mid \Theta)}{\sum_{\mathbf{h}' \in \mathbb{H}} p(\mathbf{x}, \mathbf{h}', y \mid \Theta)} \\
&= \frac{e^{c_y - \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2} \prod_{i=1}^{H} e^{h_i(b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy})}}{e^{c_y - \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2} \prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}\right]} \\
&= \frac{\prod_{i=1}^{H} e^{h_i(b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy})}}{\prod_{i=1}^{H} \left[1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}\right]} \\
&= \prod_{i=1}^{H} p(h_i \mid \mathbf{x}, y, \Theta)\,, \qquad (1.7)
\end{aligned}$$

where

$$p(h_i = 1 \mid \mathbf{x}, y, \Theta) = \frac{e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}}{1 + e^{b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy}}} = \sigma(b_i + \mathbf{w}_i^T\mathbf{x} + u_{iy})\,. \qquad (1.8)$$

This is just the logistic sigmoid function.

2 Supervised Discriminative Optimisation

Consider the problem of estimating the RBC parameters Θ from a data-set of fully labelled feature vectors, $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, with corresponding labels $\mathcal{Y} = (y_1, y_2, \ldots, y_N)$. Assuming that the data items are independent, the discriminative likelihood is given by

$$p(\mathcal{Y} \mid \mathcal{X}, \Theta) = \prod_{d=1}^{N} p(y_d \mid \mathbf{x}_d, \Theta)\,, \qquad (2.1)$$

and hence, from equation (1.4), the average discriminative log-likelihood is given by

$$\begin{aligned}
\mathcal{L}_{\mathcal{Y}|\mathcal{X}}(\Theta) &= \frac{1}{N}\ln p(\mathcal{Y} \mid \mathcal{X}, \Theta) = \frac{1}{N}\sum_{d=1}^{N} \ln p(y_d \mid \mathbf{x}_d, \Theta) \\
&= \frac{1}{N}\sum_{d=1}^{N}\left\{\sum_{y'=1}^{C}\delta_{y',y_d}\ln\left(e^{c_{y'}}\prod_{i=1}^{H}\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy'}}\right]\right)\right. \\
&\quad \left. - \ln\sum_{y'=1}^{C}e^{c_{y'}}\prod_{i=1}^{H}\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy'}}\right]\right\} \\
&= \frac{1}{N}\sum_{d=1}^{N}\left\{\sum_{y'=1}^{C}\delta_{y',y_d}\left(c_{y'}+\sum_{i=1}^{H}\ln\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy'}}\right]\right)\right. \\
&\quad \left. - \ln\sum_{y'=1}^{C}e^{c_{y'}}\prod_{i=1}^{H}\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy'}}\right]\right\}. \qquad (2.2)
\end{aligned}$$

Hence, the gradient with respect to c_y is

$$\begin{aligned}
\frac{\partial\mathcal{L}_{\mathcal{Y}|\mathcal{X}}}{\partial c_y} &= \frac{1}{N}\sum_{d=1}^{N}\left\{\delta_{y,y_d} - \frac{e^{c_y}\prod_{i=1}^{H}\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}}\right]}{\sum_{y'=1}^{C}e^{c_{y'}}\prod_{i=1}^{H}\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy'}}\right]}\right\} \\
&= \frac{1}{N}\sum_{d=1}^{N}\left\{\delta_{y,y_d} - p(y \mid \mathbf{x}_d, \Theta)\right\} \\
&= \frac{N_y}{N} - \frac{1}{N}\sum_{d=1}^{N} p(y \mid \mathbf{x}_d, \Theta), \qquad (2.3)
\end{aligned}$$

where N_y is the number of data labelled with class y.

In order to develop the remaining derivatives, we first observe that

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\left(c_y + \sum_{i'=1}^{H}\ln\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right]\right) &= \frac{\partial}{\partial\theta_i}\ln\left[1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}}\right] \\
&= \frac{e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}}}{1+e^{b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}}}\frac{\partial}{\partial\theta_i}(b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}) \\
&= p(h_i=1 \mid \mathbf{x}_d, y, \Theta)\frac{\partial}{\partial\theta_i}(b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}), \qquad (2.4)
\end{aligned}$$

from equation (1.8), and then use the fact that $\nabla f(\theta) = f(\theta)\nabla\ln f(\theta)$ to deduce that

$$\begin{aligned}
&\tfrac{\partial}{\partial\theta_i}e^{c_y}\textstyle\prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right] \\
&\quad = e^{c_y}\prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right]\frac{\partial}{\partial\theta_i}\left(c_y+\sum_{i'=1}^{H}\ln\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right]\right) \\
&\quad = e^{c_y}\prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right]p(h_i=1 \mid \mathbf{x}_d, y, \Theta)\frac{\partial}{\partial\theta_i}(b_i+\mathbf{w}_i^T\mathbf{x}_d+u_{iy}). \qquad (2.5)
\end{aligned}$$

Hence, the gradient of the log-likelihood with respect to u_{iy} is

$$\begin{aligned}
\frac{\partial\mathcal{L}_{\mathcal{Y}|\mathcal{X}}}{\partial u_{iy}} &= \frac{1}{N}\sum_{d=1}^{N}\left\{\delta_{y,y_d}\,p(h_i=1 \mid \mathbf{x}_d, y, \Theta)\right. \\
&\quad \left. - \frac{e^{c_y}\prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y}}\right]}{\sum_{y'=1}^{C}e^{c_{y'}}\prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y'}}\right]}p(h_i=1 \mid \mathbf{x}_d, y, \Theta)\right\} \\
&= \frac{1}{N}\sum_{d=1}^{N}\left\{\delta_{y,y_d} - p(y \mid \mathbf{x}_d, \Theta)\right\}p(h_i=1 \mid \mathbf{x}_d, y, \Theta). \qquad (2.6)
\end{aligned}$$

Similarly, the gradient of the log-likelihood with respect to b_i is

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathcal{Y}|\mathcal{X}}}{\partial b_i} &= \frac{1}{N}\sum_{d=1}^{N}\left\{\sum_{y'=1}^{C}\delta_{y',y_d}\, p(h_i=1 \mid \mathbf{x}_d, y', \Theta)\right. \\
&\quad \left. - \frac{\sum_{y'=1}^{C} e^{c_{y'}} \prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y'}}\right]\, p(h_i=1 \mid \mathbf{x}_d, y', \Theta)}{\sum_{y'=1}^{C} e^{c_{y'}} \prod_{i'=1}^{H}\left[1+e^{b_{i'}+\mathbf{w}_{i'}^T\mathbf{x}_d+u_{i'y'}}\right]}\right\} \\
&= \frac{1}{N}\sum_{d=1}^{N}\sum_{y'=1}^{C}\left\{\delta_{y',y_d} - p(y' \mid \mathbf{x}_d, \Theta)\right\}\, p(h_i=1 \mid \mathbf{x}_d, y', \Theta) \\
&= \sum_{y=1}^{C}\frac{\partial \mathcal{L}_{\mathcal{Y}|\mathcal{X}}}{\partial u_{iy}}\,, \qquad (2.7)
\end{aligned}$$

and the gradient with respect to $\mathbf{w}_i$ is

$$\frac{\partial \mathcal{L}_{\mathcal{Y}|\mathcal{X}}}{\partial \mathbf{w}_i} = \frac{1}{N}\sum_{d=1}^{N}\mathbf{x}_d\sum_{y'=1}^{C}\left\{\delta_{y',y_d} - p(y' \mid \mathbf{x}_d, \Theta)\right\}\, p(h_i=1 \mid \mathbf{x}_d, y', \Theta)\,. \qquad (2.8)$$

Consequently, the discriminative log-likelihood $\mathcal{L}_{\mathcal{Y}|\mathcal{X}}$ can be maximised using standard or accelerated gradient ascent. Note, however, that the parameter $\mathbf{a}$ from equation (1.1) does not appear in the RBC (1.4), and therefore cannot be optimised discriminatively.