rbm_classifiers

February 15, 2022

1 Restricted Boltzmann Classifiers

This notebook is based largely on the (Larochelle and Bengio) paper on discriminative RBMs. The related notes in (Jarrad) were an attempt to see what the Gaussian input version looked like, in contrast to the usual Bernoulli input. A thorough derivation of RBMs from first principles is given by (Jarrad$_2$). A thorough derivation of the properties of discrete Boltzmann machines from first principles is given by (Jarrad$_3$).

1.0.1 Composition of RBMs

The basic idea is to compose together two RBMs. However, rather than simply feed the output of the first RBM into the input of the second RBM, the two RBMs are merged to share the same hidden layer (see the figure below). The first RBM model takes a known input vector $\mathbf{x} = (x_1, x_2, \ldots, x_F) \in \mathcal{X}$ and outputs the expectations of the hidden vector $\mathbf{z} = (z_1, z_2, \ldots, z_H) \in \mathcal{Z}$. The second RBM model then takes $\mathbf{z}$ as input and outputs the expectations of the vector $\mathbf{y} = (y_1, y_2, \ldots, y_C) \in \mathcal{Y}$.

Since the input and output layers of an RBM are connected via undirected edges to form a bipartite graph, the first RBM (taken by itself) has the property (with caveats) that the distributions of the elements of $\mathbf{x}$ are conditionally independent given $\mathbf{z}$, and the distributions of $\mathbf{z}$ are conditionally independent given $\mathbf{x}$. Similarly, the second RBM (taken alone) has the same conditional independence property between $\mathbf{z}$ and $\mathbf{y}$.

However, the composition of the two RBMs changes these independence properties somewhat. We retain that the elements of $\mathbf{x}$ are conditionally independent given $\mathbf{z}$, and that the elements of $\mathbf{y}$ are also conditionally independent given $\mathbf{z}$. However, the converse is no longer true - the composed model now requires knowledge of both $\mathbf{x}$ and $\mathbf{y}$ for the elements of $\mathbf{z}$ to be independent. In addition, the composed model has the further property of conditional independence between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$.

Consequently, the joint probability distribution takes the form

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \doteq \frac{e^{f(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}' \in \mathcal{X}} \sum_{\mathbf{y}' \in \mathcal{Y}} \sum_{\mathbf{z}' \in \mathcal{Z}} e^{f(\mathbf{x}',\mathbf{y}',\mathbf{z}')}} = \frac{e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{x}' \in \mathcal{X}} \sum_{\mathbf{y}' \in \mathcal{Y}} \sum_{\mathbf{z}' \in \mathcal{Z}} e^{\mathbf{a}^T\mathbf{x}'+\mathbf{b}^T\mathbf{z}'+\mathbf{x}'^T\mathbf{W}\mathbf{z}'+\mathbf{c}^T\mathbf{y}'+\mathbf{z}'^T\mathbf{U}\mathbf{y}'}} \quad (1)$$

For convenience, we have assumed discrete values for $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. If continuous values are required instead, then the relevant summations will be replaced by integrations.

As a demonstration of the aforementioned conditional independence properties, observe that

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \frac{e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}'+\mathbf{z}^T\mathbf{U}\mathbf{y}'}} \tag{2}$$

$$= \frac{e^{\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\mathbf{c}^T\mathbf{y}'+\mathbf{z}^T\mathbf{U}\mathbf{y}'}} = p(\mathbf{y} \mid \mathbf{z})\,. \tag{3}$$

Further note that the exponent is a linear combination of the elements of $\mathbf{y}$. Hence, we let $\mathbf{U}_{:,j}$ denote the j-th column of $\mathbf{U}$, such that

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{\prod_{j=1}^{C} e^{y_j\left(c_j+\mathbf{z}^T\mathbf{U}_{:,j}\right)}}{\sum_{\mathbf{y}'\in\mathcal{Y}} \prod_{j=1}^{C} e^{y'_j(c_j+\mathbf{z}^T\mathbf{U}_{:,j})}} \tag{4}$$

$$= \frac{\prod_{j=1}^{C} e^{y_j\left(c_j+\mathbf{z}^T\mathbf{U}_{:,j}\right)}}{\prod_{j=1}^{C} \sum_{y'_j\in\mathcal{Y}_j} e^{y'_j(c_j+\mathbf{z}^T\mathbf{U}_{:,j})}} \tag{5}$$

$$= \prod_{j=1}^{C} p(y_j \mid \mathbf{z})\,, \tag{6}$$

where

$$p(y_j \mid \mathbf{z}) = \frac{e^{y_j\left(c_j+\mathbf{z}^T\mathbf{U}_{:,j}\right)}}{\sum_{y'_j\in\mathcal{Y}_j} e^{y'_j(c_j+\mathbf{z}^T\mathbf{U}_{:,j})}}\,. \tag{7}$$

Note that this conditional independence of elements relies on the assumption that we can partition the space via

$$\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_C\,. \tag{8}$$

Similarly, observe that

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{y}) = \frac{e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\mathbf{a}^T\mathbf{x}'+\mathbf{b}^T\mathbf{z}+\mathbf{x}'^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}} \tag{9}$$

$$= \frac{e^{\mathbf{a}^T\mathbf{x}+\mathbf{x}^T\mathbf{W}\mathbf{z}}}{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\mathbf{a}^T\mathbf{x}'+\mathbf{x}'^T\mathbf{W}\mathbf{z}}} = p(\mathbf{x} \mid \mathbf{z})\,. \tag{10}$$

Hence, we let $\mathbf{W}_{i,:}$ denote the i-th row of $\mathbf{W}$, such that

$$p(\mathbf{x} \mid \mathbf{z}) = \frac{\prod_{i=1}^{F} e^{x_i(a_i+\mathbf{W}_{i,:}\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}} \prod_{i=1}^{F} e^{x'_i(a_i+\mathbf{W}_{i,:}\mathbf{z})}} \tag{11}$$

$$= \frac{\prod_{i=1}^{F} e^{x_i(a_i+\mathbf{W}_{i,:}\mathbf{z})}}{\prod_{i=1}^{F} \sum_{x'_i\in\mathcal{X}_i} e^{x'_i(a_i+\mathbf{W}_{i,:}\mathbf{z})}} \tag{12}$$

$$= \prod_{i=1}^{F} p(x_i \mid \mathbf{z})\,, \tag{13}$$

where

$$p(x_i \mid \mathbf{z}) \;=\; \frac{e^{x_i(a_i+\mathbf{W}_{i,:}\mathbf{z})}}{\sum_{x_i' \in \mathcal{X}_i} e^{x_i'(a_i+\mathbf{W}_{i,:}\mathbf{z})}}\,. \tag{14}$$

Again, we note that this conditional independence of elements relies on the assumption that we can partition the space via

$$\mathcal{X} \;=\; \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_F\,. \tag{15}$$

Conversely, we observe that

$$\begin{aligned}
p(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \;&=\; \frac{e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{c}^T\mathbf{y}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{z}' \in \mathcal{Z}} e^{\mathbf{a}^T\mathbf{x}+\mathbf{b}^T\mathbf{z}'+\mathbf{x}^T\mathbf{W}\mathbf{z}'+\mathbf{c}^T\mathbf{y}+\mathbf{z}'^T\mathbf{U}\mathbf{y}}} && (16)\\
&=\; \frac{e^{\mathbf{b}^T\mathbf{z}+\mathbf{x}^T\mathbf{W}\mathbf{z}+\mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{\mathbf{z}' \in \mathcal{Z}} e^{\mathbf{b}^T\mathbf{z}'+\mathbf{x}^T\mathbf{W}\mathbf{z}'+\mathbf{z}'^T\mathbf{U}\mathbf{y}}} && (17)\\
&=\; \frac{\prod_{k=1}^{H} e^{z_k\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}}{\sum_{\mathbf{z}' \in \mathcal{Z}} \prod_{k=1}^{H} e^{z_k'\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}} && (18)\\
&=\; \frac{\prod_{k=1}^{H} e^{z_k\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}}{\prod_{k=1}^{H} \sum_{z_k' \in \mathcal{Z}_k} e^{z_k'\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}} && (19)\\
&=\; \prod_{k=1}^{H} p(z_k \mid \mathbf{x}, \mathbf{y})\,, && (20)
\end{aligned}$$

where

$$p(z_k \mid \mathbf{x}, \mathbf{y}) \;=\; \frac{e^{z_k\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}}{\sum_{z_k' \in \mathcal{Z}_k} e^{z_k'\left(b_k+\mathbf{x}^T\mathbf{W}_{:,k}+\mathbf{U}_{k,:}\mathbf{y}\right)}}\,. \tag{21}$$

Once again, this conditional independence of elements relies on the assumption that we can partition the space via

$$\mathcal{Z} \;=\; \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_H\,. \tag{22}$$

1.0.2 Discriminative RBM

Since $\mathbf{z}$ is unknown in practice, the ultimate purpose of the composite RBM is to predict output $\mathbf{y}$ based on known input $\mathbf{x}$. The discriminative probability is thus

$$
\begin{aligned}
p(\mathbf{y} \mid \mathbf{x}) &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}) && (23)\\
&= \frac{\sum_{\mathbf{z} \in \mathcal{Z}} e^{\mathbf{a}^T \mathbf{x} + \mathbf{b}^T \mathbf{z} + \mathbf{x}^T \mathbf{W} \mathbf{z} + \mathbf{c}^T \mathbf{y} + \mathbf{z}^T \mathbf{U} \mathbf{y}}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \sum_{\mathbf{z} \in \mathcal{Z}} e^{\mathbf{a}^T \mathbf{x} + \mathbf{b}^T \mathbf{z} + \mathbf{x}^T \mathbf{W} \mathbf{z} + \mathbf{c}^T \mathbf{y}' + \mathbf{z}^T \mathbf{U} \mathbf{y}'}} && (24)\\
&= \frac{e^{\mathbf{c}^T \mathbf{y}} \sum_{\mathbf{z} \in \mathcal{Z}} e^{\mathbf{b}^T \mathbf{z} + \mathbf{x}^T \mathbf{W} \mathbf{z} + \mathbf{z}^T \mathbf{U} \mathbf{y}}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\mathbf{c}^T \mathbf{y}'} \sum_{\mathbf{z} \in \mathcal{Z}} e^{\mathbf{b}^T \mathbf{z} + \mathbf{x}^T \mathbf{W} \mathbf{z} + \mathbf{z}^T \mathbf{U} \mathbf{y}'}} && (25)\\
&= \frac{e^{\mathbf{c}^T \mathbf{y}} \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{k=1}^{H} e^{z_k \left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y} \right)}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\mathbf{c}^T \mathbf{y}'} \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{k=1}^{H} e^{z_k \left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}' \right)}} && (26)\\
&= \frac{e^{\mathbf{c}^T \mathbf{y}} \prod_{k=1}^{H} \sum_{z_k \in \mathcal{Z}_k} e^{z_k \left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y} \right)}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\mathbf{c}^T \mathbf{y}'} \prod_{k=1}^{H} \sum_{z_k \in \mathcal{Z}_k} e^{z_k \left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}' \right)}} \,. && (27)
\end{aligned}
$$

1.0.3 Bernoulli hidden layer

It is traditional, and convenient for the purposes of tractability, to assume that the hidden layer takes binary-valued $\mathbf{z}$ vectors. Hence, we have $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_H = \{0, 1\}^H$, with the result that

$$
p(\mathbf{y} \mid \mathbf{x}) = \frac{e^{\mathbf{c}^T \mathbf{y}} \prod_{k=1}^{H} \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}} \right)}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\mathbf{c}^T \mathbf{y}'} \prod_{k=1}^{H} \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}'} \right)} \,. \tag{28}
$$

In addition, we note from the earlier derivation of $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ above that now

$$
\begin{aligned}
p(z_k = 1 \mid \mathbf{x}, \mathbf{y}) &= \frac{e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}}}{1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y}}} && (29)\\
&= \frac{1}{1 + e^{-\left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y} \right)}} && (30)\\
&= \sigma\left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:} \mathbf{y} \right) \,, && (31)
\end{aligned}
$$

where $\sigma(\cdot)$ is the sigmoid logistic function.

For later use, we can therefore define the expected value of element z_k as as

$$
\bar{z}_k(\mathbf{x}, \mathbf{y}) \doteq \mathbb{E}_{\mathbf{z} | \mathbf{x}, \mathbf{y}}[z_k] = p(z_k = 1 \mid \mathbf{x}, \mathbf{y}) \,. \tag{32}
$$

We may also collect these individual expected values together into the vector

$$
\bar{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \doteq \mathbb{E}_{\mathbf{z} | \mathbf{x}, \mathbf{y}}[\mathbf{z}] = [\bar{z}_k(\mathbf{x}, \mathbf{y})]_{k=1}^{H} \,. \tag{33}
$$

1.0.4 Restricted Boltzmann Classifier

Similarly to the hidden layer, we also assume that the output layer takes binary-valued $\mathbf{y}$ vectors. However, we further assume that each input $\mathbf{x}$ belongs to exactly one of C possible classes. Hence, $\mathbf{y}$

is a 1-of-C (or one-hot) vector, for which $C-1$ elements take the value 0 and exactly 1 element takes the value 1. Thus, we have $\mathcal{Y} = \{\mathbf{y} \in \{0,1\}^C \mid \sum_{j=1}^C y_j = 1\}$. It follows that the discriminative model above reduces to the probabilistic classifier

$$p(y_j = 1 \mid \mathbf{x}) \quad = \quad \frac{e^{c_j} \prod_{k=1}^H \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + u_{kj}}\right)}{\sum_{j'=1}^C e^{c_{j'}} \prod_{k=1}^H \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + u_{kj'}}\right)} \,. \tag{34}$$

For the purposes of comparison, we contrast this Restricted Boltzmann Classifier (RBC) with the traditional (linear) logistic classifier

$$p(y_j = 1 \mid \mathbf{x}) \quad = \quad \frac{e^{c_j + \mathbf{x}^T \mathbf{v}_j}}{\sum_{j'=1}^C e^{c_{j'} + \mathbf{x}^T \mathbf{v}_{j'}}} \,. \tag{35}$$

Hence, we observe that the presence of hidden units (even for $H = 1$, but especially for $H > 1$) provides additional nonlinearity to the logistic classifier.

It must be noted that the restriction imposed upon $\mathcal{Y}$ has the consequence that $\mathcal{Y} \neq \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_C$, and so this breaks the usual conditional independence property. Thus, from the earlier derivation of $p(\mathbf{y} \mid \mathbf{z})$ above, we instead have the probability

$$p(y_j = 1 \mid \mathbf{z}) \quad = \quad \frac{e^{c_j + \mathbf{z}^T \mathbf{U}_{:,j}}}{\sum_{j'=1}^C e^{c_{j'} + \mathbf{z}^T \mathbf{U}_{:,j'}}} \,. \tag{36}$$

However, we may still define the expected values

$$\bar{y}_j(\mathbf{z}) \quad \doteq \quad \mathbb{E}_{\mathbf{y}|\mathbf{z}}[y_j] = p(y_j = 1 \mid \mathbf{z}) \,, \tag{37}$$

and

$$\bar{\mathbf{y}}(\mathbf{z}) \quad \doteq \quad \mathbb{E}_{\mathbf{y}|\mathbf{z}}[\mathbf{y}] = [\bar{y}_j(\mathbf{z})]_{j=1}^C \,. \tag{38}$$

1.0.5 Modelling the input distribution

By design, the discriminative classifier above does not explicitly take into account the distribution of the input $\mathbf{x}$. This is manifested by the fact that, as noted by (Jarrad), the coefficients $\mathbf{a}$ do not appear in the classifier. However, there can be some implicit dependence upon the input distribution. For example, (Jarrad$_2$) showed that Gaussian inputs can be modelled with a standard (Bernoulli) RBM by replacing $\mathbf{x}$ on the right-hand side by $\tilde{\mathbf{x}} \doteq (x_1, \ldots, x_f, x_1^2, \ldots, x_F^2)$. This 'trick' can also be applied to the standard (linear) logistic classifier.

Clearly, to be able to estimate $\mathbf{a}$ from data, we need to optimise a different model than $p(\mathbf{y} \mid \mathbf{x})$, such as the joint model $p(\mathbf{x}, \mathbf{y})$. Empirically, my general observation, from past experiments on a wide variety of probabilistic classifiers, is that discriminative models have a tendency to overfit the training data, whereas joint models tend to be more robust, since they explicitly take account of the distribution of data points.

For convenience, we now make the choice that $\mathbf{x}$ is also binary-valued, such that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_F = \{0,1\}^F$. Hence, from the earlier derivation of $p(\mathbf{x} \mid \mathbf{z})$, we now have

$$p(x_i = 1 \mid \mathbf{z}) \quad = \quad \frac{e^{a_i + \mathbf{W}_{i,:}\mathbf{z}}}{1 + e^{a_i + \mathbf{W}_{i,:}\mathbf{z}}} = \sigma\left(a_i + \mathbf{W}_{i,:}\mathbf{z}\right) \,. \tag{39}$$

Thus, we may also define the expected values

$$\bar{x}_i(\mathbf{z}) \doteq \mathbb{E}_{\mathbf{x}|\mathbf{z}}[x_i] = p(x_i = 1 \mid \mathbf{z}), \tag{40}$$

and

$$\bar{\mathbf{x}}(\mathbf{z}) \doteq \mathbb{E}_{\mathbf{x}|\mathbf{z}}[\mathbf{x}] = [\bar{x}_i(\mathbf{z})]_{i=1}^{F}. \tag{41}$$

1.0.6 Optimising the joint likelihood

The RBC presented in this notebook is a discrete Bolztmann machine with a hidden layer, and we now wish to optimise the joint likelihood $p(\mathbf{x}, \mathbf{y})$. Hence, from (Jarrad$_3$) we obtain

$$\begin{aligned} \nabla \ln p(\mathbf{x}, \mathbf{y}) &\approx \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y}}\left[\nabla f(\mathbf{x}, \mathbf{y}, \mathbf{z})\right] - \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y}}\left[\mathbb{E}_{\mathbf{x}',\mathbf{y}'|\mathbf{z}}\left[\mathbb{E}_{\mathbf{z}'|\mathbf{x}',\mathbf{y}'}\left[\nabla f(\mathbf{x}', \mathbf{y}', \mathbf{z}')\right]\right]\right] && (42) \\ &\approx \nabla f(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}(\mathbf{x}, \mathbf{y})) - \nabla f(\bar{\mathbf{x}}', \bar{\mathbf{y}}', \bar{\mathbf{z}}(\mathbf{x}', \mathbf{y}')). && (43) \end{aligned}$$

Now, since we have demonstrated for the RBC that $\mathbf{x}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$, then we see that

$$\mathbb{E}_{\mathbf{x},\mathbf{y}|\mathbf{z}}[\cdot] = \mathbb{E}_{\mathbf{x}|\mathbf{z}}[\cdot]\,\mathbb{E}_{\mathbf{y}|\mathbf{z}}[\cdot], \tag{44}$$

such that $\bar{\mathbf{x}}' = \bar{\mathbf{x}}(\bar{\mathbf{z}}(\mathbf{x}, \mathbf{y}))$ and $\bar{\mathbf{y}}' = \bar{\mathbf{y}}(\bar{\mathbf{z}}(\mathbf{x}, \mathbf{y}))$.

To interpret this result, note that it corresponds to the procedure: 1. Push the known input $\mathbf{x}$ and output $\mathbf{y}$ into the hidden layer of the RBC, and compute $\bar{\mathbf{z}} = \bar{\mathbf{z}}(\mathbf{x}, \mathbf{y})$. 2. Push $\bar{\mathbf{z}}$ back out of the hidden layer into the input and output layers, and compute $\bar{\mathbf{x}}' = \bar{\mathbf{x}}(\bar{\mathbf{z}})$ and $\bar{\mathbf{y}}' = \bar{\mathbf{y}}(\bar{\mathbf{z}})$. 3. Push the reconstructed input $\bar{\mathbf{x}}'$ and output $\bar{\mathbf{y}}'$ back into the hidden layer, and compute $\bar{\mathbf{z}}' = \bar{\mathbf{z}}(\bar{\mathbf{x}}', \bar{\mathbf{y}}')$. 4. Compute the gradient as the difference of the data term $\nabla f(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}})$ and the reconstruction term $\nabla f(\bar{\mathbf{x}}', \bar{\mathbf{y}}', \bar{\mathbf{z}}')$.

In addition to approximating the gradient, we also need to approximate the joint likelihood of input $\mathbf{x}$ and output $\mathbf{y}$. Again from (Jarrad$_3$), we have

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &\approx p(\mathbf{x}, \mathbf{y} \mid \bar{\mathbf{z}}(\mathbf{x}, \mathbf{y})) && (45) \\ &= p(\mathbf{x} \mid \bar{z}(\mathbf{x}, \mathbf{y}))\, p(\mathbf{y} \mid \bar{z}(\mathbf{x}, \mathbf{y})), && (46) \end{aligned}$$

again due to the conditional independence of $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$.

1.0.7 Supervised RBC training

Putting together the various assumptions and derivations from above, the key expectations for a Bernoulli RBC are:

$$\bar{z}_k(\mathbf{x}, \mathbf{y}) = p(z_k = 1 \mid \mathbf{x}, \mathbf{y}) = \sigma\left(b_k + \mathbf{x}^T \mathbf{W}_{:,k} + \mathbf{U}_{k,:}\mathbf{y}\right), \tag{47}$$

for $k = 1, 2, \ldots, H$,

$$\bar{x}_i(\mathbf{z}) = p(x_i = 1 \mid \mathbf{z}) = \sigma\left(a_i + \mathbf{W}_{i,:}\mathbf{z}\right), \tag{48}$$

for $i = 1, 2, \ldots, F$, and

$$\bar{y}_j(\mathbf{z}) = p(y_j = 1 \mid \mathbf{z}) = \frac{e^{c_j + \mathbf{z}^T \mathbf{U}_{:,j}}}{\sum_{j'=1}^{C} e^{c_{j'} + \mathbf{z}^T \mathbf{U}_{:,j'}}}, \tag{49}$$

for $j = 1, 2, \ldots, C$.

Hence, we have

$$p(\mathbf{x} \mid \mathbf{z}) \quad = \quad \prod_{i=1}^{F} \bar{x}_i(\mathbf{z})^{\,x_i} \; [1 - \bar{x}_i(\mathbf{z})]^{\,1-x_i} \;, \tag{50}$$

and

$$p(\mathbf{y} \mid \mathbf{z}) \quad = \quad \frac{e^{\mathbf{c}^T\mathbf{y} + \mathbf{z}^T\mathbf{U}\mathbf{y}}}{\sum_{j'=1}^{C} e^{c_{j'} + \mathbf{z}^T\mathbf{U}_{:,j'}}} \,, \tag{51}$$

such that

$$p(\mathbf{x}, \mathbf{y}) \quad \approx \quad p(\mathbf{x} \mid \bar{\mathbf{z}})\, p(\mathbf{y} \mid \bar{\mathbf{z}}) \,. \tag{52}$$

Consequently, the required gradients are:

$$\frac{\partial f}{\partial \mathbf{a}} = \mathbf{x} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{a}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{x} - \bar{\mathbf{x}}' \,, \tag{53}$$

$$\frac{\partial f}{\partial \mathbf{b}} = \mathbf{z} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{b}} \ln p(\mathbf{x}, \mathbf{y}) \approx \bar{\mathbf{z}} - \bar{\mathbf{z}}' \,, \tag{54}$$

$$\frac{\partial f}{\partial \mathbf{c}} = \mathbf{y} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{c}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{y} - \bar{\mathbf{y}}' \,, \tag{55}$$

$$\frac{\partial f}{\partial \mathbf{W}} = \mathbf{x}\,\mathbf{z}^T \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{W}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{x}\,\bar{\mathbf{z}}^T - \bar{\mathbf{x}}'\,\bar{\mathbf{z}}'^T \,, \tag{56}$$

$$\frac{\partial f}{\partial \mathbf{U}} = \mathbf{z}\,\mathbf{y}^T \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{U}} \ln p(\mathbf{x}, \mathbf{y}) \approx \bar{\mathbf{z}}\,\mathbf{y}^T - \bar{\mathbf{z}}'\,\bar{\mathbf{y}}'^T \,. \tag{57}$$

1.0.8 Optimising the marginal likelihood

In the case of unsupervised training, we have no class label for input $\mathbf{x}$, and thus no explicit $\mathbf{y}$ vector. Consequently, rather than maximising the joint likelihood $p(\mathbf{x}, \mathbf{y})$, we instead maximise the marginal likelihood $p(\mathbf{x})$. Hence, from (Jarrad$_3$), we obtain

$$p(\mathbf{x}) \quad \approx \quad p(\mathbf{x} \mid \tilde{\mathbf{y}}(\mathbf{x}), \bar{\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{y}}(\mathbf{x}))) \,, \tag{58}$$

where we have now defined

$$\tilde{\mathbf{y}}(\mathbf{x}) \quad \doteq \quad \mathbb{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}] \,, \tag{59}$$

in order to distinguish it from $\bar{\mathbf{y}}(\mathbf{z})$. Using the conditional independence of $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$, we thus have

$$p(\mathbf{x}) \quad \approx \quad p(\mathbf{x} \mid \bar{\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{y}}(\mathbf{x}))) \,. \tag{60}$$

Also from (Jarrad$_3$), we obtain the gradient as

$$\nabla \ln p(\mathbf{x}) \quad \approx \quad \mathbb{E}_{\mathbf{y}|\mathbf{x}} \left[\mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y}} [\nabla f] \right] - \mathbb{E}_{\mathbf{y}|\mathbf{x}} \left[\mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y}} \left[\mathbb{E}_{\mathbf{x}'|\mathbf{y},\mathbf{z}} \left[\mathbb{E}_{\mathbf{y}'|\mathbf{x}'} \left[\mathbb{E}_{\mathbf{z}'|\mathbf{x}',\mathbf{y}'} [\nabla f] \right] \right] \right] \right] \,. \tag{61}$$

However, observe that $\mathbb{E}_{\mathbf{x}'|\mathbf{y},\mathbf{z}} = \mathbb{E}_{\mathbf{x}'|\mathbf{z}}$. Hence, the gradient is given by

$$\nabla \ln p(\mathbf{x}) \quad \approx \quad \nabla f(\mathbf{x}, \tilde{\mathbf{y}}, \bar{\mathbf{z}}) - \nabla f(\bar{\mathbf{x}}', \tilde{\mathbf{y}}', \bar{\mathbf{z}}') \,, \tag{62}$$

where now $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\mathbf{x})$, $\bar{\mathbf{z}} = \bar{\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{y}})$, $\bar{\mathbf{x}}' = \bar{\mathbf{x}}(\bar{\mathbf{z}})$, $\tilde{\mathbf{y}}' = \tilde{\mathbf{y}}(\bar{\mathbf{x}}')$, and $\bar{\mathbf{z}}' = \bar{\mathbf{z}}(\bar{\mathbf{x}}', \tilde{\mathbf{y}}')$.

The procedure for computing the unsupervised gradient is: 1. Push the known input $\mathbf{x}$ through the RBC to the output, and compute $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\mathbf{x})$. 1. Push the known input $\mathbf{x}$ and estimated output $\tilde{\mathbf{y}}$ back into the hidden layer of the RBC, and compute $\bar{\mathbf{z}} = \bar{\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{y}})$. 2. Push $\bar{\mathbf{z}}$ back out of the hidden layer into the input layer, and compute $\bar{\mathbf{x}}' = \bar{\mathbf{x}}(\bar{\mathbf{z}})$. 3. Push the reconstructed $\bar{\mathbf{x}}'$ through the RBC to the output, and compute $\tilde{\mathbf{y}}' = \tilde{\mathbf{y}}(\mathbf{x}')$. 3. Push the reconstructed input $\bar{\mathbf{x}}'$ and output $\tilde{\mathbf{y}}'$ back into the hidden layer, and compute $\bar{\mathbf{z}}' = \bar{\mathbf{z}}(\bar{\mathbf{x}}', \bar{\mathbf{y}}')$. 4. Compute the gradient as the difference of the data term $\nabla f(\mathbf{x}, \tilde{\mathbf{y}}, \bar{\mathbf{z}})$ and the reconstruction term $\nabla f(\bar{\mathbf{x}}', \tilde{\mathbf{y}}', \bar{\mathbf{z}}')$.

1.0.9 Unsupervised RBC training

From the previous section, the required gradients are:

$$\frac{\partial f}{\partial \mathbf{a}} = \mathbf{x} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{a}} \ln p(\mathbf{x}) \approx \mathbf{x} - \bar{\mathbf{x}}' \,, \tag{63}$$

$$\frac{\partial f}{\partial \mathbf{b}} = \mathbf{z} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{b}} \ln p(\mathbf{x}) \approx \bar{\mathbf{z}} - \bar{\mathbf{z}}' \,, \tag{64}$$

$$\frac{\partial f}{\partial \mathbf{c}} = \mathbf{y} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{c}} \ln p(\mathbf{x}) \approx \tilde{\mathbf{y}} - \tilde{\mathbf{y}}' \,, \tag{65}$$

$$\frac{\partial f}{\partial \mathbf{W}} = \mathbf{x}\,\mathbf{z}^T \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{W}} \ln p(\mathbf{x}) \approx \mathbf{x}\,\bar{\mathbf{z}}^T - \bar{\mathbf{x}}'\,\bar{\mathbf{z}}'^T \,, \tag{66}$$

$$\frac{\partial f}{\partial \mathbf{U}} = \mathbf{z}\,\mathbf{y}^T \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{U}} \ln p(\mathbf{x}) \approx \bar{\mathbf{z}}\,\tilde{\mathbf{y}}^T - \bar{\mathbf{z}}'\,\tilde{\mathbf{y}}'^T \,. \tag{67}$$

We also require

$$\tilde{y}_j(\mathbf{x}) \doteq p(y_j = 1 \mid \mathbf{x}) = \frac{e^{c_j} \prod_{k=1}^{H} \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + u_{kj}}\right)}{\sum_{j'=1}^{C} e^{c_{j'}} \prod_{k=1}^{H} \left(1 + e^{b_k + \mathbf{x}^T \mathbf{W}_{:,k} + u_{kj'}}\right)} \,, \tag{68}$$

such that

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{j=1}^{C} \tilde{y}_j(\mathbf{x})^{\,y_j} \,. \tag{69}$$

The log-likelihood score is given by

$$\ln p(\mathbf{x}) \approx \ln p(\mathbf{x} \mid \bar{\mathbf{z}}) \,. \tag{70}$$

1.0.10 Sequential RBC

(Jarrad$_2$) examined the case where the input $\mathbf{x}$ to an RBM implicitly took the form of a Markov sequence. Following similar reasoning here, we obtain

$$p(\mathbf{x}) = p(x_1)\, p(x_2 \mid x_1) \ldots p(x_F \mid x_1, \ldots, x_{F-1}) \tag{71}$$

$$= \prod_{i=1}^{F} p(x_i \mid \mathbf{x}_{1:i-1}) \,, \tag{72}$$

where now

$$
\begin{aligned}
p(x_i \mid \mathbf{x}_{1:i-1}) &= \sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} p(x_i, \mathbf{y}, \mathbf{z} \mid \mathbf{x}_{1:i-1}) && (73)\\
&= \sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} p(x_i \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{x}_{1:i-1}, \mathbf{y})\, p(\mathbf{y} \mid \mathbf{x}_{1:i-1}) && (74)\\
&= \mathbb{E}_{\mathbf{y}|\mathbf{x}_{1:i-1}}\left[\mathbb{E}_{\mathbf{z}|\mathbf{x}_{1:i-1},\mathbf{y}}\left[p(x_i \mid \mathbf{z})\right]\right] && (75)\\
&\approx p\left(x_i \mid \bar{\mathbf{z}}\left(\mathbf{x}_{1:i-1}, \tilde{\mathbf{y}}(\mathbf{x}_{1:i-1})\right)\right) , && (76)
\end{aligned}
$$

due to the conditional independence of $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$.

We require

$$
\bar{x}_i(\mathbf{z}) \doteq p(x_i = 1 \mid \mathbf{z}) = \sigma\left(a_i + \mathbf{W}_{i,:}\mathbf{z}\right) , \tag{77}
$$

for $i = 1, 2, \ldots, F$, and

$$
\tilde{y}_j(\mathbf{x}_{1:i-1}) \doteq p(y_j = 1 \mid \mathbf{x}_{1:i-1}) = \frac{e^{c_j}\prod_{k=1}^{H}\left(1 + e^{b_k + \mathbf{x}_{1:i-1}^T\mathbf{W}_{1:i-1,k} + u_{kj}}\right)}{\sum_{j'=1}^{C} e^{c_{j'}}\prod_{k=1}^{H}\left(1 + e^{b_k + \mathbf{x}_{1:i-1}^T\mathbf{W}_{1:i-1,k} + u_{kj'}}\right)} , \tag{78}
$$

for $j = 1, 2, \ldots, C$, and

$$
\bar{z}_k(\mathbf{x}_{1:i-1}, \mathbf{y}) \doteq p(z_k = 1 \mid \mathbf{x}_{1:i-1}, \mathbf{y}) = \sigma\left(b_k + \mathbf{x}_{1:i-1}^T\mathbf{W}_{1:i-1,k} + \mathbf{U}_{k,:}\mathbf{y}\right) , \tag{79}
$$

for $k = 1, 2, \ldots, H$.

TODO Finsh the derivatives of the log-likelihood.

1.0.11 Logistic RBM

For the purposes of comparison, we now turn to a version of the RBC without the hidden $\mathbf{z}$ layer. Hence, the joint probability is now

$$
p(\mathbf{x}, \mathbf{y}) \doteq \frac{e^{g(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{g(\mathbf{x}',\mathbf{y}')}} = \frac{e^{\mathbf{a}^T\mathbf{x} + \mathbf{c}^T\mathbf{y} + \mathbf{x}^T\mathbf{V}\mathbf{y}}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\mathbf{a}^T\mathbf{x}' + \mathbf{c}^T\mathbf{y}' + \mathbf{x}'^T\mathbf{V}\mathbf{y}'}} , \tag{80}
$$

where we have intoduced the new weight matrix V to account for the interactions between input $\mathbf{x}$ and output $\mathbf{y}$.

The joint log-likelihood is thus

$$
\begin{aligned}
\ln p(\mathbf{x}, \mathbf{y}) &= g(\mathbf{x}, \mathbf{y}) - \ln \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{g(\mathbf{x}',\mathbf{y}')} && (81)\\
\Rightarrow \nabla \ln p(\mathbf{x}, \mathbf{y}) &= \nabla g(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathbf{x}',\mathbf{y}'}\left[\nabla g(\mathbf{x}', \mathbf{y}')\right] && (82)\\
&\approx \nabla g(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}'|\mathbf{x}}\left[\mathbb{E}_{\mathbf{x}'|\mathbf{y}'}\left[\nabla g(\mathbf{x}', \mathbf{y}')\right]\right] && (83)\\
&\approx \nabla g(\mathbf{x}, \mathbf{y}) - \nabla g\left(\bar{\mathbf{x}}\left(\tilde{\mathbf{y}}(\mathbf{x})\right), \tilde{\mathbf{y}}(\mathbf{x})\right) , && (84)
\end{aligned}
$$

using CEA and MFA, from (Jarrad$_3$).

Consequently, the joint gradients are therefore:

$$\frac{\partial g}{\partial \mathbf{a}} = \mathbf{x} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{a}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{x} - \bar{\mathbf{x}}(\tilde{\mathbf{y}}(\mathbf{x}))\,, \tag{85}$$

$$\frac{\partial g}{\partial \mathbf{c}} = \mathbf{y} \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{c}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{y} - \tilde{\mathbf{y}}(\mathbf{x})\,, \tag{86}$$

$$\frac{\partial g}{\partial \mathbf{V}} = \mathbf{x}\,\mathbf{y}^T \quad \Rightarrow \quad \frac{\partial}{\partial \mathbf{V}} \ln p(\mathbf{x}, \mathbf{y}) \approx \mathbf{x}\,\mathbf{y}^T - \bar{\mathbf{x}}\left(\tilde{\mathbf{y}}(\mathbf{x})\right)\,\tilde{\mathbf{y}}(\mathbf{x})^T\,. \tag{87}$$

The predictive output of the model is

$$p(\mathbf{y} \mid \mathbf{x}) \;=\; \frac{e^{f(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{f(\mathbf{x},\mathbf{y}')}} = \frac{e^{\mathbf{c}^T\mathbf{y} + \mathbf{x}^T\mathbf{V}\mathbf{y}}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\mathbf{c}^T\mathbf{y}' + \mathbf{x}^T\mathbf{V}\mathbf{y}'}}\,. \tag{88}$$

For binary one-hot outputs, we thus have

$$\tilde{y}_j(\mathbf{x}) \;\doteq\; p(y_j = 1 \mid \mathbf{x}) = \frac{e^{c_j + \mathbf{x}^T\mathbf{V}_{:,j}}}{\sum_{j'=1}^{C} e^{c'_j + \mathbf{x}^T\mathbf{V}_{:,j'}}}\,, \tag{89}$$

for $j = 1, 2, \ldots, C$.

Conversely, the predictive input of the model is

$$p(\mathbf{x} \mid \mathbf{y}) \;=\; \frac{e^{f(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{x}' \in \mathcal{X}} e^{f(\mathbf{x}',\mathbf{y})}} = \frac{e^{\mathbf{a}^T\mathbf{x} + \mathbf{x}^T\mathbf{V}\mathbf{y}}}{\sum_{\mathbf{x}' \in \mathcal{X}} e^{\mathbf{a}^T\mathbf{x}' + \mathbf{x}'^T\mathbf{V}\mathbf{y}}}\,. \tag{90}$$

For binary inputs, we thus have

$$\bar{x}_i(\mathbf{y}) \;\doteq\; p(x_i = 1 \mid \mathbf{y}) = \sigma(a_i + \mathbf{V}_{i,:}\mathbf{y})\,, \tag{91}$$

for $i = 1, 2, \ldots, F$.

In order to score the gradient updates, first observe that

$$p(\mathbf{x}, \mathbf{y}) \;=\; p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\,, \tag{92}$$

where

$$p(\mathbf{y} \mid \mathbf{x}) \;=\; \prod_{j=1}^{C} \tilde{y}_j(\mathbf{x})^{y_j}\,, \tag{93}$$

for one-hot outputs. Next, note that

$$p(\mathbf{x}) \;=\; \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y}) \tag{94}$$

$$\;=\; \mathbb{E}_{\mathbf{y}}[p(\mathbf{x} \mid \mathbf{y})] \approx \mathbb{E}_{\mathbf{y}|\mathbf{x}}[p(\mathbf{x} \mid \mathbf{y})] \tag{95}$$

$$\;\approx\; p\left(\mathbf{x} \mid \tilde{\mathbf{y}}(\mathbf{x})\right)\,, \tag{96}$$

where

$$p(\mathbf{x} \mid \mathbf{y}) \;=\; \prod_{i=1}^{F} \bar{x}_i(\mathbf{y})^{\,x_i}\, \left[1 - \bar{x}_i(\mathbf{y})\right]^{1 - x_i}\,, \tag{97}$$

for binary inputs.

[ ]: