Lab Exercise 2: Restricted Boltzmann Machine

Caio Corro

1 Introduction

In this lab exercise, you will build a Restricted Boltzmann Machine (RBM) with continuous observed variables and binary latent variables. Hence, we have the following sample spaces:

- $\boldsymbol{x} \in \mathbb{R}^2$ for the observed variables,
- $\boldsymbol{y} \in \{0, 1\}^n$ for the latent variables.

The joint probability distribution over all random variables is defined as follows:

$$p(\boldsymbol{x}, \boldsymbol{y}; \theta) = \frac{\exp(w(\boldsymbol{x}, \boldsymbol{y};\ \theta))}{\int_{\boldsymbol{x}'} \sum_{\boldsymbol{y}'} \exp(w(\boldsymbol{x}', \boldsymbol{y}';\ \theta)) d\boldsymbol{x}'} = \frac{\exp(w(\boldsymbol{x}, \boldsymbol{y};\ \theta))}{Z(\theta)} = Z(\theta)^{-1} \exp(w(\boldsymbol{x}, \boldsymbol{y};\ \theta)) = \exp(w(\boldsymbol{x}, \boldsymbol{y};\ \theta) - c(\theta))$$

where $w : \mathbb{R}^2 \times \{0, 1\}^n \to \mathbb{R}$ is the energy function parameterized by θ, $Z(\theta)$ is the partition function and $c(\theta) = \log Z(\theta)$ is the log-partition (or cumulant) function. Note that the integral and the sum are over all possible values for the observed and latent variables, respectively. As mentioned in the course, in a RBM we have the following conditional independence by construction (i.e. by how we define the function w):

- $p(\boldsymbol{x}|\boldsymbol{y}; \theta) = p(x_1|\boldsymbol{y}; \theta) p(x_2|\boldsymbol{y}; \theta)$
- $p(\boldsymbol{y}|\boldsymbol{x}; \theta) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}; \theta)$

Moreover, both $p(\boldsymbol{x}|\boldsymbol{y}; \theta)$ and $p(\boldsymbol{y}|\boldsymbol{x}; \theta)$ are well known probability distributions. In the example of the course, they were both independent Bernoullis. In this lab exercise, $p(\boldsymbol{x}|\boldsymbol{y}; \theta)$ are independent Gaussians and $p(\boldsymbol{y}|\boldsymbol{x}; \theta)$ are independent Bernoullis (see below for a derivation). This fact allows us to rely on efficient Gibbs sampling in order to sample from the joint distribution $p(\boldsymbol{x}, \boldsymbol{y}; \theta)$: we start from a random assignation of either $\boldsymbol{y}$ or $\boldsymbol{x}$, and then iteratively sample from $p(X|\boldsymbol{y}; \theta)$ and $p(Y|\boldsymbol{x}; \theta)$. For example:

1. $\forall i : \boldsymbol{y}_i^{(1)} \sim \mathbb{B}(0.5)$ where $\mathbb{B}(0.5)$ is an unbiased Bernoulli distribution,
2. $\boldsymbol{x}^{(2)} \sim p(X|\ \boldsymbol{y}^{(1)}; \theta)$,
3. $\boldsymbol{y}^{(2)} \sim p(Y|\ \boldsymbol{x}^{(2)}; \theta)$,
4. $\boldsymbol{x}^{(3)} \sim p(X|\ \boldsymbol{y}^{(2)}; \theta)$,
5. $\boldsymbol{y}^{(3)} \sim p(Y|\ \boldsymbol{x}^{(3)}; \theta)$,
6. ...

And we keep one or several couple of values $\langle \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)} \rangle$ from this Markov chain as samples.

2 Weighting function

In this lab exercise, the RBM weighting function is defined as follows:

$$w(\boldsymbol{x}, \boldsymbol{y}; \theta) = -\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{2} \sum_{j=1}^{n} \frac{x_i}{\sigma_i^2} W_{ij} y_j + \sum_{j=1}^{n} d_j y_j$$

where $\theta = \{\boldsymbol{b}, \sigma, \boldsymbol{W}, \boldsymbol{d}\}$ is the set of parameters that must be learned from data:

- $\boldsymbol{b} \in \mathbb{R}^2$,
- $\sigma \in \mathbb{R}^2_{++}$ (that is $\sigma_1 > 0$ and $\sigma_2 > 0$),
- $\boldsymbol{W} \in \mathbb{R}^{2 \times n}$,
- $\boldsymbol{d} \in \mathbb{R}^n$.

We can compute analytically the conditional probability of x_1:

$$\begin{aligned}
p(x_1|x_2,\boldsymbol{y};\theta) &= \frac{p(x_1,x_2,\boldsymbol{y};\theta)}{p(x_2,\boldsymbol{y};\theta)} \\
&= \frac{p(x_1,x_2,\boldsymbol{y};\theta)}{\int_{x_1'} p(x_1',x_2,\boldsymbol{y};\theta)dx_1'} \\
&= \frac{Z(\theta)^{-1}\exp\left(-\frac{(x_1-b_1)^2}{2\sigma_i^2} - \frac{(x_2-b_2)^2}{2\sigma_i^2} + \sum_{j=1}^n \frac{x_1}{\sigma_1^2}W_{1j}y_j + \sum_{j=1}^n \frac{x_2}{\sigma_2^2}W_{2j}y_j + \sum_{j=1}^n d_j y_j\right)}{\int_{x_1'} Z(\theta)^{-1}\exp\left(-\frac{(x_1'-b_1)^2}{2\sigma_i^2} - \frac{(x_2-b_2)^2}{2\sigma_i^2} + \sum_{j=1}^n \frac{x_1'}{\sigma_1^2}W_{1j}y_j + \sum_{j=1}^n \frac{x_2}{\sigma_2^2}W_{2j}y_j + \sum_{j=1}^n d_j y_j\right)dx_1'}
\end{aligned}$$

Remember the exponential rule $\exp(a+b) = \exp(a)\exp(b)$. In the denominator, we can get out of the integral the partition function term $Z(\theta)^{-1}$ and all terms in the exponential that do not depend on x_1' thanks to the exponential rule. Then, these terms will cancel with their counterparts in the nominator, so we end up with:

$$= \frac{\exp\left(-\frac{(x_1-b_1)^2}{2\sigma_i^2} + \sum_{j=1}^n \frac{x_1}{\sigma_1^2}W_{1j}y_j\right)}{\int_{x_1'}\exp\left(-\frac{(x_1'-b_1)^2}{2\sigma_i^2} + \sum_{j=1}^n \frac{x_1'}{\sigma_1^2}W_{1j}y_j\right)dx_1'}$$

which shows the conditional independence of x_1 given x_2 as the latter value does not appear in the equation anymore. Let $a = \sum_{j=1}^n W_{1j}y_j$. Then we can rewritte the argument of the exponential in the numerator argument as:

$$\frac{ax_1}{\sigma_1^2} - \frac{(x_1-b_1)^2}{2\sigma_1^2} = \frac{2ax_1 - x_1^2 - b_1^2 + 2b_1x_1}{2\sigma_1^2} = \frac{2ax_1 - x_1^2 - b_1^2 + 2b_1x_1 + a^2 - a^2 + 2ab_1 - 2ab_1}{2\sigma_1^2} = \frac{-(x_1-a-b_1)^2 + a^2 + 2ab_1}{2\sigma_1^2}$$

By using this relation and similarly in the denominator, we can rewritte the conditional probability as:

$$= \frac{\exp\left(\frac{-(x_1-\sum_{j=1}^n y_jW_{1j}-b_1)^2 + (\sum_{j=1}^n W_{1j}y_j)^2 + 2\sum_{j=1}^n W_{1j}y_jb_1}{2\sigma_1^2}\right)}{\int_{x_1'}\exp\left(\frac{-(x_1'-\sum_{j=1}^n W_{1j}y_j-b_1)^2 + (\sum_{j=1}^n W_{1j}y_j)^2 + 2\sum_{j=1}^n W_{1j}y_jb_1}{2\sigma_1^2}\right)dx_1'}$$

Again, we can simplify by removing terms that do not depend on x_1' or x_1 by using the exponential rule and taking them out of the integral in the denominator:

$$\begin{aligned}
&= \frac{\exp\left(\frac{-(x_1-\sum_{j=1}^n W_{1j}y_j-b_1)^2}{2\sigma_1^2}\right)}{\int_{x_1'}\exp\left(\frac{-(x_1'-\sum_{j=1}^n W_{1j}y_j-b_1)^2}{2\sigma_1^2}\right)dx_1'} \\
&= \frac{\exp\left(-\frac{1}{2}\left(\frac{x_1-\sum_{j=1}^n W_{1j}y_j-b_1}{\sigma_1}\right)^2\right)}{\int_{x_1'}\exp\left(-\frac{1}{2}\left(\frac{x_1'-\sum_{j=1}^n W_{1j}y_j-b_1}{\sigma_1}\right)^2\right)dx_1'}
\end{aligned}$$

Now, this value looks like a PDF where the denominator act as a normalizer that ensures that the integral of the PDF is equal to 1. You can observe that the numeror is an unnormalized Gaussian, therfore the normalization term is equal to $\sigma_1\sqrt{2\pi}$. This results can also be shown using closed form formula for Gaussian function integration.[1]By rewritting the denominator as $\sigma_1\sqrt{2\pi}$ we have:

$$\begin{aligned}
&= \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_1-(\sum_{j=1}^n W_{1j}y_j+b_1)}{\sigma_1}\right)^2\right) \\
&= \mathcal{N}\left(x_1\middle|b_1+\sum_{j=1}^n W_{1j}y_j, \sigma_1^2\right)
\end{aligned}$$

that is $p(x_1|x_2,\boldsymbol{y};\theta) = p(x_1|\boldsymbol{y};\theta)$ is a Gaussian with mean $b_1 + \sum_{j=1}^n W_{1j}y_j$ and standard deviation σ_1^2. We can do a similar computation for the other observed variable:

$$p(x_2|x_1,\boldsymbol{y};\theta) = p(x_2|\boldsymbol{y};\theta) = \mathcal{N}\left(x_2\middle|b_2+\sum_{j=1}^n W_{2j}y_j, \sigma_2^2\right)$$

[1]See `https://en.wikipedia.org/wiki/Gaussian_integral`

Similarly, we can compute the conditional probability of an observation. If we take the specific case $p(y_1|x, y_2...y_n; \theta)$:

$$\begin{aligned} p(y_1|x, y_2...y_n; \theta) &= \frac{\exp(w(x, [y_1, y_2...y_n]; \theta))}{\sum_{y_1'} \exp(w(x, [y_1', y_2...y_n]; \theta))} \\ &= \frac{\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{2}\sum_{j=1}^{n} \frac{x_i}{\sigma_i^2} W_{ij} y_j + \sum_{j=1}^{n} d_j y_j\right)}{\sum_{y_1'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1} y_1' + d_j y_1' + \sum_{i=1}^{2}\sum_{j=2}^{n} \frac{x_i}{\sigma_i^2} W_{ij} y_j + \sum_{j=2}^{n} d_j y_j\right)} \end{aligned}$$

As previously, there are many terms that cancels:

$$= \frac{\exp\left(\sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1} y_1 + d_1 y_1\right)}{\sum_{y_1'} \exp\left(\sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1} y_1' + d_1 y_1'\right)}$$

The variable y_1' can either be 0 or 1 in the denominator's sum. In the case it is 0, the exponential simplifies as $\exp(0) = 1$. Moreover, we can factorize by y_1 in the numerator's exponential, therefore:

$$= \frac{\exp\left(y_1 \left(\sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1} + d_1\right)\right)}{1 + \exp\left(\sum_{i=1}^{2} \frac{x_1}{\sigma_i^2} W_{i1} + d_1\right)}$$

Hence, we can see that y_1 is conditionally independent of other latent variables $p(y_1|x, y_2...y_n; \theta) = p(y_1|x; \theta)$. Moreover, from the formula we deduce that $p(y_1|x; \theta)$ is a Bernoulli distribution with parameter $\mu = \text{sigmoid}(\sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1} + d_1)$ which we write as:

$$= \mathcal{B}\left(y_1 \middle| \text{sigmoid}(d_1 + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{i1})\right)$$

A similar computation can be done for other latent variables. Note that for both condition distributions over observed and latent variables we can use tensor operations instead of computing parameters for each variable independently. Try to visualize how does it works!

3 Training

The parameter θ must be set so that the probability of the training data is maximized. To do this, we reformulate the probability of observation using the log-partition function of $p(\boldsymbol{y}|\boldsymbol{x})$, denoted $c(\boldsymbol{x}; \theta)$, which can be computed in closed form:

$$\begin{aligned} p(\boldsymbol{x}; \theta) &= \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}; \theta) \\ &= \sum_{\boldsymbol{y}} \frac{\exp(w(\boldsymbol{x}, \boldsymbol{y}; \theta))}{\int_{\boldsymbol{x}'} \sum_{\boldsymbol{y}'} \exp(w(\boldsymbol{x}', \boldsymbol{y}'; \theta)) d\boldsymbol{x}'} \\ &= \frac{\sum_{\boldsymbol{y}} \exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{2}\sum_{j=1}^{n} \frac{x_i}{\sigma_i^2} W_{ij} y_j + \sum_{j=1}^{n} d_j y_j\right)}{\int_{\boldsymbol{x}'} \sum_{\boldsymbol{y}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^{2}\sum_{j=1}^{n} \frac{x_i'}{\sigma_i^2} W_{ij} y_j' + \sum_{j=1}^{n} d_j y_j'\right) d\boldsymbol{x}'} \end{aligned}$$

We can use the fact that $\exp(a + b) = \exp(a)\exp(b)$ to rewritte both exponentials:

$$= \frac{\sum_{\boldsymbol{y}} \left[\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \exp\left(d_j y_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij} y_j\right)\right]}{\int_{\boldsymbol{x}'} \left[\sum_{\boldsymbol{y}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \left(d_j y_j' + \sum_{i=1}^{2} \frac{x_i'}{\sigma_i^2} W_{ij} y_j'\right)\right] d\boldsymbol{x}'}$$

Note that the first exponential term in the nominator does not depend on $\boldsymbol{y}$, so we can take it out of the sum, and similarly in the denominator:

$$= \frac{\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \sum_{\boldsymbol{y}} \left[\prod_{j=1}^{n} \exp\left(d_j y_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij} y_j\right)\right]}{\int_{\boldsymbol{x}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2}\right) \sum_{\boldsymbol{y}'} \left[\prod_{j=1}^{n} \left(d_j y_j' + \sum_{i=1}^{2} \frac{x_i'}{\sigma_i^2} W_{ij} y_j'\right)\right] d\boldsymbol{x}'}$$

Now we can rewritte using the distributivity of multiplication. The sum will simplify as a sum over a single latent variable, see the course for a more detailled explanation:

$$= \frac{\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \left[\sum_{y_j} \exp\left(d_j y_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij} y_j\right)\right]}{\int_{\boldsymbol{x}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \left[\sum_{y_j'} \left(d_j y_j' + \sum_{i=1}^{2} \frac{x_i'}{\sigma_i^2} W_{ij} y_j'\right)\right] d\boldsymbol{x}'}$$

The latent variable y_j can only take two values (0 and 1) in the sum in the numerator. When $y_j = 0$, we have $\exp(0) = 1$. By applying the same trick in the nominator and denominator, we have:

$$= \frac{\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \left[1 + \exp\left(d_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]}{\int_{\boldsymbol{x}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n} \left[1 + \exp\left(d_j + \sum_{i=1}^{2} \frac{x_i'}{\sigma_i^2} W_{ij}\right)\right] d\boldsymbol{x}'}$$

Now observe that if we can take the exponential-log of the numerator and then use the fact that $\log(ab) = \log a + \log b$ to change the product by a sum (moreover the log-exp in the first term then cancel - again, check the course for an explanation of this trick) we have:

$$\begin{aligned}
&= \frac{\exp\left(-\sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n} \log\left[1 + \exp\left(d_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]\right)}{\int_{\boldsymbol{x}'} \exp\left(-\sum_{i=1}^{2} \frac{(x_i' - b_i)^2}{2\sigma_i^2}\right) + \sum_{j=1}^{n} \log\left[1 + \exp\left(d_j + \sum_{i=1}^{2} \frac{x_i'}{\sigma_i^2} W_{ij}\right)\right] d\boldsymbol{x}'} \\
&= \frac{\exp\left(c(\boldsymbol{x}; \theta)\right)}{\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}'; \theta)\right) d\boldsymbol{x}'}
\end{aligned}$$

where $c(\boldsymbol{x}; \theta) = \sum_{i=1}^{2} \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n} \log\left[1 + \exp\left(d_j + \sum_{i=1}^{2} \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]$ is the log-partition function of $p(\boldsymbol{y}|\boldsymbol{x})$. Note that $c(\boldsymbol{x}; \theta)$ can be computed efficiently because we transformed the original sum over an exponential number of values $\sum_{\boldsymbol{y}}$ (sum over 2^n possible assignation for $\boldsymbol{y}$!) as n sums. The marginalization in the denominator is still intractable, but during training we will approximate its contribution to the gradient via Monte Carlo estimation.

Let $D = \{\boldsymbol{x}^{(1)} ... \boldsymbol{x}^{(d)}\}$ be the training set. During training, we seek to maximize the log probability of the dataset:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \sum_{i=1}^{d} \log p(\boldsymbol{x}^{(d)}; \theta) \\
&= \arg\min_{\theta} \sum_{i=1}^{d} \mathcal{L}(\theta, \boldsymbol{x}^{(d)})
\end{aligned}$$

where $\mathcal{L}(\theta, \boldsymbol{x}) = -\log p(\boldsymbol{x}; \theta)$ is the loss function. The gradient of the loss function is defined as:

$$\begin{aligned}
\nabla \mathcal{L}(\theta, \boldsymbol{x}) &= \nabla - \log p(\boldsymbol{x}; \theta) \\
&= \nabla - \log \frac{\exp\left(c(\boldsymbol{x};\ \theta)\right)}{\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}'} \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \nabla \log \left(\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}' \right) \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \frac{1}{\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}'} \nabla \left(\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}' \right)
\end{aligned}$$

Similarly to sums, the gradient of an integral is the integral of the gradient:

$$\begin{aligned}
&= -\nabla c(\boldsymbol{x};\ \theta) + \frac{1}{\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}'} \int_{\boldsymbol{x}'} \nabla \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}' \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \frac{1}{\int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) d\boldsymbol{x}'} \int_{\boldsymbol{x}'} \exp\left(c(\boldsymbol{x}';\ \theta)\right) \nabla c(\boldsymbol{x}';\ \theta) d\boldsymbol{x}' \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \int_{\boldsymbol{x}'} \frac{\exp\left(c(\boldsymbol{x}';\ \theta)\right)}{\int_{\boldsymbol{x}''} \exp\left(c(\boldsymbol{x}'';\ \theta)\right) d\boldsymbol{x}''} \nabla c(\boldsymbol{x}';\ \theta) d\boldsymbol{x}' \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \int_{\boldsymbol{x}'} p(\boldsymbol{x}'; \theta) \nabla c(\boldsymbol{x}';\ \theta) d\boldsymbol{x}' \\
&= -\nabla c(\boldsymbol{x};\ \theta) + \mathbb{E}_{x' \sim p(\boldsymbol{x}'; \theta)} \left[\nabla c(\boldsymbol{x}';\ \theta)\right]
\end{aligned}$$

Computing the expectation exactly is intractable but we can rely on Monte Carlo estimation:

$$\simeq -\nabla c(\boldsymbol{x};\ \theta) + \frac{1}{K} \sum_{k=1}^{K} \nabla c(\tilde{\boldsymbol{x}}^{(k)};\ \theta)$$

where $\tilde{\boldsymbol{x}}^{(k)} \sim p(\tilde{\boldsymbol{x}}^{(k)}; \theta)$ are K samples from the model distribution. To sample from the distribution, we can rely on Gibbs sampling Markov Chain Monte Carlo as it is easy to sample from both conditional distributions. In practice, we will rely on the contrastive divergence objective, i.e. we take a single sample using a single step of a Markov chain initialized with the datapoint $\boldsymbol{x}$.