COMP64101 - Lecture 4 - Exercise sheet with answers

Exercise 1 [Restricted Boltzmann Machine] The restricted Boltzmann machine (RBM for short) is an undirected graphical model for vectors of binary variables $\mathbf{v} = (v_1, \ldots, v_n)^\top$ and $\mathbf{h} = (h_1, \ldots, h_m)^\top$ with a probability mass function:

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)$$

where the matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ and vectors $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are parameters of this distribution. The variables v_j and h_i both take values in $\{0, 1\}$. The v_j are called the "visible" variables since they are assumed to be observed, while the h_i are the hidden variables since it is assumed that we cannot measure them.

(a) Use graph separation to show that the joint conditional $p(\mathbf{h}|\mathbf{v})$ factorises as

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}).$$

(b) Show that for each i we have

$$p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-b_i - \sum_j W_{ji} v_j\right)}$$

where $\mathbf{W} = (W_{ji})$, meaning W_{ji} is the element of $\mathbf{W}$ located at the intersection of row j and column i, so that $\sum_j W_{ji} v_j$ is the inner product (scalar product) between the i-th column of $\mathbf{W}$ and $\mathbf{v}$.

Answers:

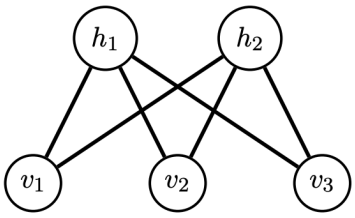

Figure 1: Graph for $p(\mathbf{v}, \mathbf{h})$.

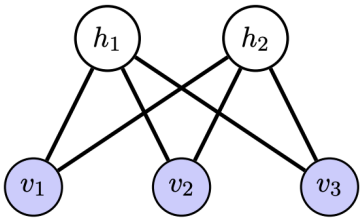

Figure 2: Graph for $p(\mathbf{h}|\mathbf{v})$.

(a) Figure 1 shows the undirected graph for $p(\mathbf{v}, \mathbf{h})$ with $n = 3$, $m = 2$. We note that the graph is bi-partite: there are only direct connections between the h_i and the v_j. Conditioning on $\mathbf{v}$ thus blocks all trails between the h_i (graph on Figure 2). This means that the h_i are independent from each other given $\mathbf{v}$ so that

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}).$$

(b) For the conditional pmf $p(h_i|\mathbf{v})$ any quantity that does not depend on h_i can be considered to be part of the normalisation constant. A general strategy is to first work out $p(h_i|\mathbf{v})$ up to the normalisation constant and then to normalise it afterwards.

We begin with $p(\mathbf{h}|\mathbf{v})$:

$$\begin{aligned} p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\ &\propto p(\mathbf{h},\mathbf{v}) \\ &\propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right) \\ &\quad \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{b}^\top \mathbf{h}\right) \\ &\quad \exp\left(\sum_i \sum_j v_j W_{ji} h_i + \sum_i b_i h_i\right) \end{aligned}$$

As we are interested in $p(h_i|\mathbf{v})$ for a fixed i, we can drop all the terms not depending on that h_i, so that

$$p(h_i|\mathbf{v}) \propto \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right)$$

Since h_i only takes two values, 0 and 1, normalisation is here straightforward. Call the unnormalised pmf $\tilde{p}(h_i|\mathbf{v})$, so we have

$$\tilde{p}(h_i|\mathbf{v}) = \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right)$$

We then have

$$\begin{aligned} p(h_i|\mathbf{v}) &= \frac{\tilde{p}(h_i|\mathbf{v})}{\tilde{p}(h_i=0|\mathbf{v}) + \tilde{p}(h_i=1|\mathbf{v})} \\ &= \frac{\tilde{p}(h_i|\mathbf{v})}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} = \frac{\exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \end{aligned}$$

Therefore

$$\begin{aligned} p(h_i=1|\mathbf{v}) &= \frac{\exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \\ &= \frac{1}{\exp\left(-\sum_j v_j W_{ji} - b_i\right) + 1} = \frac{1}{1 + \exp\left(-b_i - \sum_j v_j W_{ji}\right)} \end{aligned}$$

Remark 1: The probability $p(h_i=0|\mathbf{v})$ equals $1 - p(h_i=1|\mathbf{v})$, so then

$$p(h_i=0|\mathbf{v}) = 1 - p(h_i=1|\mathbf{v}) = \cdots (\text{algebra}) \cdots = \frac{1}{1 + \exp\left(b_i + \sum_j v_j W_{ji}\right)}$$

Remark 2: The function $x \mapsto 1/(1+\exp(-x))$ is called the logistic function. It is a sigmoid function and is thus sometimes denoted by $\sigma(x)$, so $\sigma(x) = 1/(1+\exp(-x))$. For other versions of the sigmoid function, see https://en.wikipedia.org/wiki/Sigmoid_function

With the notation $\sigma(\cdot)$ for the sigmoid, we have

$$p(h_i=1|\mathbf{v}) = \sigma\left(b_i + \sum_j v_j W_{ji}\right)$$

Exercise 2 [Restricted Boltzmann Machine, continued] For this exercise, we continue considering the RBM from Exercise 1. Use a symmetry argument to show that

(a) The joint conditional $p(\mathbf{v}|\mathbf{h})$ factorises as

$$p(\mathbf{v}|\mathbf{h}) = \prod_{j=1}^{n} p(v_j|\mathbf{h}).$$

(b) For each j we have

$$p(v_j = 1|\mathbf{h}) = \frac{1}{1 + \exp\left(-a_j - \sum_i W_{ji} h_i\right)}.$$

Answers:

(a) Similar to the argument used in part (a) of Exercise 1, but the roles of v and h swapped.

(b) Since $\mathbf{v}^\top \mathbf{W}\mathbf{h}$ is a scalar, we have $(\mathbf{v}^\top \mathbf{W}\mathbf{h})^\top = \mathbf{v}^\top \mathbf{W}\mathbf{h}$, so that $\mathbf{h}^\top \mathbf{W}^\top \mathbf{v} = \mathbf{v}^\top \mathbf{W}\mathbf{h}$. Therefore:

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right) = \exp\left(\mathbf{h}^\top \mathbf{W}^\top \mathbf{v} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right).$$

To derive the result, we note that $\mathbf{v}$ and $\mathbf{a}$ now take the place of $\mathbf{h}$ and $\mathbf{b}$ from before, and that we now have $\mathbf{W}^\top$ rather than $\mathbf{W}$. In the equation (Exercise 1 part (b))

$$p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-b_i - \sum_j W_{ji} v_j\right)}$$

we thus replace

$$h_i \text{ with } v_j, \qquad b_i \text{ with } a_j, \qquad W_{ji} \text{ with } W_{ij}$$

to obtain $p(v_j = 1|\mathbf{h})$. In terms of the sigmoid function, we have

$$p(v_j = 1|\mathbf{h}) = \sigma\left(a_j + \sum_i W_{ji} h_i\right).$$

Exercise 3 [Sampling from a restricted Boltzmann machine] For this exercise, we continue considering the RBM from Exercise 1. Explain how to use Gibbs sampling to generate samples from the marginal $p(\mathbf{v})$ for any given values of $\mathbf{W}$, $\mathbf{a}$, and $\mathbf{b}$. Notice that the marginal has the form:

$$p(\mathbf{v}) = \frac{\sum_{\mathbf{h}} \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)}{\sum_{\mathbf{h},\mathbf{v}} \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)}$$

Hint: Feel free to use the identities established in the previous two exercises.

Answer: In order to generate samples $\mathbf{v}^{(k)}$ from $p(\mathbf{v})$ we generate samples $(\mathbf{v}^{(k)}, \mathbf{h}^{(k)})$ from $p(\mathbf{v}, \mathbf{h})$ and then ignore the $\mathbf{h}^{(k)}$.

Gibbs sampling is a MCMC method to produce a sequence of samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ that follow a target pdf/pmf $p(\mathbf{x})$ in the limit of infinite samples. In practice, we are happy if the chain is run long enough. Assuming that $\mathbf{x}$ is d-dimensional, we generate the next sample $\mathbf{x}^{(k+1)}$ in the sequence from the previous sample $\mathbf{x}^{(k)}$ by:

1. picking (randomly) an index $i \in \{1, \ldots, d\}$;
2. sampling $x_i^{(k+1)}$ from $p(x_i|\mathbf{x}_{\neg i}^{(k)})$ where $\mathbf{x}_{\neg i}^{(k)}$ is the $((d-1)$-dimensional) vector obtained by removing the ith coordinate from the (d-dimensional) vector $\mathbf{x}^{(k)}$, i.e. $\mathbf{x}_{\neg i}^{(k)} = (x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k)}, \ldots, x_d^{(k)})$;
3. setting $\mathbf{x}^{(k+1)} = (x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_d^{(k)})$.

For the RBM, the tuple $(\mathbf{h}, \mathbf{v})$ corresponds to $\mathbf{x}$ so that x_i in the above steps can either be a hidden variable or a visible. Hence

$$p(x_i|\mathbf{x}_{\neg i}) = \begin{cases} p(h_i|\mathbf{h}_{\neg i}, \mathbf{v}) & \text{if } x_i \text{ is a hidden variable,} \\ p(v_i|\mathbf{h}, \mathbf{v}_{\neg i}) & \text{if } x_i \text{ is a visible variable.} \end{cases}$$

To compute the conditionals on the right hand side, we use the hint (take for granted all the four identities established in the previous two exercises).

Given the independencies between the hidden given the visible and vice versa, we have

$$p(h_i|\mathbf{h}_{\neg i}, \mathbf{v}) = p(h_i|\mathbf{v})$$
$$p(v_i|\mathbf{h}, \mathbf{v}_{\neg i}) = p(v_i|\mathbf{h})$$

so that the expressions for $p(h_i = 1|\mathbf{v})$ and $p(v_i = 1|\mathbf{h})$ allow us to implement the Gibbs sampler.

Given the independencies, it makes further sense to sample the $\mathbf{h}$ and $\mathbf{v}$ variables in blocks: first we sample all the h_i given $\mathbf{v}$, and then all the v_i given the $\mathbf{h}$ (or vice versa). This is also known as block Gibbs sampling.

In summary, given a sample $(\mathbf{h}^{(k)}, \mathbf{v}^{(k)})$, we thus generate the next sample $(\mathbf{h}^{(k+1)}, \mathbf{v}^{(k+1)})$ in the sequence as follows:

- For all h_i for $i = 1, \ldots, m$:
  - compute $p_i^h = p(h_i = 1|\mathbf{v}^{(k)})$
  - sample u_i from a uniform distribution on $[0, 1]$ and set $h_i^{(k+1)}$ to 1 if $u_i \leq p_i^h$.
- For all v_j for $j = 1, \ldots, n$:
  - compute $p_j^v = p(v_j = 1|\mathbf{h}^{(k+1)})$
  - sample u_j from a uniform distribution on $[0, 1]$ and set $v_j^{(k+1)}$ to 1 if $u_j \leq p_j^v$.

As final step, after sampling S pairs $(\mathbf{h}^{(k)}, \mathbf{v}^{(k)})$ for $k = 1, \ldots, S$ the set of visibles $\mathbf{v}^{(k)}$ form samples from the marginal $p(\mathbf{v})$.