1 REPLICATED SOFTMAX MODEL

1.1 Replicated Softmax Model

bib:salakhutdinov-hinton2009 proposed the Replicated Softmax to model and extract low-dimensional latent semantic representations of an unstructured collection of documents.

Within Natural Language Processing research field it is possible to access the topic a document based on its words probability distribution, this is known as Topic Modelling.

Replicated Softmax model uses a Restricted Boltzmann Machine structure.

Considering a dictionary that has K words, a document with D words, and $\mathbf{h} \in \{0,1\}^F$ stochastic hidden units. Let $\mathbf{V}$ be an $D \times K$ observed binary matrix whose entries $v_{ik} = 1$ means the i^{th} word of the document is the k^{th} word of the dictionary, and $v_{ik} = 0$ otherwise. For a state $\{\mathbf{V}, \mathbf{h}\}$, the energy is defined as follows:

$$E(\mathbf{V}, \mathbf{h}; \boldsymbol{\theta}) = -\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk} h_j v_{ik} - \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik} b_{ik} - \sum_{j=1}^{F} h_j a_j, \tag{1}$$

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}, \mathbf{a}\}$ is the set of model parameters. $\mathbf{W}$ is the weight tensor, each element W_{ijk} is a symmetric interaction between a visible unit i that takes on value k and a hidden unit j, b_{ik} is the bias of unit i that takes on value k, and a_j is the bias of hidden unit j.

The normalized joint probability of state $\{\mathbf{V}, \mathbf{h}\}$ is given by:

$$P(\mathbf{V}, \mathbf{h}; \boldsymbol{\theta}) = \frac{e^{-E(\mathbf{V}, \mathbf{h}; \boldsymbol{\theta})}}{Z(\boldsymbol{\theta})}, \tag{2}$$

where Z is the partition function, the normalization constant to keep the probability within interval $[0, 1]$:

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{V}'}\sum_{\mathbf{h}'} e^{-E(\mathbf{V}', \mathbf{h}'; \boldsymbol{\theta})}. \tag{3}$$

The marginal probabilities the model given by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}, \mathbf{a}\}$ assings to a visible binary matrix $\mathbf{V}$ and to a hidden binary vector $\mathbf{h}$ are, respectively:

$$P(\mathbf{V}; \boldsymbol{\theta}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{V}, \mathbf{h}; \boldsymbol{\theta})}, \tag{4}$$

$$P(\mathbf{h}; \boldsymbol{\theta}) = \frac{1}{Z}\sum_{\mathbf{V}} e^{-E(\mathbf{V}, \mathbf{h}; \boldsymbol{\theta})}. \tag{5}$$

From the joint and marginal probability, we are able to derive the conditional

probabilities for the hidden units $P(\mathbf{h}|\mathbf{V};\boldsymbol{\theta})$ given the visible matrix $\mathbf{V}$:

$$P(\mathbf{h}|\mathbf{V};\boldsymbol{\theta}) = \frac{P(\mathbf{V},\mathbf{h};\boldsymbol{\theta})}{P(\mathbf{V};\boldsymbol{\theta})}. \tag{6}$$

For the sake of simplicity we will omit the term $\boldsymbol{\theta}$ by considering we are deriving the equations for a fixed model given by parameters $\boldsymbol{\theta}$, thus equation (6) becomes:

$$\begin{aligned}
P(\mathbf{h}|\mathbf{V}) &= \frac{\frac{1}{Z}e^{-E(\mathbf{V},\mathbf{h})}}{\frac{1}{Z}\sum_{\mathbf{h}'} e^{-E(\mathbf{V},\mathbf{h}')}} \qquad (7)\\
&= \frac{\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik} + \sum_{j=1}^{F} h_j a_j\right)}{\sum_{\mathbf{h}'}\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h'_j v_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik} + \sum_{j=1}^{F} h'_j a_j\right)}\\
&= \frac{\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v_{ik} + \sum_{j=1}^{F} h_j a_j\right)\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik}\right)}{\left[\sum_{\mathbf{h}'}\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h'_j v_{ik} + \sum_{j=1}^{F} h'_j a_j\right)\right]\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik}\right)}\\
&= \frac{\exp\left[\sum_{j=1}^{F} h_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}{\sum_{\mathbf{h}'}\exp\left[\sum_{j=1}^{F} h'_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}\\
&= \frac{\prod_{j=1}^{F}\exp\left[h_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}{\sum_{h'_1\in\{0,1\}}\cdots\sum_{h'_F\in\{0,1\}}\left\{\prod_{j=1}^{F}\exp\left[h'_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]\right\}}\\
&= \frac{\prod_{j=1}^{F}\exp\left[h_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}{\left\{\sum_{\substack{h'_1\\ \in\{0,1\}}}\exp\left[h'_1\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{i1k}v_{ik} + a_1\right)\right]\right\}\cdots\left\{\sum_{\substack{h'_F\\ \in\{0,1\}}}\exp\left[h'_F\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{iFk}v_{ik} + a_F\right)\right]\right\}}\\
&= \frac{\prod_{j=1}^{F}\exp\left[h_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}{\left[1+\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{i1k}v_{ik} + a_1\right)\right]\cdots\left[1+\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{iFk}v_{ik} + a_F\right)\right]}\\
&= \prod_{j=1}^{F}\frac{\exp\left[h_j\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}{\left[1+\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]}
\end{aligned}$$

$$P(\mathbf{h}|\mathbf{V}) = \prod_{j=1}^{F} P(h_j|\mathbf{V}), \tag{8}$$

where, considering the probability of $h_j = 1$ given $\mathbf{V}$, than equation (8) for a hidden unit j is written as:

$$P(h_j = 1|\mathbf{V}) = \frac{\exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)}{\left[1 + \exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right)\right]} = \sigma\left(\sum_{i=1}^{D}\sum_{k=1}^{K} W_{ijk}v_{ik} + a_j\right), \tag{9}$$

which is also known as the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$. Thus, as in an common restricted Boltzmann Machine, each of the hidden units have an activation probability given by the logistic function.

On the other hand, the conditional probability of the visible units given the state of the hidden units $P(\mathbf{V}|\mathbf{h};\boldsymbol{\theta})$ is given by:

$$P(\mathbf{V}|\mathbf{h};\boldsymbol{\theta}) = \frac{P(\mathbf{V},\mathbf{h};\boldsymbol{\theta})}{P(\mathbf{h};\boldsymbol{\theta})}. \tag{10}$$

Again we consider a fixed model $\boldsymbol{\theta}$, thus this term is omitted from now on. The following derivation of equation (10) is not quite straight forward, and we will try to keep it as clear as possible:

$$\begin{aligned}
P(\mathbf{V}|\mathbf{h}) &= \frac{\frac{1}{Z}e^{-E(\mathbf{V},\mathbf{h})}}{\frac{1}{Z}\sum_{\mathbf{V}'} e^{-E(\mathbf{V}',\mathbf{h})}} \qquad (11)\\
&= \frac{\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik} + \sum_{j=1}^{F} h_j a_j\right)}{\sum_{\mathbf{V}'}\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v'_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v'_{ik}b_{ik} + \sum_{j=1}^{F} h_j a_j\right)}\\
&= \frac{\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}b_{ik}\right)\exp\left(\sum_{j=1}^{F} h_j a_j\right)}{\left[\sum_{\mathbf{V}'}\exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ijk}h_j v'_{ik} + \sum_{i=1}^{D}\sum_{k=1}^{K} v'_{ik}b_{ik}\right)\right]\exp\left(\sum_{j=1}^{F} h_j a_j\right)}\\
&= \frac{\exp\left[\sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik}\left(\sum_{j=1}^{F} W_{ijk}h_j + b_{ik}\right)\right]}{\sum_{\mathbf{V}'}\exp\left[\sum_{i=1}^{D}\sum_{k=1}^{K} v'_{ik}\left(\sum_{j=1}^{F} W_{ijk}h_j + b_{ik}\right)\right]}
\end{aligned}$$

$$= \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\sum_{\mathbf{V}'}\left\{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v'_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]\right\}}$$

$$= \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\sum_{\mathbf{V}'_1}\sum_{\mathbf{V}'_2}\cdots\sum_{\mathbf{V}'_D}\left\{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v'_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]\right\}}$$

$$= \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\left\{\sum_{\mathbf{V}'_1}\prod_{k=1}^{K}\exp\left[v'_{1k}\left(\sum_{j=1}^{F}W_{1jk}h_j + b_{1k}\right)\right]\right\}\cdots\left\{\sum_{\mathbf{V}'_D}\prod_{k=1}^{K}\exp\left[v'_{Dk}\left(\sum_{j=1}^{F}W_{Djk}h_j + b_{Dk}\right)\right]\right\}}$$

$$P(\mathbf{V}|\mathbf{h}) = \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\prod_{i=1}^{D}\left\{\sum_{\mathbf{V}'_i}\prod_{k=1}^{K}\exp\left[v'_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]\right\}} \tag{12}$$

To contunue with derivation of equation (12), we will take a closer look at the denominator. For a document $\mathbf{V}$, in which there are D words positions that are chosen from a dictionary of K available words, when $v_{ik} = 1$, it means that the i^{th} word of the document is the k^{th} word of the dictionary, and 0 otherwise. Whenever we sum over all K words of the dictionary, this sum is 1 for each word i of the document, $\sum_{k=1}^{K} v_{ik} = 1$. In this case we exclude empty documents, and even for a single word document, the summation above will be 1. In order to help with readability, we consider the simplification presented on (JORG THESIS, 2011), we take $r_{ik} = \sum_{j=1}^{F} W_{ijk}h_j + b_{ik}$, then, for a particular word i of the document, the denominator of equation (12) becomes:

$$\sum_{\mathbf{V}'_i}\prod_{k=1}^{K}\exp\left[v'_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right] = \sum_{\mathbf{V}'_i}\prod_{k=1}^{K}\exp\left(v'_{ik}r_{ik}\right)$$

$$= \sum_{\mathbf{V}'_i}\left[\exp\left(v_{i1}r_{i1}\right)\exp\left(v_{i2}r_{i2}\right)\ldots\exp\left(v_{iK}r_{iK}\right)\right]$$

$$= \left[\exp\left(r_{i1}(1)\right)\exp\left(r_{i2}(0)\right)\ldots\exp\left(r_{iK}(0)\right)\right] + \left[\exp\left(r_{i1}(0)\right)\exp\left(r_{i2}(1)\right)\ldots\exp\left(r_{iK}(0)\right)\right] +$$

$$\ldots + \left[\exp\left(r_{i1}(0)\right)\exp\left(r_{i2}(0)\right)\ldots\exp\left(r_{iK}(1)\right)\right]$$

$$= \sum_{q=1}^{K}\exp\left(r_{iq}\right)$$

$$= \sum_{q=1}^{K}\exp\left(\sum_{j=1}^{F}W_{ijq}h_j + b_{iq}\right). \tag{13}$$

Equation (13) can be replaced at equation (12) denominator:

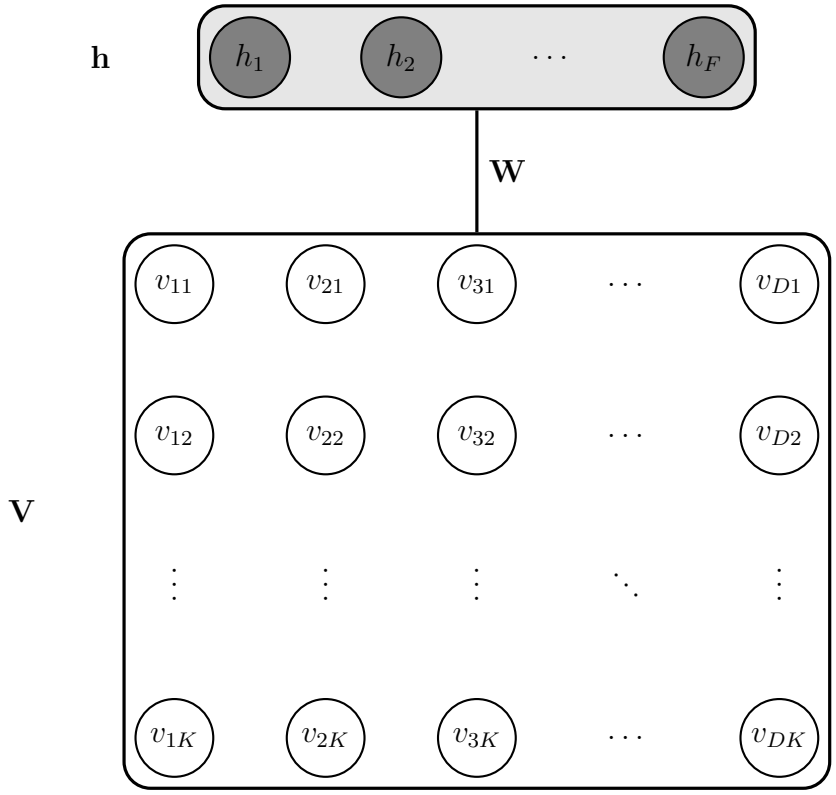
$$
\begin{aligned}
P(\mathbf{V}|\mathbf{h}) &= \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\prod_{i=1}^{D}\left\{\sum_{\mathbf{V}_i'}\prod_{k=1}^{K}\exp\left[v_{ik}'\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]\right\}} \\
&= \frac{\prod_{i=1}^{D}\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\prod_{i=1}^{D}\left[\sum_{q=1}^{K}\exp\left(\sum_{j=1}^{F}W_{ijq}h_j + b_{iq}\right)\right]} \\
P(\mathbf{V}|\mathbf{h}) &= \prod_{i=1}^{D}\frac{\prod_{k=1}^{K}\exp\left[v_{ik}\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)\right]}{\sum_{q=1}^{K}\exp\left(\sum_{j=1}^{F}W_{ijq}h_j + b_{iq}\right)}.
\end{aligned}
\tag{14}
$$

Considering we would like to know the probability of having visible unit i on at word k, $v_{ik} = 1$, given the hidden units state $\mathbf{h}$:

$$
P(v_{ik} = 1|\mathbf{h}) = \frac{\exp\left(\sum_{j=1}^{F}W_{ijk}h_j + b_{ik}\right)}{\sum_{q=1}^{K}\exp\left(\sum_{j=1}^{F}W_{ijq}h_j + b_{iq}\right)},
\tag{15}
$$

which is the softmax function.

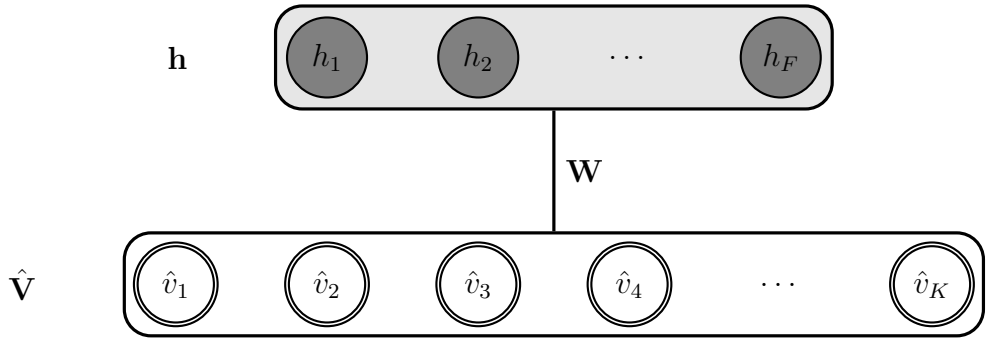
Figure 1 - REPLICATED SOFTMAX MODEL



Legend: Replicated Softmax model. The top layer represents the hidden layer, denoted by vector $\mathbf{h} = (h_1, h_2, \ldots, h_F)$, of stochastic binary units. The bottom layer, denoted by matrix $\mathbf{V}$, represents the visible layer with units v_{ik}, $i = 1, \ldots, D$ and $k = 1, \ldots, K$. Between visible and hidden units are the weights w_{ijk}, which is a symmetric connection, denoted by tensor $\mathbf{W}$.

Source: Author

Figure 2 - REPLICATED SOFTMAX MODEL SIMPLIFICATION



Legend: Replicated Softmax model. The top layer represents the hidden layer, denoted by vector $\mathbf{h} = (h_1, h_2, \ldots, h_F)$, of stochastic binary units, the same hidden layer as in Figure (1). The bottom layer, denoted by vector $\hat{\mathbf{V}}$, represents the visible layer with units $\hat{v}_k$, $k = 1, \ldots, K$. These *hat* units represents the frequency count of each k word of the dictionary that appears in document $\mathbf{V}$. Between visible and hidden units are the weights w_{jk}, which is a symmetric connection, denoted by matrix $\mathbf{W}$. These weights are known as **shared weights**, because, even though documents of different sizes are analysed, the dictionary remains the same, and then the weights for each word in the dictionary is shared among the documents. It is important to remember that word order is not taken into consideration.

Source: Author