Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto de Matemática e Estatística

Artur Chiaperini Grover

Inteligência Computacional Usando Máquinas de Boltzmann

Rio de Janeiro
2019

Artur Chiaperini Grover

Inteligência Computacional Usando Máquinas de Boltzmann

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Curso, da Universidade do Estado do Rio de Janeiro.

Orientador: Dra. Roseli Suzi Wedemann

Rio de Janeiro

2019

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/D

D979 Chiaperini Grover, Artur
Inteligência Computacional Usando Máquinas de Boltzmann / Artur Chiaperini Grover. – Rio de Janeiro, 2019-
22 f.

Orientador: Dra. Roseli Suzi Wedemann
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística, Programa de Pós-Graduação em Curso, 2019.

1. máquina de Boltzmann.. 2. máquina restrita de Boltzmann.. 3. terceira palavra chave.. I. Dra. Roseli Suzi Wedemann. II. Universidade do Estado do Rio de Janeiro. III. Instituto de Matemática e Estatística. IV. Título

CDU 02:141:005.7

Assinatura Data

Artur Chiaperini Grover

Inteligência Computacional Usando Máquinas de Boltzmann

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Curso, da Universidade do Estado do Rio de Janeiro.

Aprovada em 22 de Novembro de 2019.
Banca Examinadora:

Dra. Roseli Suzi Wedemann (Orientador)
IME/CComp – UERJ

Dra. Patrícia Nunes da Silva
IME/CComp – UERJ

Dra. Zochil González Arenas
IME/CComp – UERJ

Rio de Janeiro
2019

If no one died, it is just another story to be told.
[Daniel Mirolhaum]

RESUMO

CHIAPERINI GROVER, A. C. G. *Inteligência Computacional Usando Máquinas de Boltzmann*. 2019. 22 f. Dissertação (Mestrado em Curso) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019.

Texto do resumo em português.

Palavras-chave: máquina de Boltzmann. máquina restrita de Boltzmann. terceira palavra chave.

ABSTRACT

CHIAPERINI GROVER, A. C. G. *Computational Intelligence Using Boltzmann Machines*. 2019. 22 f. Dissertação (Mestrado em Curso) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019.

Abstract in English.

Keywords: Boltzmann machine. restricted Boltzmann machine. third keyword.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BM	Boltzmann Machine
RBM	Restricted Boltzmann Machine
SNN	Stochastic Neural Network
ANN	Artificial Neural Network
KL	Kullback-Leibler

CONTENTS

INTRODUCTION

Nowadays, modelling intelligent complex systems uses two main paradigms, commonly referred to as Symbolic and Connectionist, as basic guidelines for achieving your goals of creating intelligent machines and understanding human cognition. These two approaches depart from different positions, each advocating advantages over the other in reproducing intelligent activity. The traditional symbolic approach argues that the algorithmic manipulation of symbolic systems is an appropriate context for modelling cognitive processes. On the other hand, connectionists restrict themselves to brain-inspired architectures and argue that this approach has the potential to overcome the rigidity of symbolic systems by more accurately modeling cognitive tasks that can only be solved, in the best case, approximately. In addition, connectionist paradigm focus on parallel models which have simple and uniform processing elements interconnected (SUN, 2001).

Years of experimentation with both paradigms lead us to the conclusion that the solution lies between these two extremes, and that the approaches must be integrated and unified. In order to establish a proper link between them, much remains to be researched. If in the 1980s the discussion of intelligence was placed at the distinct poles of the symbolists and connectionists, today the connectionists are divided by the reductionist arguments of the structuralists. For this structuralist current, the failure of the symbolists was due to the fact that their models despised brain architecture, and therefore connectionism must continue to explore more deeply the structural aspects of the thinking organ.

The connectionist and structuralist aspects are approached, respectively, through the paradigm of artificial neural networks and realistic models of the brain, within the area called Computational Neuroscience. Through our models, we investigate ancient questions of Artificial Intelligence regarding the understanding of computability aspects of the human mind.

In this project we will continue with the study and implementation of artificial neural network known as Boltzmann Machine and its derivatives models, such as the Restricted Boltzmann Machine, which have been used to solve artificial intelligence problems, in areas such as: computer vision, encoder problem, statistical pattern recognition, speech recognition, and combinatorial optimization (HERTZ; KROGH; PALMER, 1991). Boltzmann machines is a branch of stochastic methods, where randomness is present. (DUDA; HART; STOCK, 2000).

1 BOLTZMANN MACHINES

In this chapter will expose the theory behind the Boltzmann Machine. A few considerations regarding the notation used in this exposition is required before stepping forward into the main content. A single random variables is denoted by x. A vector of random variables of size n is represented by $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n)$, where each x_i, $i \in \{1, 2, \ldots, n\}$, represents a single unit of the network. The value a random variable can assume is represented by x. Assuming a discrete scenario, the probability of a random variable x_i of assuming a certain value x_i is $P(\mathrm{x}_i = x_i)$.

We begin this chapter by briefly introducing the Hopfield Network. The reason is because the Boltzmann Machine is a generalization of the Hopfield Network. In Hopfield Networks units are deterministic while, in Boltzmann Machines, units are stochastic.

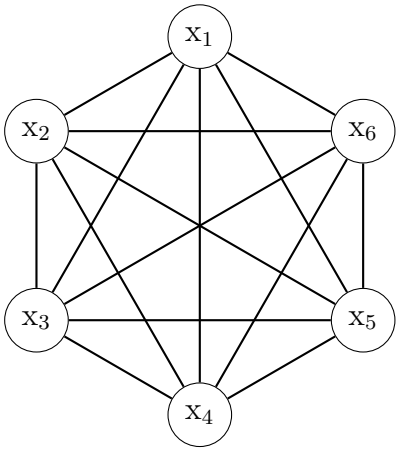
1.1 Hopfield Networks

Hopfield networks are a simple neural network architecture, often referred to as an associative memory network (HERTZ; KROGH; PALMER, 1991). Each unit x_i in a Hopfield network is a binary unit that can assume a value $x_i \in \{0, 1\}$. As Géron (2017) mentioned, this kind of network is first taught a few patterns, and then when exposed to new patterns it will output the closest learned pattern. In the context of image recognition, we consider that each pixel of a binary image maps to one neuron of the network, then the previous statement can be read as the case where the training set contains images of characters, for instance, which are the stored memories of the network; if a new image of a character is presented, the network will recall from the memory the closest character.

In Figure 1, we can see a diagram of the architecture of a Hopfield network. It is a fully connected graph of binary units, which means that each unit is connected to every other unit of the network. It is a different network arrangement compared to perceptron, for instance, where there is no back-coupling (HOPFIELD, 1982).

Hopfield network is an energy-based model, because there is a global energy function associated to the network. This global energy function evolves to a low-energy state during training phase, i.e., the network connections between units are modified so that the energy decreases; the less energy, the better. When connections ω between units, also know as weights or *synaptic strenght*, are symmetrical, i.e., $\omega_{ij} = \omega_{ji}$, where the subscript identify the connection between unit i and j, then Hopfield (1982) presented that the

Figure 1 - Hopfiel Network diagram.



Legend: Each white circle represents a single unit of the Hopfield Network.
Source: Author.

energy function is give by

$$H(\mathbf{x}) = -\frac{1}{2}\sum_i \sum_j \omega_{ij} x_i x_j - \sum_i \phi_i x_i, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the vector of binary values that each unit of the network has for in a particular state, and n is the number of units in the network. Equation (1) shows that the energy is the sum of many contributions. Each contribution depends on one connection weight ω_{ij} and the binary state in which each neuron is, x_i and x_j, first term of the equation. And the second term only involves the state of each individual unit weighted by bias ϕ_i.

If we want to store a pattern in a Hopfield network, we require the network to learn the right connection between units, so that this pattern can be accessed later on. To compute de proper connection in a network with N units, we have to compute each ω_{ij}

$$\omega_{ij} = \frac{1}{N} x_i x_j. \tag{2}$$

The quadratic energy function, equation (1), makes it possible for each unit to compute locally how its state affects the global energy, in another words, how each unit affects the global energy when its state is changed. Define the energy gap ΔH_i for a unit x_i, as the measure of the global energy function difference when the unit x_i has its state

changed.

$$\Delta H_i = H(x_i = 0) - H(x_i = 1) = \sum_j \omega_{ij} x_j + \phi_i, \tag{3}$$

the energy gap equation can also be read as the difference between the energy when x_i is off, $x_i = 0$, and the energy when x_i is on, $x_i = 1$. In addition, it can also be computed by differenciating the energy function H, equation (1). [Derivation to be added to appendix]. The Hopfield network will go down hill in this global energy. To find the energy minimum, start from a random state, update each unit one at a time in random order. Update each unit to whichever of its two states gives the lowest global energy.

According to Hopfield (1982), memories can be seen as energy minima. Furthermore, by just knowing some parts of an energy minima, its possible to access that memory. An interesting analogy, is that Hopfield networks are like reconstructing a dinosaur from a few bones.

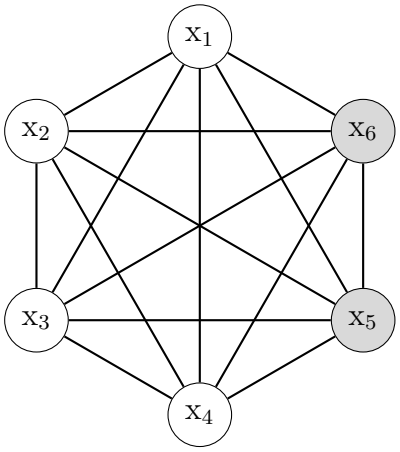
1.2 **Boltzmann Machines**

Boltzmann Machines are a type of stochastic neural networks where the connections between units, which are described by ω, are symmetrical, i.e., $\omega_{ij} = \omega_{ji}$ (HERTZ; KROGH; PALMER, 1991). This kind of stochastic neural networks are capable of learning internal representation and to model an input distribution. Boltzmann Machines were named after the Boltzmann distribution. Due to its stochatics behaviour, the probability of the state of the system to be found in a certain configuration is given by previous mentioned distribution (HERTZ; KROGH; PALMER, 1991). According to (MONTÚFAR, 2018), Boltzmann machines can be seen as an extension of Hopfield networks to include hidden units and units with a stochastic behaviour.

Boltzmann Machines have two kind of units x_i: the visible and hidden units. The visible units are linked to the external world and they correspond to the components of an observation, where data is input. On the other hand, the hidden units do not have any connection outside of the network and they model the dependencies between the components of the observations (FISCHER; IGEL, 2012), i.e., they are responsible for finding the data relation from the input. In Boltzmann machines, there is no connection restriction, this means that every unit, visible or hidden, can be connected to every other unit as in a complete graph, this pattern is not mandatory as some of the connections may not exists depending on the network layout. Regardless of how the connections are, if there is a connection bewteen two units, this connection is symmetric. Training a Boltzmann Machines means finding the right connections ω_{ij} between the units.

In Figure 2, we have a diagram of a Boltzmann machine layout. We can see that

Figure 2 - Boltzmann machine diagram.



Legend: Gray circles represent the hidden units of the Boltzmann Machine, while the white circles are the visible units.

Source: Author.

each unit is connected to every other unit in the network regardless of it being a visible or hidden unit.

Stochastics units x_i compose the Boltzmann machine. Stochastics units are random variables x that can assume a binary value with a certain probability. We will consider that a random variable x_i can assume a value of $x_i = 1$ with probability $g(h_i)$, and $x_i = 0$, otherwise, i.e.,

$$x_i = \begin{cases} 1 \text{ with probability } g(h_i) \\ 0 \text{ with probability } 1 - g(h_i) \end{cases}, \tag{4}$$

where the probability $g(h_i)$ is given by

$$g(h_i) = \frac{1}{1 + e^{-2\beta h_i}}, \tag{5}$$

and

$$h_i = \sum_j \omega_{ij} x_j + \phi_i, \tag{6}$$

equation (6) stands for the input a unit x_i receives.

Likewise the Hopfield network, due to the symmetrical connections, there is an energy function which is also given by equation (1). This energy function has minimum

when there is a stable state characterised by

$$x_i = sgn(h_i). \tag{7}$$

For stochastics neural networks the probability P of finding the system in a given state $\mathbf{x}$ after the equilibrium is reached is

$$P(\mathbf{x}) = \frac{1}{Z} e^{-\beta H(\mathbf{x})}, \tag{8}$$

where

$$Z = \sum_{\mathbf{x}'} e^{-\beta H(\mathbf{x}')} \tag{9}$$

is the partition function. The vector $\mathbf{x}$ represents the state in which the units of the network are, for instance, the following state, in a 3 units stochastic network, $\mathbf{x} = (x_1, x_2, x_3) = (1, 0, 1)$.

Now let us consider the Boltzmann machine case where there are two kinds of units, i.e., the visible and the hidden units. Let us identify the state of the visible units by an index v and the state of the hidden units by an index h. Considering a particular system with N visible units and K hidden units, the whole system have 2^{N+K} possibilities of states in which it can be found.

The joint probability P_{vh} is the probability of finding the visible and hidden units in the states v and h, respectively. This probability distribution is given by the Boltzmann distribution

$$P_{vh} = \frac{1}{Z} e^{-\beta H_{vh}}, \tag{10}$$

where

$$Z = \sum_u \sum_k e^{-\beta H_{uk}}, \tag{11}$$

and

$$H_{vh} = -\sum_i \sum_j \omega_{ij} x_i^{(vh)} x_j^{(vh)} - \sum_i \phi_i x_i^{(vh)}. \tag{12}$$

Equations (10), (11) and (12) are ajustments to equations (8), (9) and (1), respectively, where now the different kind of units state is taken into consideration. In equation (11), the indexes u and k refer to visible and hidden units states, respectively, different indexes are used to avoid bewilderment.

As previously mentioned, the problem a Boltzmann Machine is trying to solve is

determining the connections ω_{ij} between units such that the visible units have a certain probability distribution. In order to do that, we need to find the marginal probability of the state v in which the visible units are found regardless of the state h of the hidden units. The marginal probability P_v is given by

$$P_v = \sum_h P_{vh} = \sum_h \frac{e^{-\beta H_{vh}}}{Z}. \tag{13}$$

Although we know that P_v is a function of the connections ω_{ij}, and that this is the probability of finding the visible units in the state v. We want the states to have a certain probability R_v, i.e., a desired probability. This means that ideally we would like to match the empirical distribution of the data, even though we do not have access to the correct distribution, only to what the observed data has given us as an input to training the model.

One way to evaluate the difference between two probability distribution, for example, P_v and R_v, is using the Kullback-Leibler divergence $D_{KL}(R_v||P_v)$, which can also be referred to as relative entropy, E, which will be our cost function. Further comments about the ullback-Leibler divergence can be found on appendix A.

$$E = \sum_v R_v \ln\left(\frac{R_v}{P_v}\right). \tag{14}$$

The relative entropy E has the property of always being equal or greater than zero. It reaches zero only if $P_v = R_v$, which means that we are able to retrieve the exactly desired probability distribution at the visible units from the input data. We have to minimise E using the gradient descent (HERTZ; KROGH; PALMER, 1991), relative to the weights ω_{ij} and the bias ϕ_i,

$$\Delta\omega_{ij} = -\eta\frac{\partial E}{\partial\omega_{ij}}, \tag{15}$$

and

$$\Delta\phi_i = -\eta\frac{\partial E}{\partial\phi_i}, \tag{16}$$

where

$$\begin{aligned} E &= \sum_v R_v \ln\left(\frac{R_v}{P_v}\right) \\ &= \sum_v \left[\ln\left(R_v\right) - \ln\left(P_v\right)\right]. \end{aligned} \tag{17}$$

In the following steps, we present the derivation of the gradient descent, based on

the derivation in (HERTZ; KROGH; PALMER, 1991)

$$\begin{aligned}
\Delta\omega_{ij} &= -\eta\frac{\partial E}{\partial\omega_{ij}} \\
&= -\eta\frac{\partial}{\partial\omega_{ij}}\left[\sum_v R_v\left(\ln\left(R_v\right)-\ln\left(P_v\right)\right)\right] \\
&= \eta\frac{\partial}{\partial\omega_{ij}}\left[\sum_v R_v\ln\left(P_v\right)\right] \\
&= \eta\sum_v R_v\frac{\partial}{\partial\omega_{ij}}\left[\ln\left(P_v\right)\right] \\
\Rightarrow\Delta\omega_{ij} &= \eta\sum_v\frac{R_v}{P_v}\frac{\partial P_v}{\partial\omega_{ij}}.
\end{aligned} \tag{18}$$

To continue with the computation of $\Delta\omega_{ij}$, we have to find the derivative of $\partial P_v/\partial\omega_{ij}$, from the marginal probability, equation 13,

$$P_v = \frac{\sum_h e^{-\beta H_{vh}}}{\sum_u\sum_k e^{-\beta H_{uk}}}, \tag{19}$$

thus the derivative of P_v follows

$$\begin{aligned}
\frac{\partial P_v}{\partial\omega_{ij}} &= \frac{\partial}{\partial\omega_{ij}}\left[\frac{\sum_h e^{-\beta H_{vh}}}{\sum_u\sum_k e^{-\beta H_{uk}}}\right] \\
&= \frac{1}{\sum_u\sum_k e^{-\beta H_{uk}}}\sum_h(-\beta)e^{-\beta H_{vh}}\frac{\partial H_{vh}}{\partial\omega_{ij}} \\
&- \sum_h e^{-\beta H_{vh}}\frac{1}{\left(\sum_u\sum_k e^{-\beta H_{uk}}\right)^2}\sum_u\sum_k e^{-\beta H_{uk}}(-\beta)\frac{\partial H_{uk}}{\partial\omega_{ij}}.
\end{aligned} \tag{20}$$

Following equation (20), we need to compute the term $\partial H_{vh}/\partial\omega_{ij}$,

$$\begin{aligned}
\frac{\partial H_{vh}}{\partial\omega_{ij}} &= \frac{\partial}{\partial\omega_{ij}}\left[-\frac{1}{2}\sum_m\sum_n\omega_{mn}x_m^{(vh)}x_n^{(vh)} - \sum_m\phi_{mm}x_m^{(vh)}\right] \\
&= \frac{\partial}{\partial\omega_{ij}}\left[-\frac{1}{2}\sum_{m\neq i,j}\sum_{n\neq i,j}\omega_{mn}x_m^{(vh)}x_n^{(vh)} - \frac{1}{2}\omega_{ij}x_i^{(vh)}x_j^{(vh)} - \frac{1}{2}\omega_{ji}x_j^{(vh)}x_i^{(vh)} - \sum_m\phi_m x_m^{(vh)}\right],
\end{aligned} \tag{21}$$

as the connections between units are symmetric, i.e., $\omega_{ij} = \omega_{ji}$, we can simplify equa-

tion (21),

$$\begin{aligned}
\frac{\partial H_{vh}}{\partial \omega_{ij}} &= \frac{\partial}{\partial \omega_{ij}}\left[-\frac{1}{2}\sum_{m\neq i,j}\sum_{n\neq i,j}\omega_{mn}x_m^{(vh)}x_n^{(vh)} - \omega_{ij}x_i^{(vh)}x_j^{(vh)} - \sum_m \phi_m x_m^{(vh)}\right] \\
&= \frac{\partial}{\partial \omega_{ij}}\left[-\frac{1}{2}\sum_{m\neq i,j}\sum_{n\neq i,j}\omega_{mn}x_m^{(vh)}x_n^{(vh)}\right] + \frac{\partial}{\partial \omega_{ij}}\left[-\omega_{ij}x_i^{(vh)}x_j^{(vh)}\right] + \frac{\partial}{\partial \omega_{ij}}\left[-\sum_m \phi_m x_m^{(vh)}\right] \\
\Rightarrow &\frac{\partial H_{vh}}{\partial \omega_{ij}} = -x_i^{(vh)}x_j^{(vh)}.
\end{aligned} \tag{22}$$

Analagous to $\partial H_{vh}/\partial \omega_{ij}$, we have the derivative of H_{uk} which is

$$\frac{\partial H_{uk}}{\partial \omega_{ij}} = -x_i^{(uk)}x_j^{(uk)}. \tag{23}$$

Going back to equation (20), we can replace the derivatives of H, and solve the derivative of the marginal probability P_v,

$$\begin{aligned}
\frac{\partial P_v}{\partial \omega_{ij}} &= \frac{1}{Z}\sum_h e^{-\beta H_{vh}}(-\beta)(-x_i^{(vh)}x_j^{(vh)}) - \frac{1}{Z^2}\sum_h e^{-\beta H_{vh}}\sum_u\sum_k e^{-\beta H_{uk}}(-\beta)(-x_i^{(uk)}x_j^{(uk)}) \\
&= \beta\left[\sum_h \frac{e^{-\beta H_{vh}}}{Z}x_i^{(vh)}x_j^{(vh)} - \sum_h \frac{e^{-\beta H_{vh}}}{Z}\sum_u\sum_k \frac{e^{-\beta H_{uk}}}{Z}x_i^{(uk)}x_j^{(uk)}\right] \\
&= \beta\left[\sum_h P_{vh}x_i^{(vh)}x_j^{(vh)} - P_v\sum_u\sum_k P_{uk}x_i^{(uk)}x_j^{(uk)}\right] \\
\Rightarrow &\frac{\partial P_v}{\partial \omega_{ij}} = \beta\left[\sum_h P_{vh}x_i^{(vh)}x_j^{(vh)} - P_v\langle x_i x_j\rangle\right].
\end{aligned} \tag{24}$$

Given the derivative of P_v in relation to ω_{ij}, we can compute the learning term $\Delta\omega_{ij}$, from equation (18),

$$\begin{aligned}
\Delta\omega_{ij} &= \eta\sum_v \frac{R_v}{P_v}\beta\left[\sum_h P_{vh}x_i^{(vh)}x_j^{(vh)} - P_v\langle x_i x_j\rangle\right] \\
&= \eta\beta\left[\sum_v\sum_h \frac{R_v}{P_v}P_{vh}x_i^{(vh)}x_j^{(vh)} - \sum_v \frac{R_v}{P_v}P_v\langle x_i x_j\rangle\right] \\
&= \eta\beta\left[\sum_v\sum_h R_v\frac{P_{vh}}{P_v}x_i^{(vh)}x_j^{(vh)} - \sum_v R_v\langle x_i x_j\rangle\right] \\
&= \eta\beta\left[\sum_v\sum_h R_v P_{h|v}x_i^{(vh)}x_j^{(vh)} - \langle x_i x_j\rangle\right].
\end{aligned} \tag{25}$$

In the above derivation, we have used the following relations,

$$P_{h|v} = \frac{P_{vh}}{P_v}, \tag{26}$$

which is the conditional probability equation. In our scenario this equation means that the probability distribution of the hidden units in state h given the state v of the visible units is the joint probability distribution of both states, v and h, divided by the marginal probability distribution of the visible units in state v.

The second term in equation (25), is the average of units i and j over all combinations of states v and h of the system. In other words, we would have to compute all possible combination of states of visible and hidden units, v and h, and then average over the specific units i and j.

The first term can be simplified by

$$\sum_v R_v \sum_h P_{h|v} x_i^{(vh)} x_j^{(vh)} = \sum_v R_v \langle x_i x_j \rangle^{(v)} = \langle \langle x_i x_j \rangle^{(v)} \rangle. \tag{27}$$

Then equation (25) becomes

$$\Delta\omega_{ij} = \eta\beta \left[\langle \langle x_i x_j \rangle^{(v)} \rangle_{clamped} - \langle x_i x_j \rangle_{free} \right], \tag{28}$$

it is important to notice that the subscripts *clamped* means that we have to fix a certain v state on the visible units otherwise the second term in the equation does not have a reference. On the other hand, the subscript *free* identify the ...

FUTURE DEVELOPMENT

To be studied...

Following up the current status of the project, the next steps includes:

1. A deeper understanding of the Restricted Boltzmann Machines and how its learning algorithm works. Also
2. Find a proper contextt where the RBM can be applied and tested.

BIBLIOGRAPHY

DUDA, R. O.; HART, P. E.; STOCK, D. G. *Patter Classification*. 2. ed. USA: Wiley-Interscience, 2000.

FISCHER, A.; IGEL, C. An introduction to restricted boltzmann machines. In: ALVAREZ, L. et al. (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 14–36.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 1. ed. USA: O'Reilly Media Inc., 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. (Adaptive computation and machine learning series).

HERTZ, J.; KROGH, A.; PALMER, R. D. *Introduction to the theory of neural computation*. USA: Westview Press, 1991. v. 1. (Santa Fe Institute Series, v. 1).

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, v. 79, p. 2554–2558, Apr 1982.

MONTÚFAR, G. Restricted boltzmann machines: Introduction and review. *arXiv e-prints*, p. arXiv:1806.07066, Jun 2018.

SUN, R. Artificial intelligence: connectionist and symbolic approaches. In: SMELSER, N. J.; BALTES, P. B. (Ed.). *International Encyclopedia of the Social and Behavioral Sciences*. Oxford: Pergamon/Elsevier, 2001. p. 783–789.

APPENDIX A – Kullback-Leibler Divergence or Relative Entropy

A.1 Kullback-Leibler Divergence

The entropy computes how uncertainty events are. If a system is quite stable, the entropy will be close to zero and will not vary a lot, while if the system is quite unstable, then the entropy will be very large. The Kullback-Leibler divergence, or relative entropy, is the measure of how different two separate probability distributions, R and P, are from each other (GOODFELLOW; BENGIO; COURVILLE, 2016). Considering these probability distribution are both discrete and are defined on the same probability space χ, the Kullback-Leibler divergence $D_{KL}(R||P)$ is given by

$$D_{KL}(R||P) = \sum_{x\in\chi} R(x) \ln\left(\frac{R(x)}{P(x)}\right), \tag{29}$$

the equation above can be read as the divergence from P to R.

It is important to notice that this divergence measurement is not symmetric,

$$\begin{aligned} D_{KL}(R||P) &\neq D_{KL}(P||R), \\ \sum_{x\in\chi} R(x) \ln\left(\frac{R(x)}{P(x)}\right) &\neq \sum_{x\in\chi} P(x) \ln\left(\frac{P(x)}{R(x)}\right), \end{aligned} \tag{30}$$

and that KL divergence, relative entropy, is always a non-negative value,

$$\begin{aligned} E &= D_{KL}(R||P) \\ D_{KL}(R||P) &= \sum_{x\in\chi} R(x) \ln\left(\frac{R(x)}{P(x)}\right) \\ &\geq \sum_{x\in\chi} R(x) \left(1 - \frac{P(x)}{R(x)}\right) \\ &= \sum_{x\in\chi} (R(x) - P(x)) \\ &= \sum_{x\in\chi} R(x) - \sum_{x\in\chi} P(x) = 1 - 1 \\ &\Rightarrow D_{KL}(R||P) \geq 0. \end{aligned} \tag{31}$$

Notice that $D_{KL}(R||P)$ will only reach zero if and only if $P(x) = R(x)$, which means that we are able to retrieve the exactly desired probability distribution at the visible units from the input data.