Universidade do Estado do Rio de Janeiro

centro

unidade patrono

Artur Chiaperini Grover

Computational Intelligence Using Deep Neural Networks (BM)

Rio de Janeiro

aaaa

Artur Chiaperini Grover

Computational Intelligence Using Deep Neural Networks (BM)

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: Cargo Titulação Nome Sobrenome
Coorientador: Cargo Titulação Nome Sobrenome

Rio de Janeiro
aaaa

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

D979 Chiaperini Grover, Artur
Computational Intelligence Using Deep Neural Networks (BM) / Artur Chiaperini Grover. – Rio de Janeiro, aaaa-
34 f.

Orientador: Cargo Titulação Nome Sobrenome
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, unidade, Programa de Pós-Graduação em Curso, aaaa.

1. primeira palavra chave.. 2. segunda palavra chave.. 3. terceira palavra chave.. I. Cargo Titulação Nome Sobrenome. II. Universidade do Estado do Rio de Janeiro. III. unidade. IV. Título

CDU 02:141:005.7

Artur Chiaperini Grover

Computational Intelligence Using Deep Neural Networks (BM)

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 dd de Mês de aaaa.

Banca Examinadora:

Cargo Titulação Nome Sobrenome (Orientador)
Unidade – Instituição

Cargo Titulação Nome Sobrenome (Coorientador)
Unidade – Instituição

primeiro membro titular da banca
instituição

segundo membro titular da banca
instituição

terceiro membro titular da banca
instituição

primeiro membro suplente da banca
instituição

segundo membro suplente da banca
instituição

terceiro membro suplente da banca
instituição instituição instituição instituição

Rio de Janeiro

aaaa

DEDICATÓRIA

Texto da dedicatória

AGRADECIMENTOS

Texto de agradecimento

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

RESUMO

CHIAPERINI GROVER, A. C. G. *Computational Intelligence Using Deep Neural Networks (BM)*. aaaa. 34 f. Dissertação (Mestrado em Curso) – unidade, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, aaaa.

Texto do resumo em português.

Palavras-chave: primeira palavra chave. segunda palavra chave. terceira palavra chave.

ABSTRACT

CHIAPERINI GROVER, A. C. G. *Inteligência Computacional Usando Redes Neuronais Profundas*. aaaa. 34 f. Dissertação (Mestrado em Curso) – unidade, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, aaaa.

Abstract in English.

Keywords: first keyword. second keyword. third keyword.

LIST OF FIGURES

LIST OF TABLES

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

sigla1	por extenso
sigla2	por extenso
sigla3	por extenso

LISTA DE SÍMBOLOS

$simbolo1$	significado e/ou valor
$simbolo2$	significado e/ou valor
$simbolo3$	significado e/ou valor

CONTENTS

INTRODUÇÃO

Texto da introdução. Texto, texto texto (NADA, a), texto nada (a). Texto nada em b, texto NADA, texto.

1 BOLTZMANN MACHINES

EXPLICAR OS ELEMENTOS DE PROBABILIDADE COM OS QUAIS ESTAMOS LIDANDO:QUEM SÃO AS VARIÁVEIS ALEATÓRIAS E COMO VAMOS DOMINÁ-LAS NO TEXTO, COMENTAR QUE ESTAMOS TRATANDO COM VARIÁVEIS DISCRETAS, COMO VAMOS IDENTIFICAR AS PROBABILIDADES,...

EXPLICAR O QUE OS ÍNDICES REPResentam, e as variiáveis.

EXPLICAR OS TERMOS ω e ϕ. ω É para os pesos, e ϕ para o bias de cada unidade.

EXPLICAR MINHA NOTAÇÃO sobre o sobre-escrito entre parênteses!!!

In this chapter will expose the Boltzmann Machine theory. 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 $P(\mathrm{x}_i = x_i)$.

1.1 Hopfield Networks

Hopfield networks are a simple neural network architecturei that can be used for storing memories of patterns of activities. Energy-based model, because property derives from global energy function. Binary units with recurrent connections between them. Is the connections are symmetric, there is a global energy function: each binary configuration of the whole netwark has an energy. Binary threshold decision rule causes the network to settle to a minimum of this energy function. Energy function is the sum of many contributions. Each contribution depends on one connection weight and the ninary state of two neurons is. (Energy is bad, so that is why we have a minus sign. The less energy, the better.)

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

where the first term is the symmetric connection between neurons ω_{ij} and the activity of the two connected neurons x_i and x_j. The second term only involves the state of individual units.

The quadraticc energy function makes it possible for each unit to compute locally how it's 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, this measure is the global energy function difference when the unit x_i has its state changed.

$$\Delta H_i = H(x_i = 0) - H(x_i = 1) = \sum_j \omega_{ij} x_j + \phi_i \tag{2}$$

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

1.2 Boltzmann Machines

Boltzmann Machines (BM) are a type of stochastic neural networks (SNN) where the connections between units, which are described by ω, are symmetrical, i.e., $\omega_{ij} = \omega_{ji}$ [HERTZ]. 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]. According to [MONTUFAR, 2018], BM can be seen as an extension of Hopfield networks to include hidden units.

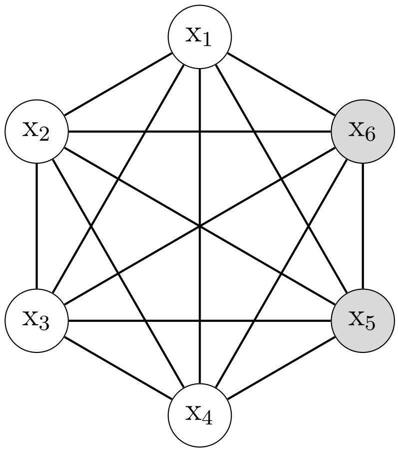
Boltzamann Machines have visible and hidden units. The visible units are linked to the external world and they correspond to the components of an observation. On the other hand, the hidden units do not have any connection outside of the network and model the dependencies between the components of the observations [FISCHER, 2012]. In BM, 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.

Training Boltzmann Machines means finding the right connection between the units.

Boltzmann Machines (BM) are stochastics neural networks with symmetric connections, i.e., $\omega_{ij} = \omega_{ji}$. Boltzmann Machines use the Boltzmann distribution to determine the probability of the state of the system of the network. BM ressambles the Hopfield networks with the inclusion of hidden units. Finding the right connections between the hidden units without knowing it from the training patterns what the hidden units represent is part of the solving the Boltzmann Machine problem.

Units x_i in BM are split into two kinds: visible and hidden units. The visible

Figure 1 - Boltzmann machine diagram.



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

units have connection to the outside world and are the units that receive the data input. On the other hand, the hidden units do not have any connection to the outside of the network and they are resposible to find the data relation from the input. In a BM, the connections between units can be complete or not. Regardless of how the connections are, every connection in a BM is symmetric.

BM are made of stochastics units x_i. Stochastics units are random variables that can assume a binary value with a certain probability. We will consider that a random variable x_i can assume a value $x_i \in \{0, 1\}$, iė;

$$x_i = \begin{cases} 1 \text{ with probability } g(h_i) \\ 0 \text{ with probability } 1 - g(h_i) \end{cases}, \tag{3}$$

where the probability is given by

$$g(h_i) = \frac{1}{1 + e^{-2\beta h_i}}, \tag{4}$$

and

$$h_i = \sum_j \omega_{ij} x_j. \tag{5}$$

Due to the symmetrical connections, there is an energy function give by

$$H(\mathbf{x}) = -\sum_i \sum_j \omega_{ij} x_i x_j - \sum_i \phi_i x_i, \tag{6}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and n is equal to the number of units in the network, and the above energy function has minimum when there is a stable state characterised by

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

The probability P of finding the system in a given state $\mathbf{x}$ after the equilibrium is reached can be computed as follows:

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

where

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

is the partition function.

The learning process of a Boltzmann Machine consists in ajusting the connections ω_{ij} in such a way that the state of the visible units have a particular desired probability distribution.

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 system which has 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 probabiblity measument is given by the Boltzmann distribution:

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

where

$$Z = \sum_{u} \sum_{k} e^{-\beta H_{uk}}, \tag{11}$$

and

$$H_{vh} = -\sum_{i} \sum_{j} \omega_{ij} x_i^{(vh)} x_j^{(vh)} - \sum_{i} \phi_i x_i^{(vh)}. \tag{12}$$

As metioned above, 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, which can also be referred to relative entropy, E, which will be our cost function. (EXPLICAR FUNÇÃO DE CUSTO e D_{KL}!!!).

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

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 probability distribution of the input data at the visible units.

$$\begin{aligned}
E &= \sum_v R_v \ln\left(\frac{R_v}{P_v}\right) \\
&\geq \sum_v R_v \left(1 - \frac{P_v}{R_v}\right) \\
&= \sum_v (R_v - P_v) \\
&= \sum_v R_v - \sum_v P_v = 1 - 1 \\
&\Rightarrow E \geq 0.
\end{aligned} \tag{15}$$

From the gradient descent equation

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

where

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

In the following steps, we present the gradient descent derivation

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

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 the 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., $\omega_{ij} = \omega_{ji}$, then we can simplify equation 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 H_{uk} derivative, 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 ...

2 TÍTULO DO CAPÍTULO 1

Texto do capítulo. Texto, texto, Figura 2. Texto Figura 3(a).

Figure 2 - Título da figura.

Legend: Texto da legenda.
Source: Citação da fonte ou 'O autor.'.

Table 1 - Título da tabela.

X	Y
1,20	15,7
1,23	15,6
1,19	15,3
1,26	15,1
1,22	15,5
1,16	15,3
1,37	15,7

Legend: Texto da legenda.
Source: Citação da fonte ou 'O autor.'.

Figure 3 - Título da figura.

(a)

(b)

(c)

Legend: Texto da legenda. (a) Texto da imagem. (b) Texto da imagem. (c) Texto da imagem.

Source: Citação da fonte ou 'O autor'.

3 TÍTULO DO CAPÍTULO 2

Texto do capítulo. Texto, texto Algoritmo 1. Texto.

Algoritmo 1 - Título do algoritmo.

DOCUMENTAÇÃO

TÍTULO

Nome do algoritmo

PROPÓSITO

Propósito do algoritmo.

MÉTODO

Método utilizado no algoritmo.

ENTRADAS

a, m: multiplicador e módulo
n0: semente
i: contador auxiliar

SAÍDAS

n: número aleatório

OBSERVAÇÕES, RESTRIÇÕES, REQUISITOS

Observações, restrições e requisitos.

```
ALGORITMO IDENTIFICAÇÃO
  declarar a, m, i numéricos
  declarar n0, n numéricos
  m ← 13
  n0 ← 1
  para a de 2 até m − 1 , fazer      {para cada possível valor de 'a'}
  |  escrever "a = ", a, ": n = {"
  |  n ← n0      {reinicia a geração com a semente n0}
  |  para i de 0 até m − 1 , fazer
  |  |  n ← resto(a ∗ n, m)      {gerador de números aleatórios}
  |  |  se (n == n0), então      {se fim da sequencia ...}
  |  |  |  escrever n,"}"
  |  |  |  parar
  |  |  senão
  |  |  |  escrever n
  |  |  fim se
  |  fim para
  fim para
```

— *continua* —

Algoritmo 1 - Título do algoritmo. (continuação)

— *continuação* —
$a \leftarrow 1$
enquanto $(a < 10)$, **fazer** {*comentário*}
| **escrever** a
| $a \leftarrow a + 1$
fim enquanto
$a \leftarrow 1$
repetir {*comentário*}
| **escrever** a
| $a \leftarrow a + 1$
até que $(a \geq 10)$

$a \leftarrow 1$
fazer {*comentário*}
| **escrever** a
| $a \leftarrow a + 1$
enquanto $(a < 10)$
FIM ALGORITMO
FIM DOCUMENTAÇÃO

CONCLUSÃO

Texto da conclusão.

BIBLIOGRAPHY

[S.l.: s.n.].

[S.l.: s.n.].

GLOSSÁRIO

termo	significado
termo	significado
termo	significado

APPENDIX A – Primeiro apêndice

A.1 **Primeira seção**

Texto da primeira seção.

A.1.1 Primeira subseção

Texto da primeira subseção.

A.1.1.1 Primeira subsubseção

Texto da primeira subsubseção.

APPENDIX B – Segundo apêndice

B.1 **Primeira seção**

Texto da primeira seção.

B.1.1 Primeira subseção

Texto da primeira subseção.

B.1.1.1 Primeira subsubseção

Texto da primeira subsubseção.

ANNEX A – Primeiro anexo

A.1 **Primeira seção**

Texto da primeira seção.

A.1.1 Primeira subseção

Texto da primeira subseção.

A.1.1.1 Primeira subsubseção

Texto da primeira subsubseção.

ANNEX B – Segundo anexo

B.1 **Primeira seção**

Texto da primeira seção.

B.1.1 Primeira subseção

Texto da primeira subseção.

B.1.1.1 Primeira subsubseção

Texto da primeira subsubseção.