Capítulo 10

Modelos de Markov não observáveis

A metodologia de reconhecimento de fala baseada nos modelos de Markov não-observáveis (HMM - *Hidden Markov Models*) é uma das mais utilizadas. A teoria dos HMM foi publicada por Baum em 1966 [Baum (66)], sendo a primeira aplicação em reconhecimento de fala proposta por Jelinek logo em 1969 [Jelinek (69)]. No entanto, só a partir de 1975 estas aplicações começaram a ser reportadas com regularidade [Baker (75)], [Jelinek (75)], [Bahl (75)] e só na década de 80 apareceram na literatura um conjunto de artigos explicitando a teoria básica [Levinson (83)], [Juang (84)], [Rabiner (89)], [Rabiner (93)], que permitiu que esta metodologia se tornasse tão popular.

Neste capítulo faremos uma breve introdução aos fundamentos dos HMM discretos e apresentaremos o princípio da sua aplicação no reconhecimento de sinais de fala.

10.1 Processo Discreto de Markov

Considere-se um sistema que num determinado instante de tempo se encontra no estado i de entre N estados possíveis S_1, S_2, ... S_N. A intervalos de tempo regulares o sistema evolui para outro estado ou eventualmente permanece no mesmo, em função de uma probabilidade de transição entre estados. Designaremos os diversos instantes de tempo por t=1,2,.... e o estado no instante t por q_t. A descrição probabilística deste processo estocástico requer o conhecimento dos estados ocupados nos instantes passados, ou seja

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) \qquad 1 \le i,j,k \le N. \tag{10.1}$$

Num processo de Markov de primeira ordem a descrição probabilística é condicionada apenas ao estado no instante anterior, podendo ser representado através de uma matriz de transição entre estados $A=\{a_{ij}\}$, independente do instante de tempo, em que cada elemento é definido por:

$$a_{ij} = P\left(q_t = S_j \middle| q_{t-1} = S_i\right) \qquad 1 \le i,j \le N. \tag{10.2}$$

Esta matriz verifica as restrições estocásticas de definição de probabilidades, nomeadamente:

$$\begin{aligned} a_{ij} &\ge 0 \\ \sum_{j=1}^{N} a_{ij} &= 1 \end{aligned} \tag{10.3}$$

Como exemplo, considere-se o modelo de Markov com 3 estados, para descrever de um modo simplificado o estado do tempo [Rabiner(89)]. Neste modelo, cada estado corresponde à observação, uma vez por dia, das seguintes condições atmosféricas:

Estado 1: Dia chuvoso; *Estado* 2: Dia nublado; *Estado* 3: Dia com sol;

Assumindo que o estado do tempo num dia apenas depende do estado do tempo no dia anterior e que a matriz de transição de estados é dada por:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \qquad (10.4)$$

obtém-se a seguinte cadeia de Markov, ilustrada na figura 10.1

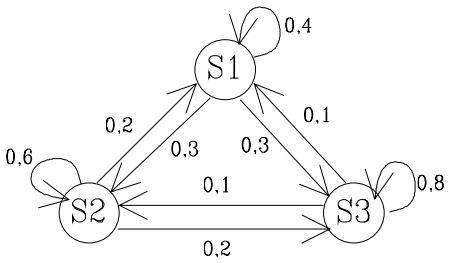

Figura 10.1
Exemplo de Cadeia de Markov com 3 estados

Admitindo que o tempo num determinado dia é de sol (estado 3), pode perguntar-se, por exemplo, qual a probabilidade de os 7 dias seguintes serem dias de sol-sol-chuva-chuva-sol-nublado-sol. Se definirmos a sequência de observações *O* correspondente à sequência de estados $\{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$, a probabilidade da sequência *O* dado o modelo é dada por:

$$
\begin{aligned}
P(O \mid \mathit{Modelo}) &= P(S_3 \mid S_3)P(S_3 \mid S_3)P(S_1 \mid S_3) \\
&\times P(S_1 \mid S_1)P(S_3 \mid S_1)P(S_2 \mid S_3)P(S_3 \mid S_2) \\
&= a_{33}a_{33}a_{31}a_{11}a_{13}a_{32}a_{23} \\
&= 0.8 \times 0.8 \times 0.1 \times 0.4 \times 0.3 \times 0.1 \times 0.2 \\
&= 1.536 \times 10^{-4}
\end{aligned}
\quad . \quad (10.5)
$$

O processo descrito é denominado de modelo de Markov observável, uma vez que a cada observação corresponde um estado. Este modelo é no entanto bastante restritivo e incapaz de ser utilizado em muitos problemas reais. Para tornar o modelo mais flexível, associa-se a cada estado uma função distribuição de probabilidade de observações. Assim, cada estado pode gerar uma observação, de entre um conjunto, de acordo com esta distribuição. A mesma sequência de observações pode assim ser gerada, com probabilidades diferentes, através de sequências diferentes de estados. A sequência de estados que gera uma sequência de observações não é conhecida, pelo que este modelo se denomina de não-observável. Estes modelos encontram aplicações na solução de uma grande variedade de problemas.

Para ilustrar melhor este duplo processo estocástico, considere-se um conjunto de *N* urnas, cada uma com *M* bolas de cores diferentes. Estas urnas estão colocadas por detrás de uma cortina, apenas visível a um indivíduo que as manuseia. Este indivíduo, de acordo com determinado processo aleatório, escolhe uma urna inicial e retira dela uma bola mostrando-a através da cortina. A cor da bola é a única observação para quem está na sala. Seguidamente, a bola é colocada na urna de onde foi retirada e é escolhida outra urna através de um processo aleatório que depende apenas da última urna escolhida. Para este caso, os estados correspondem às urnas escolhidas e as cores das bolas às observações, sendo a probabilidade de cada cor definida diferentemente para cada urna.

10.2 Elementos de um HMM

Um modelo de Markov não-observável é caracterizado através dos seguintes elementos:

1) O número N de estados. Cada estado é denotado de $S=\{S_1,S_2,...,S_N\}$ e o estado no instante t denotado por q_t.

2) A distribuição de probabilidades inicial para cada estado $\pi=\{\pi_i\}$

$$\pi_i = P(q_1 = S_i) \quad 1 \le i \le N. \tag{10.6}$$

3) A distribuição de probabilidades de transição entre estados definida pela matriz $A=\{a_{ij}\}$ em que

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i) \quad 1 \le i,j \le N. \tag{10.7}$$

4) O número M de símbolos distintos observáveis por estado. Estes símbolos denotam-se de $V=\{V_1,V_2,....,V_M\}$. Dado que neste caso existe um número finito de símbolos, o modelo representado denomina-se de modelo discreto.

5) A distribuição de probabilidade dos símbolos observáveis para cada estado S_j , $B=\{b_j(k)\}$ em que:

$$b_j(k) = P(v_k \text{no instante } t \mid q_t = S_j) \quad 1 \le j \le N, 1 \le k \le M. \tag{10.8}$$

A especificação do modelo pode ser descrita através da notação abreviada,

$$\lambda = (A, B, \pi). \tag{10.9}$$

10.3 Geração de sequências de símbolos

Dado um modelo λ definido por *A, B,* e π, pode-se gerar uma sequência com *T* observações $O=\{o_1, o_2, ...o_T\}$, em que cada observação o_t é um símbolo de *V*, através das seguintes etapas:

1) Faz-se $t=1$ e escolhe-se um estado inicial $q_1=S_i$ através da distribuição de probabilidades inicial π;

2) Gera-se uma observação o_t em função da distribuição de probabilidades de símbolos do estado S_i, isto é $o_t=V_k$, com probabilidade $b_i(k)$ definida pela matriz *B*;

3) Transita-se para novo estado $q_t=S_j$ de acordo com a distribuição de probabilidades de transição entre estados definida pela matriz *A*;

4) Se $t > T$ a sequência está gerada. Caso contrário incrementa-se *t* e retorna-se ao ponto 2;

10.4 Os três problemas básicos dos HMM

Para a aplicação dos modelos de Markov não-observáveis em reconhecimento, existem três problemas a serem resolvidos:

1) Determinação da probabilidade de uma sequência de observações: Dada uma sequência de *T* observações $O=\{o_1, o_2, ...o_T\}$ e o modelo λ, qual a probabilidade, $P(O|\lambda)$, de esta sequência ter sido gerada pelo modelo?

2) Determinação da sequência de estados: Dada uma sequência de *T* observações $O=\{o_1, o_2, ...o_T\}$ e o modelo λ, qual a sequência de estados $Q=\{q_1, q_2, ..., q_T\}$ mais provável ?

3) Estimação de parâmetros: Dada uma sequência (ou conjunto de sequências) de observações $O=\{o_1\ o_2, \ldots o_T\}$, de que forma se ajusta os parâmetros do modelo $\lambda=(A,B,\pi)$ de modo a maximizar a probabilidade da sequência dado o modelo, $P(O|\lambda)$?

10.5 Determinação da probabilidade de uma sequência de observações

O cálculo da probabilidade de uma sequência O dado o modelo, é utilizada no reconhecimento. Por exemplo, no reconhecimento de fonemas, deverá existir um modelo λ_i que represente cada fonema. Para uma sequência de observações, é dado como reconhecido o fonema correspondente ao modelo com maior probabilidade $P(O|\lambda_i)$, utilizando o método de classificação do máximo *a posteriori*.

Para o cálculo desta probabilidade repare-se que, assumindo conhecida a sequência de estados $Q=\{q_1, q_2,.., q_T\}$, a probabilidade da sequência de observações ter sido gerada pelo modelo é dada por,

$$P(O \mid Q,\lambda) = \prod_{t=1}^{T} P(O \mid q_t,\lambda) = b_{q1}(o_1) b_{q2}(o_2) \ldots b_{qT}(o_T), \quad (10.10)$$

e por outro lado, a probabilidade da sequência de estados O dado o modelo é:

$$P(Q \mid \lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots a_{qT-1qT}. \quad (10.11)$$

A probabilidade conjunta da sequência de observações e da sequência de estados dado o modelo resulta do produto dos dois termos anteriores,

$$P(O,Q \mid \lambda) = P(O \mid Q,\lambda) P(Q \mid \lambda). \quad (10.12)$$

Finalmente, a probabilidade da sequência de observações dado o modelo resulta da soma, para todas as sequências de estados possíveis, desta probabilidade conjunta:

$$P(O|\lambda) = \sum_{\text{todos os } Q} P(O,Q|\lambda) . \qquad (10.13)$$

O cálculo de $P(O|\lambda)$ através da equação 10.13 é extremamente pesada computacionalmente, envolvendo um número $(2T\text{-}1)N^T$ multiplicações e N^T-1 adições. Mesmo para apenas três estados e 10 observações por estado, este valor é de 1180979. Felizmente é possível calcular esta probabilidade de um modo eficiente, através de um processo recursivo [Baum(66)], a que se dá o nome de algoritmo progressivo-regressivo (*forward-backward procedure*). Considerando a variável progressiva $\alpha_t(i)$ definida como a probabilidade de observação parcial da sequência $\{o_1, o_2, \ldots o_t\}$ até ao instante t, conjuntamente com a ocorrência do estado S_i no instante t, dado o modelo,

$$\alpha_t(i) = P(o_1 o_2 \ldots . o_t, q_t = S_i | \lambda) . \qquad (10.14)$$

Esta variável pode ser calculada recursivamente através de:

1) Inicialização:

$$\alpha_1(i) = \pi_i b_i(o_1) \quad 1 \le i \le N . \qquad (10.15)$$

2) Recursão:

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(\mathrm{i}) \mathrm{a}_{\mathrm{ij}} \quad 1 \le t \le T-1, 1 \le j \le N . \qquad (10.16)$$

A probabilidade da sequência de observações é dada pela soma da variável progressiva para todos os estados S_i no instante final T,

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) . \qquad (10.17)$$

O cálculo de $P(O|\lambda)$ utilizando este método recursivo necessita apenas de $N(N+1)(T-1)+N$ multiplicações e $N(N-1)(T-1)$ adições, o que para $N=3$ e $T=10$ perfaz 165 operações, contra as 1180979 necessárias para o cálculo através da equação 10.13.

Pode-se também considerar uma variável regressiva (*backward*) $\beta_t(i)$, que representa a probabilidade de ocorrência da sequência parcial de observações entre $t+1$ e T, $\{o_{t+1}o_{t+2},...o_T\}$, dado o modelo e dado que ocorreu o estado S_i no instante t

$$\beta_t(i) = P(o_{t+1}o_{t+2}...o_T \mid q_t = S_i, \lambda). \tag{10.18}$$

Esta variável pode ser calculada recursivamente através de:

1) Inicialização

$$\beta_T(i) = 1 \quad 1 \le i \le N. \tag{10.19}$$

2) Recursão

$$\beta_t(i) = \sum_{i=1}^{N} \beta_{t+1}(j)\mathrm{a_{ij}}b_j(o_{t+1}) \quad \mathrm{t = T - 1}, T-2....1,\ 1 \le j \le N. \tag{10.20}$$

e a probabilidade $P(O|\lambda)$ é dada por:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_1(i)\pi_i. \tag{10.21}$$

Repare-se que, a probabilidade $P(O|\lambda)$ em qualquer instante t, pode ser também calculada com ambas as variáveis progressiva e regressiva, através de:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_t(i)\alpha_t(i). \tag{10.22}$$

10.6 Determinação da sequência de estados

Este problema prende-se com a determinação da sequência de estados correspondente a uma dada sequência de observações. Tal como já referido, uma mesma sequência de observações pode ter sido gerada por diferentes sequências de estados. Assim, a determinação da sequência de estados correspondente a uma sequência de observações terá que obedecer a um determinado critério. Critérios diferentes conduzirão em geral a soluções diferentes.

Um dos critérios possível é escolher em cada instante *t* o estado com maior probabilidade. A probabilidade do estado S_i estar ocupado no instante *t* é dada por:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}, \qquad (10.23)$$

sendo a melhor sequência de estados utilizando este critério dada por:

$$q_t = \arg\max_{1\le i\le N}\left(\gamma_t(i)\right) \qquad 1\le t\le T. \qquad (10.24)$$

Embora este método maximize o número de estados com maior probabilidade em cada instante, pode gerar uma sequência de estados não válida, bastando para isso que a probabilidade de transição entre dois estados seja zero.

Uma outra solução é escolher a sequência de estados que gera a sequência de observações em causa com maior probabilidade, $P(Q|O,\lambda)$, que é equivalente a maximizar $P(Q,O|\lambda)$. Esta maximização é realizada de forma eficiente pelo algoritmo de Viterbi:

1) Inicialização:

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N\,, \tag{10.25a}$$

$$\psi_1(i) = 0. \quad 1 \le i \le N\,. \tag{10.25b}$$

2) Recursão:

$$\delta_t(j) = \max_{1 \le i \le N}\left(b_j(o_t)\delta_{t-1}(i)a_{ij}\right) \quad 2 \le t \le T, 1 \le j \le N\,, \tag{10.26a}$$

$$\psi_t(i) = \arg\max_{1 \le i \le N}\left(\delta_{t-1}(i)a_{ij}\right) \quad 2 \le t \le T, 1 \le j \le N\,. \tag{10.26b}$$

3) Terminação:

$$P^* = \max_{1 \le i \le N}\left(\delta_t(i)\right)\,, \tag{10.27a}$$

$$q^*_T = \arg\max_{1 \le i \le N}\left(\delta_t(i)\right)\,. \tag{10.27b}$$

4) Escolha da melhor sequência:

$$q^*_t = \psi_{t+1} q^*_{t+1} \quad t = T\text{-}1, T\text{-}2, \ldots\ldots, 1\,. \tag{10.28}$$

10.7 Estimação de parâmetros

A determinação dos parâmetros do modelo, de forma a maximizar a probabilidade $P(O|\lambda)$, não tem uma solução óptima conhecida. A solução mais utilizada envolve a criação de um modelo inicial (por exemplo de um modo aleatório) e um método de reestimação iterativo, em que cada novo modelo gera a sequência de observações, com maior probabilidade que o modelo anterior. Utilizando o conceito de frequência de ocorrência, o novo modelo $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$ é calculado a partir de (reestimação de Baum-Welch):

$$\overline{\pi}_i = \frac{\textit{número de vezes no estado } S_i \textit{ no instante } t=1}{\textit{número total de ocupações no instate } t=1}, \qquad (10.29a)$$

$$\overline{\alpha}_{ij} = \frac{\textit{número de transições do estado } S_i \textit{ para o estado } S_j}{\textit{número total de transições do estado } S_i}, \qquad (10.29b)$$

$$\overline{\beta}_i(k) = \frac{\textit{número de vezes no estado } S_j \textit{ se observou } v_k}{\textit{número total de vezes no estado } S_j}, \qquad (10.29c)$$

sendo os valores do lado direito destas equações calculadas a partir do modelo presente λ. Foi provado por Baum que este procedimento melhora a probabilidade de observação da sequência, ou seja:

$$P(O \mid \overline{\lambda}) \geq P(O \mid \lambda). \qquad (10.30)$$

A reestimação é efectuada até ser atingido um determinado critério de paragem, *e.g.*, não existam melhorias consideráveis entre duas iterações. De forma a concretizar as equações 10.29(a-c), define-se a variável intermédia $\xi_t(i,j)$, como a probabilidade conjunta de ocupar o estado S_i no instante t e ocupar o estado S_j no instante t+1:

$$\begin{aligned} \xi_t(i,j) &= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)} \\ &= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \end{aligned} . \qquad (10.31)$$

A probabilidade de ocupar o estado S_i no instante t dado pela equação 10.23 pode ser calculada utilizando 10.31, somando $\xi_t(i,j)$ para todos os j:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \qquad (10.32)$$

As equações de reestimação podem então ser rescritas utilizando estas variáveis auxiliares:

$$\overline{\pi}_i = \gamma_1(i), \tag{10.33a}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{10.33b}$$

$$\overline{b}_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{10.33c}$$

10.8 Aplicação dos HMM em reconhecimento.

No reconhecimento baseado em HMM, existem modelos probabilísticos das entidades do vocabulário a reconhecer. O reconhecimento é efectuado determinando a probabilidade da entidade a reconhecer sido gerada por cada um dos modelos.

Para a construção de um reconhecedor de sinais de fala utilizando HMM, deve-se inicialmente construir um conjunto de modelos, um para cada classe de sons (fonemas, palavras, etc.) a reconhecer, através dos seguintes passos que constituem a fase de treino:

1) definir o conjunto de classes de sons a reconhecer que corresponderá ao número *L* de modelos a treinar;

2) escolher uma topologia (o tipo de modelo, o número de estados e o número de observações por estado);

3) obter, para cada classe, um conjunto com dimensão razoável de dados de treino;

4) treinar os modelos utilizando, por exemplo, a reestimação de Baum-Welch;

Para o reconhecimento de um som, começa-se por extrair a sequência de observações correspondente ao sinal de fala. Seguidamente é calculada a probabilidade da sequência de observações, dado cada um dos modelos. Atribuí-se à sequência de observações a som (classe) associado ao modelo que obteve a máxima probabilidade.

$$P(O \mid \lambda_j) = \max P(O \mid \lambda_i) \quad 1 \le i \le L \,. \tag{10.34}$$

O esquema de blocos do reconhecedor utilizando estes modelos é apresentado na figura 10.2.

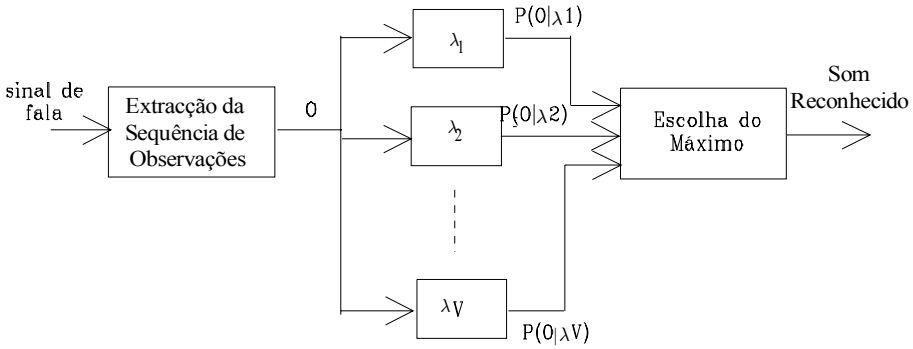

Figura 10.2
Esquema de blocos de um reconhecedor de entidades isoladas.

As características do sinal de entrada que servem como observações, obtidos trama-a-trama, são normalmente parâmetros espectrais derivados de LPC tais como os *cepstrum*, a energia, e as respectivas variações em relação à trama anterior (delta *cepstrum* e delta energia). Sendo estes valores contínuos, é necessário proceder à sua quantificação vectorial, tornando-os um conjunto finito de

símbolos, resultando numa degradação da percentagem de reconhecimento, a menos que se utilize um livro de código bastante grande. Outra solução para este problema é a utilização de modelos contínuos, onde as distribuições associadas às observações são caracterizadas por uma mistura de funções densidade de probabilidade, normalmente com distribuição gaussiana:

$$b_j(O) = \sum_{m=1}^{M} c_{jm} \aleph(O, \mu_{jm}, U_{jm}) \quad 1 \le j \le N , \qquad (10.35)$$

em que O é vector a ser modelado, c_{jm} é o peso ou coeficiente da m-ésima mistura no estado S_j. e $\aleph(c_{jm}, \mu_{jm}, U_{jm})$, com média μ_{jm} e covariância U_{jm}. A função densidade de probabilidade da equação 10.34 pode ser usada, desde que com o número suficiente de misturas, para aproximar qualquer função contínua. A equação da reestimação de $b_j(k)$ dada pela equação 10.33c desdobra-se nas equações para reestimar c_{jm}, μ_{jm} e U_{jm} [Rabiner (85)].

Nas aplicações dos HMM para o reconhecimento de fala, não se usa normalmente modelos ergódicos (completamente ligados) mas sim modelos esquerda-direita, ou seja, modelos em que de um estado S_i só é possível transitar para o estado S_{i+1}, ou permanecer no mesmo estado, como mostra a figura 10.3.

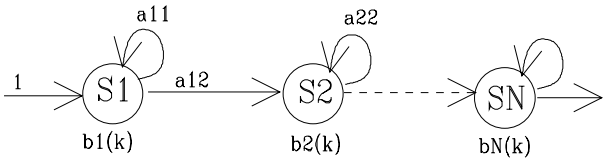

Figura 10.3
Modelo Esquerda-Direita normalmente utilizado
nas aplicações de reconhecimento de fala

Esta topologia traz algumas simplificações na estrutura dos parâmetros do modelo, que se explicitam seguidamente.

1) A distribuição de estados inicial π tem apenas um valor não nulo (igual a 1), correspondente ao estado S_1

2) A matriz de distribuição das probabilidades de transição entre estados, tem, por cada linha, apenas dois valores não nulos. Os valores correspondentes a a_{ii} e a $a_{i(i+1)}$

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{10.35}$$

sendo o último estado um estado absorvente, só podendo transitar para si mesmo.