Aprendizagem Espectral em Modelos Ocultos de Markov

Jonas Rocha Lima Amaro

Instituto de Matemática Pura e Aplicada

Resumo

Modelos Ocultos de Markov são modelos versáteis para representar processos estocásticos e com isso, muitas perguntas podem ser feitas a cerca do processo partindo dos parâmetros do modelo. Neste trabalho, o leitor será introduzido à abordagem discreta de Modelos Ocultos de Markov e também à uma forma eficiente de resolver dois problemas: Calcular probabilidades de sequências; e calcular probabilidade da próxima observação dada a sequência de observações que o antecede. Finalmente, apresentaremos estudos comparativos que trazem em média uma redução de XX& no tempo a execução de rotinas se comparada à implementação mais comum.

Introdução

Cadeia de Markov

Cadeia de Markov é um processo estocástico no qual a probabilidade do próximo estado depende apenas do estado anterior. Se o processo ainda tiver uma quantidade finita de estados e o tempo for discreto, é possível representar a cadeia como grafo direcionado completo como representado na Figura 1. Cada estado corresponde a um nó do grafo, partindo de cada nó o estado seguinte é uma variável aleatória com as probabilidades representadas nos pesos de cada aresta. Por tanto, no caso ilustrado na Figura 1

$$\text{Para todo } e, e' \in E = \{A, B, C\}$$
$$P(e'|e) \in [0, 1] \text{ e}$$
$$P(A|e) + P(B|e) + P(C|e) = 1$$

Como se espera de uma função probabilidade com domínio discreto. A partir do grafo é bastante direto representar os parâmetros de forma matricial

$$T = \begin{pmatrix} P(A|A) & P(B|A) & P(C|A) \\ P(A|B) & P(B|B) & P(C|B) \\ P(A|C) & P(B|C) & P(C|C) \end{pmatrix}$$

Por outro lado, há redundâncias na matriz, pois cada linha soma 1, por isso é suficiente 6 parâmetros para determinar uma cadeia de 3 estados. Além disso, existe uma distribuição de probabilidade para o valor inicial da sequência que será representado como $\pi_0(e)$ que é equivalente ao vetor

$$\Pi_0 = \begin{pmatrix} \pi_0(A) \\ \pi_0(B) \\ \pi_0(C) \end{pmatrix}$$

De forma geral, em uma *Cadeia de Markov* com n estados, são necessários $n^2 - n$ parâmetros e

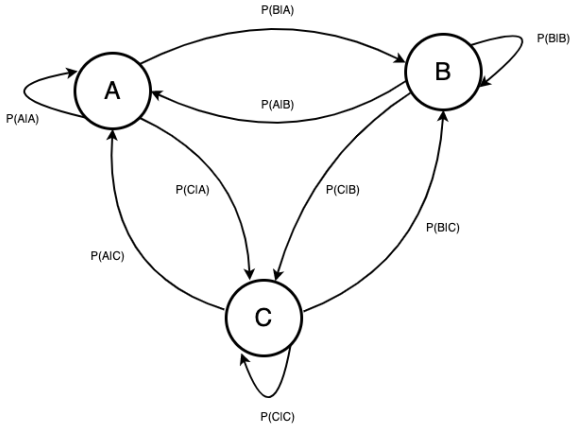

Figura 1: *Exemplo de diagrama de estados de uma Cadeia de Markov com* 3

cada probabilidade condicional é representada numa matriz $T \in \mathbb{R}^{n \times n}$ com entradas não negativas e que

$$T\mathbf{1}_n = \mathbf{1}_n$$

Isto é, cada linha soma 1. Além disso, também vetor $\Pi_0 \in \mathbb{R}^n$ com a distribuição de probabilidades para o estado inicial que também somam 1:

$$\Pi_0^{\mathrm{T}}\mathbf{1}_n = 1$$

E da mesma forma que a matriz transição, à redundância no vetor Π_0, pois é suficiente determinar apenas as probabilidades de $n - 1$ estados.

Concluí-se que uma *Cadeia de Markov* de n estados é definida por $n^2 - 1$ parâmetros.

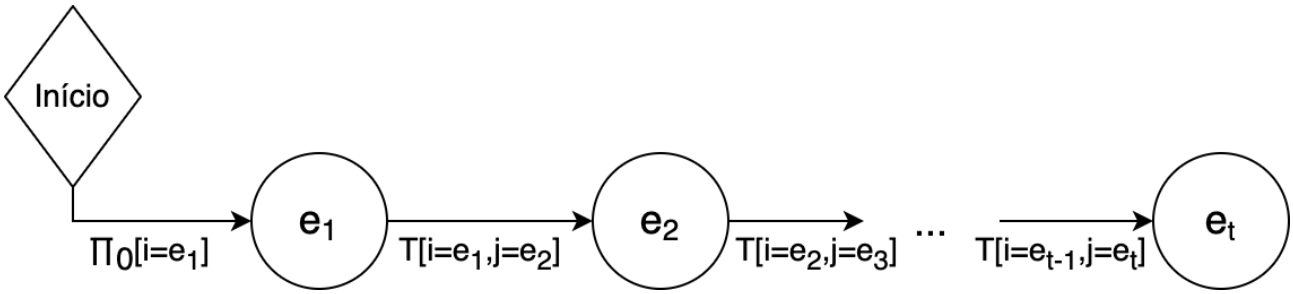

Figura 2: *Na visualização sequencial da Cadeia de Markov de o primeiro estado é escolhido de acordo com a probabilidade correspondente da coluna i no vetor Π_0 e cada estado seguinte é escolhido com a probabilidade presente na coluna j e linha i que corresponde ao estado atual da matriz transição T*

Estimativa de parâmetros

Considere o seguinte conjunto sequências

$$ABCABACCACAB$$
$$BACABACAAC$$
$$BBABAAAACCACABABABAC$$

Assumindo a premissa de que tais sequências foram geradas por um processo modelável como *Cadeia de Markov*, os parâmetros de tal cadeia serão estimados. É urgente se observar que as sequências apresentam apenas 3 estados. Por tanto, a matriz transição $T\mathbb{R}^{3\times 3}$ e $\Pi_0 \in \mathbb{R}^3$.

No total, há 20 ocorrências da letra A seguida de algum estado, 10 da B, e 9 da C. Ao se contar cada ocorrência das tuplas AA, AB, AC, e assim por diante, dividindo pela ocorrência do respectivo estado antecessor. Assim a seguinte matriz transição é obtida

$$T = \begin{pmatrix} \frac{4}{20} & \frac{8}{20} & \frac{8}{20} \\ \frac{8}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{7}{9} & \frac{0}{9} & \frac{2}{9} \end{pmatrix}$$

Já para a estimativa do vetor probabilidade de estado inicial Π_0, o processo é imediato, é a frequência pela quantidade de sequências

$$\Pi_0 = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix}$$

Probabilidade de uma sequência

O interesse nesta subseção é de calcular, partindo dos parâmetros que descrevem uma *Cadeia de Markov*, a probabilidade de cada sequência de estados. Seja $(e_i)_{i=1}^t$ um processo de Markov que gere uma sequência de estados de uma *Cadeia de Markov* que tenha probabilidade inicial Π_0 e uma matriz transferência de estados T, como ilustrado na figura 2. Como Bishop [1] apresenta, a transferência de estado só depende do estado anterior, que simplifica bastante o cálculo da probabilidade conjunta a partir das probabilidades condicionais

$$P((e_k)_{k=1}^t) = P(e_1)\prod_{k=2}^{t} P(e_k|e_{k-1})$$
$$\Rightarrow P((e_k)_{k=1}^t) = \Pi_0[i=e_1]\prod_{k=2}^{t} T[i=e_{k-1},\, j=e_k]$$

No caso do modelo da seção anterior, a sequência $BACABBAC$ tem probabilidade

$$P(BACABBAC) = P(B)P(A|B)P(C|A)P(A|C)$$
$$P(B|A)P(B|B)P(A|B)P(C|A)$$
$$= 2/3 * 8/10 * 8/20 * 7/9 * 8/20 * 1/10 * 8/10 * 8/20$$
$$\approx 2.212 * 10^{-3}$$

O leitor mais atento observou que a medida que as sequências crescem elas se tornam mais improváveis, por isso deve haver uma precaução na manipulação dessas probabilidades e evitar comparar sequências de tamanhos diferentes.

Modelo Oculto de Markov

Imagine uma sequência $(e_i)_{i=1}^t$ gerada por processo estocástico à moda da *Cadeia de Markov* como já vimos na seção anterior, adicionando um elemento: os estados não explícitos nas sequências, cada estado emite uma símbolo observável o_i, com o detalhe que a cada estado tem a sua própria distribuição de probabilidade para esses símbolos, como ilustrada na figura. Chama-se de *Modelo Oculto de Markov* o processo estocástico que gera tais sequências de elementos observáveis.

Perceba que a *Cadeia de Markov* pode ser interpretada como um *Modelo Oculto de Markov* degenerado, no qual o espaço dos símbolos observáveis é o próprio espaço dos estados e cada estado emite com probabilidade 1 elemento que o representa.

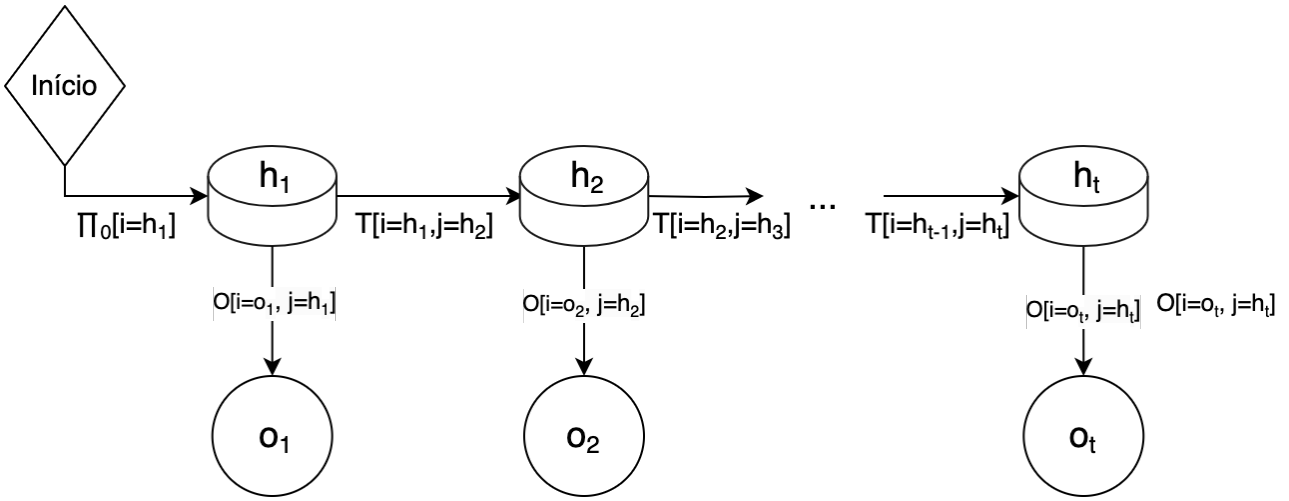

Figura 3: *Na visualização sequencial de um Modelo Oculto de Markov, o primeiro estado oculto é sorteado com a probabilidade presente na i-ésima coluna do vetor Π_0. Em seguida é sorteado o primeiro símbolo observável com probabilidade da coluna j que corresponde o estado sorteado h_1 e linha i correspondente ao símbolo observável o_1 na matriz O. O próximo estado é selecionado com probabilidade com a linha i correspondente à h_1 e coluna $j = h_2$. O padrão se repete até um período t*

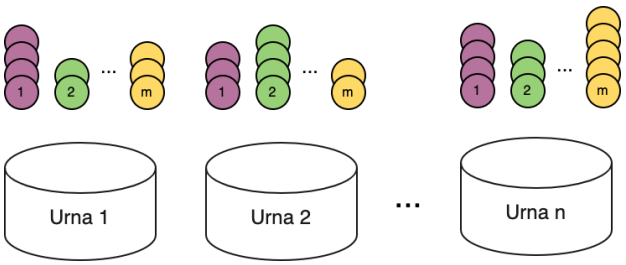

Figura 4: *Os elementos do sorteio são n urnas com bolas de m cores diferentes organizadas de forma que cada urna tem a sua proporção de bolas em cada cor*

Modelo das Urnas e Bolas

Vamos ilustrar o *Modelo Oculto de Markov* com um exemplo que leva elementos apresentado por Jack Ferguson. Imagine um sorteio seriado organizado com n urnas cada uma com sua própria distribuição de bolas com m cores diferentes e um sorteador, como ilustrado na figura 4.

Este sorteador irá escolher secretamente uma urna e a partir desta urna irá sortear com reposição uma bola e somente esta bola será revelada. No próximo sorteio a escolha da próxima urna dependerá da primeira, como é feito nas *Cadeias de Markov*, e desta urna será apresentada a segunda bola. Assim se repete o processo.

Note que se o caso fosse de apenas dois estados fossem observados como uma cara ou coroa, da mesma forma que Rabiner (1989) [2] ilustra, um modelo contendo apenas um estado que entregasse a mesma distribuição das observações entre cara e coroa, seria mais informativo que um modelo com complexas relações entre urnas e bolas.

Parâmetros do Modelo

Além de todos os elementos já presentes na *Cadeia de Markov*:

- quantidade de estados h_i da cadeia $n \in \mathbb{N}$;
- distribuição de probabilidade inicial $\Pi_0 \in \mathbb{R}^n$;
- matriz probabilidade de transferência $T \in \mathbb{R}^{n \times n}$

são adicionados os elementos referentes aos símbolos visíveis

- quantidade de símbolos observáveis o_i emitido pelo processo $m \in \mathbb{N}$
- matriz distribuição de probabilidade dos símbolos observáveis em cada estado, ou matriz emissão $O \in \mathbb{R}^{m \times n}$

A iteração desses elementos estão representados na figura 3 que ilustra a operação sequencial de um processo modelado por um *Modelo Oculto de Markov*

Apesar de um dado processo que se comporte como um *Modelo Oculto de Markov* esteja bem descrito com esses elementos, mais a frente será apresentada uma representação alternativa que será capaz de responder as principais perguntas a serem feitas sobre um dado processo que só se sabe das sequências de símbolos observáveis.

Principais perguntas ao Modelo

Dentre as possíveis perguntas que podem ser feitas a cerca do funcionamento do modelo, serão destacadas 3 perguntas essenciais que Rabiner (1989) [2]

1. Dada uma sequência de observações $(x_i)_{i=1}^t$ e um modelo (Π_0, T, O) qual é a probabilidade do modelo emitir a sequência

2. Dada uma sequência de observações $(x_i)_{i=1}^t$ e um modelo (Π_0, T, O) qual é a sequência de estados ocultos $(y_i)_{i=1}^t$ que melhor descreve as observações feitas
3. Como ajustar os parâmetros do modelo (Π_0, T, O) para maximizar a probabilidade de emitir as sequências de símbolos observados.

A primeira pergunta já teve uma investigação inicial para o caso da *Cadeia de Markov*. É importante avaliar essas probabilidades para se fazer comparações entre modelos. Em contraste ao caso da cadeia, a primeira pergunta não é trivialmente calculado de forma eficiente para os modelos ocultos. O complicador é que se o modelo tiver m possíveis estados e t termos, haverá m^t possíveis sequências de estados ocultos a serem avaliados para se calcular tal probabilidade. Um algoritmo muito mais eficiente para essa estimativa é através do Procedimento *Backward-Forward* de Baum (1967) [3]. Rabiner estima que a complexidade desse procedimento é de $O(n^2t)$. Também será exposto na seção seguinte uma solução baseada na características espectrais do modelo como apresentado por Hsu, et al. (2012) [4]

Já a segunda pergunta, pode ser resolvida pelo *Algoritmo de Viterbi* [5] que escolhe o estado seguinte que maximiza o $P((h_i)_{i=1}^t | (o_i)_{i=1}^t, ((\Pi_0, T, O)))$. A complexidade computacional do algoritmo também é $O(n^2t)$, como estima C. Zhu[6]. Essa análise permite avaliar as estruturas do modelo. Quando o modelo está sendo treinado, os parâmetros a serem aprendidos são criados sem significado explícito, esta avaliação nos permite entender que estados ocultos capturam que aspectos do modelo. É importante destacar que, apesar da popularidade de *Viterbi* não existe uma métrica única para avaliar essa resposta, o próprio Rabiner [2] ilustra outro critério, escolher a sequência de estados ocultos os quais cada estado é individualmente o mais provável dado o tempo, porém o autor também aponta que é possível que tal abordagem possa chegar em sequências possuem estados ocultos consecutivos com probabilidade de transferência nula entre eles.

A terceira, finalmente, está vinculada aos métodos iterativos de otimização, tal como *Baum-Weich*[7], que o próprio Rabiner expõe em seu tutorial[2]. Em contraste, a estratégia de aprendizagem espectral de Hsu, não depende deste problema, pois para ser treinada, será necessário medir frequências relativas aos três primeiros símbolos observáveis.

Apesar de não citado por Rabiner, também destaco que modelos estimados devem ser capazes de gerar sequências semelhantes ao que gerador, que geralmente é chamado na literatura de *Random Walk*. Note que este problema não é trivial se existirem sequências proibidas pela estrutura geradora

Aprendizagem Espectral

No trabalho *A Spectral Algorithm for Learning Hidden Markov Models*, D. Hsu, S. M. Kakade e T. Zhang apresentam uma abordagem de representação e estimativa de *Modelos Ocultos de Markov*. Serão apresentadas nessa as principais ideias e notações da abordagem deles.

Já nas discussões preliminares os autores apresentam a *Condição 1* para o funcionamento do algoritmo: $\Pi_0 > 0$ e $rank(O) = rank(T) = n$. Segundo os autores, essa condição serve para evitar que distribuição de um dos estados ocultos possa ser gerada pela combinação convexa de outros estados ocultos.

Apresentação do Modelo

Dos problemas apresentados na seção anterior, este modelo é capaz de resolver todos menos problema da maximização da probabilidade já foi discutido na seção anterior

Sejam $P_1 \in \mathbb{R}^m$, $P_{2,1} \in \mathbb{R}^{m \times m}$ e $P_{3,x,1} \in \mathbb{R}^{m \times m}$ no qual x é um elemento do conjunto dos símbolos observáveis. Seguiremos usando m como o número de símbolos observáveis. Define-se cada entrada do vetor e das matrizes como

$$P_1[i] = Pr(o_1 = i)$$
$$P_{2,1}[i,j] = Pr(o_2 = i, o_1 = j)$$
$$P_{3,x,1}[i,j] = Pr(o_3 = i, o_2 = x, o_1 = j)$$

No qual Pr é a frequência conjunta na qual cada símbolo observável é emitido. Note que só são necessárias as três primeiras entradas da sequência para esta definição.

Além disso, defini-se a matriz $U \in \mathbb{R}^{m \times n}$ como a matriz tal que $U^{\mathrm{T}}O$ é invertível, em seguida os autores sugerem um candidato natural: a matriz esquerda de vetores singulares reduzida de $P_{2,1}$, convido o leitor a consultar a demonstração do *Lema 2* que prova este fato.

Dadas as definições, é possível apresentar a nova representação do modelo

$$\vec{b}_1 = U^{\mathrm{T}} P_1$$
$$\vec{b}_\infty = (P_{2,1}^{\mathrm{T}} U)^+ P_1$$
$$B_x = (U^{\mathrm{T}} P_{3,x,1})(U^{\mathrm{T}} P_{2,1})^+$$

Distribuição de Probabilidade Conjunta Finalmente, a distribuição a ser estimada por essa representação é dada por

$$Pr((o_i)_{i=1}^t) = \vec{b}_\infty^{\mathrm{T}} \prod_{i=1}^{t} B_{o_{t+1-i}} \vec{b}_1$$

A demonstração da validade dessa relação é provada no *Lema 3*.

Demais problemas O *Lema 4* apresenta uma forma de calcular a recorrência de $\vec{b}_t$

$$\vec{b}_{t+1} = \frac{B_{o_t}\vec{b}_t}{\vec{b}_\infty^{\mathrm{T}} B_{o_t}\vec{b}_t}$$

Com esta identidade é possível calcular a relação de $\vec{b}_t$ e as distribuições de probabilidade do estado oculto, dada a sequência observada

$$\vec{b}_t = (U^{\mathrm{T}} Pr(h_t|(o_i)_{i=1}^{t-1})$$

O problema do *Random Walk* pode ser reduzido ao problema da probabilidade condicional, isto é, dada uma sequência de observações, esta probabilidade é dada por

$$Pr(o_t|(o_i)_{i=1}^{t-1}) = \vec{b}_\infty^{\mathrm{T}} B_{o_t}\vec{b}_t$$

Este trabalho, se dedicou em medir principalmente o sucesso da estimativa da *Distribuição de Probabilidade Conjunta* empiricamente

Experimentos

Discussões

Além dos resultados apresentados, esta técnica corre o risco de cair em soluções com probabilidades negativas, como apontam H. Zhao e P. Poupart [8]. Disto pode-se ou tentar adicionar novas restrições para a solução do problema, impedindo que se chegue num resultado negativo, ou até progredir na direção de uma teoria de probabilidade que comporte probabilidades negativas, como A. Blass e Y. Gurevich [9] apresentam.

Referências

[1] Christopher M. Bishop. *Pattern recognition and machine learning/*. Textbook for graduates. New York: Springer, c2006. Pp. 607–610. URL: http://www.loc.gov/catdir/enhancements/fy0818/2006922522-t.html.

[2] L.R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". Em: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286. DOI: 10.1109/5.18626.

[3] Leonard E. Baum e John A. Eagon. "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". Em: *Bulletin of the American Mathematical Society* 73 (1967), pp. 360–363.

[4] Daniel Hsu, Sham M. Kakade e Tong Zhang. "A spectral algorithm for learning Hidden Markov Models". Em: *Journal of Computer and System Sciences* 78.5 (2012). JCSS Special Issue: Cloud Computing 2011, pp. 1460–1480. ISSN: 0022-0000. DOI: https://doi.org/10.1016/j.jcss.2011.12.025. URL: https://www.sciencedirect.com/science/article/pii/S0022000012000244.

[5] A. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". Em: *IEEE Transactions on Information Theory* 13.2 (1967), pp. 260–269. DOI: 10.1109/TIT.1967.1054010.

[6] Chenguang Zhu. "Chapter 2 - The basics of natural language processing". Em: *Machine Reading Comprehension*. Ed. por Chenguang Zhu. Elsevier, 2021, pp. 27–46. ISBN: 978-0-323-90118-5. DOI: https://doi.org/10.1016/B978-0-323-90118-5.00002-3. URL: https://www.sciencedirect.com/science/article/pii/B9780323901185000023.

[7] A. P. Dempster, N. M. Laird e D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". Em: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[8] H. Zhao e Pascal Poupart. "A Sober Look at Spectral Learning". Em: *ArXiv* abs/1406.4631 (2014).

[9] Andreas Blass e Yuri Gurevich. "Negative probabilities: what are they for?*". Em: *Journal of Physics A: Mathematical and Theoretical* 54.31 (ago. de 2021), p. 315303. DOI: 10.1088/1751-8121/abef4d. URL: https://dx.doi.org/10.1088/1751-8121/abef4d.