



ALÉM DA INDEPENDÊNCIA PROBABILÍSTICA ENTRE EVENTOS

LISBETH KAISERLIAN CORDANI — IME/USP

SIMONE BEGA HARNIK — BACHARELANDA EM ESTATÍSTICA NO IME/USP

No artigo *Afinal a aula é de probabilidade ou estatística?*, publicado na RPM 88, apresentamos sequências de resultados referentes a lançamentos de moeda, em uma situação de eventos independentes onde, ao lançarmos uma moeda duas vezes, o resultado do primeiro lançamento (se é cara ou coroa) não influencia em nada o resultado do segundo lançamento. Lançamentos de dados ou retiradas de bolas em urnas (com reposição) são outros exemplos considerados em independência de eventos.

O propósito de desenvolver modelos como os citados é usá-los em problemas da vida real. Mas nem sempre isso faz sentido, e talvez seja mais realista pensar que independência seja uma restrição muito forte para vários fenômenos, como os biológicos, físicos, econômicos, dentre outros. Isso porque parece plausível que o futuro dependa de certo modo do passado (seja ele remoto ou não). No capítulo introdutório de [1], há exemplos em diversos campos de aplicação.

Então, se modelos com estrutura de independência podem ser considerados como uma aproximação grosseira da realidade, vamos propor uma abordagem que inclua *dependência*, na sua versão bem simplificada. Usaremos aqui um exemplo recorrente nos textos introdutórios de *processos estocásticos*, que são situações em que o futuro (amanhã) depende somente do presente (hoje). Alguns elementos de probabilidade serão trabalhados nesse contexto, depois da apresentação do exemplo.

Uma aplicação simples, extraída de [3], ajuda a entender essa estrutura probabilística.

Suponhamos (por simplicidade) que, no verão do Rio de Janeiro, só existam duas possibilidades para o tempo na cidade: C = chuvoso ou E = ensolarado. Podemos pensar que estamos interessados no fenômeno *chuva* e redefinir as possibilidades de tempo como 1 = chuvoso e 0 = ensolarado. Se considerarmos dez dias seguidos, poderemos ter uma configuração do tipo $(C, C, E, E, C, E, C, E, E, E)$ ou, de acordo com a notação alternativa, $(1, 1, 0, 0, 1, 0, 1, 0, 0, 0)$. Podemos, ainda, usar a notação $(X_1, X_2, \dots, X_{10})$, em que X_i é o tempo no dia i ($i = 1, 2, 3, \dots, 10$), isto é, $X_i = 1$ (chuvoso) ou $X_i = 0$ (ensolarado). A sequência de dias pode ser predeterminada ou não.

Vamos, então, usar um esquema que considere o tempo de amanhã relacionado ao de hoje e introduzir alguns elementos, em termos de probabilidades. Sem entrar no mérito da escolha dos valores, que podem ser estimados por registros históricos ou por simples intuição, vamos supor, então, que, se hoje (o primeiro dia de nosso estudo) é um dia chuvoso ($X_1 = 1$), a probabilidade de amanhã ser um dia chuvoso é de 60% (0,6); e a de ser um dia ensolarado é de 40% (0,4). Voltamos a dizer que em nosso *modelo* simplificado só existem essas duas possibilidades (chuvoso ou ensolarado). A aposta na chuva para amanhã (dado que choveu hoje) é maior do que a aposta para o dia ensolarado, embora não saibamos o que *de fato* irá acontecer. Probabilidade é *prospecção*. Nessa estrutura, observa-se que $60\% + 40\% = 100\%$ (ou $0,6 + 0,4 = 1,0$), ou seja, esgotamos as possibilidades, *dado* que choveu hoje.

Esquemmatizando o que vimos para a situação descrita:

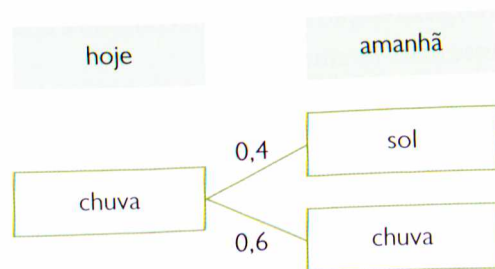


figura 1: Esquema para as probabilidades de chuva ou sol para amanhã, dado que hoje foi um dia chuvoso

Até agora nosso esquema só previa um estado para hoje, que era o chuvoso. Mas pode ser que hoje tenha sido um dia ensolarado ($X_1 = 0$). Do mesmo modo que fizemos para um dia chuvoso, vamos então supor (especulação ou motivos meteorológicos!) que há probabilidade de 80% (0,8) de que amanhã continue sendo ensolarado – teremos praia? Portanto, restam 20% de probabilidade para que amanhã seja chuvoso. Aqui, diferentemente do que especulamos na figura 1, a aposta no dia ensolarado para amanhã (dado que tivemos sol) é maior do que a aposta para chuva. Mas, de novo, não sabemos o que de fato irá acontecer. Reparem que novamente $80\% + 20\% = 100\%$, esgotando as possibilidades (dado que fez sol hoje).

Reescrevendo após essas considerações, temos a figura 2, com o esquema completo.

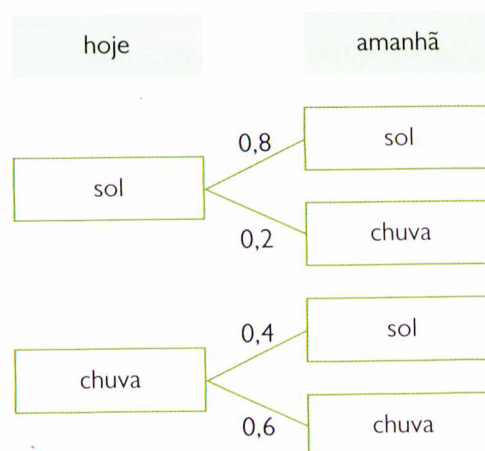


figura 2: Esquema para as probabilidades de chuva ou sol para amanhã, contemplando as duas realidades de hoje

Um conceito muito comum em situações desse tipo, dentro da teoria de probabilidades, é o conceito de Probabilidade Condicional. Os valores declarados na figura 2 só fazem sentido quando se referem ao que aconteceu hoje. Por exemplo, 80% é a probabilidade de fazer sol amanhã, *condicionada* ao fato de que hoje fez sol. Escrevemos, então, em uma notação mais formal,

Probabilidade (amanhã será ensolarado | hoje foi ensolarado) = 0,8,

em que o símbolo “|”, condicional, pode também ser lido como “dado que”. Uma notação mais compacta pode ser: $P(\text{sol amanhã} | \text{sol hoje}) = 0,8$.

Para dois eventos A e B quaisquer (do mesmo espaço amostral), usamos para a probabilidade condicional a notação genérica $P(A|B)$. É importante dizer que, se a informação de que B ocorreu não interfere na probabilidade de ocorrência de A , ou seja, se $P(A|B) = P(A)$, dizemos que os eventos A e B são independentes. Não é desse caso que estamos tratando, pois partimos da premissa de que o que ocorre amanhã depende do que ocorreu hoje, portanto estamos em uma estrutura de dependência. Mais informações sobre probabilidade condicional o leitor pode encontrar, por exemplo, em [5].

Resumindo as suposições acima com a notação de probabilidade condicional, temos:

$$\begin{aligned} P(\text{sol amanhã} | \text{sol hoje}) &= 0,8 \\ P(\text{sol amanhã} | \text{chuva hoje}) &= 0,4 \\ P(\text{chuva amanhã} | \text{sol hoje}) &= 0,2 \\ P(\text{chuva amanhã} | \text{chuva hoje}) &= 0,6 \end{aligned}$$

Podemos, com auxílio dos símbolos matemáticos, resumir ainda, tomando por

Variáveis	Estados do tempo
hoje = X_1	chuva = 1
amanhã = X_2	sol = 0

e escrevendo

$$\begin{aligned} P(X_2 = 0 | X_1 = 0) &= 0,8 & P(X_2 = 1 | X_1 = 0) &= 0,2 \\ P(X_2 = 0 | X_1 = 1) &= 0,4 & P(X_2 = 1 | X_1 = 1) &= 0,6. \end{aligned}$$

Observe que as probabilidades condicionais escritas imediatamente acima são as mesmas que as escritas sem o auxílio da linguagem matemática. O que fizemos foi apenas a substituição pelos símbolos.

A estrutura que acabamos de descrever, que considera a dependência probabilística entre eventos, pode ser estudada por meio das chamadas *cadeias de Markov* – nome que homenageia o matemático russo Andrei Andreyevich Markov. Segundo [8], no começo do século XX, Markov quis investigar a alternância de vogais e consoantes no poema “Onéguine”, de Pushkin, e elaborou um modelo de

ensaios sucessivos, em que cada um só dependia do ensaio imediatamente anterior.

Um dos casos mais simples de cadeias de Markov é o que definimos acima: uma sequência de resultados (chuva, sol, sol, chuva, sol, ...) cujas probabilidades de ocorrência do evento de amanhã dependem exclusivamente do resultado do evento de hoje; indo mais um passo, ainda nessa nossa cadeia de Markov, diríamos que as probabilidades de ocorrência do evento de depois de amanhã dependem apenas do resultado do evento de amanhã, e assim por diante. Ou seja, a distribuição de probabilidades de um momento do tempo $i + 1$ depende somente do resultado observado no tempo i . Outra maneira de dizer isso ([6]) seria que, em uma cadeia de Markov, o futuro é *condicionalmente* independente do passado, dado que conhecemos o presente.

Como já dissemos, se a estrutura fosse de independência, a probabilidade de alguma coisa (relacionada ao tempo) acontecer amanhã não dependeria absolutamente em nada do que aconteceu hoje, ou seja, $P(\text{chuva amanhã} | \text{chuva hoje})$ seria igual à $P(\text{chuva amanhã})$, não importando o que tenha acontecido hoje. Mas não é essa a interpretação que estamos dando aqui, pois estamos construindo uma estrutura de dependência *markoviana*.

Usualmente, as quatro probabilidades condicionais são organizadas em uma matriz P , chamada *matriz de transição* (ideia de mudança ao longo do tempo) da cadeia de Markov, cujas entradas são as *probabilidades de transição*. Nas linhas representamos a condição no tempo i (por exemplo, hoje), e, nas colunas, a condição no tempo $i + 1$ (por exemplo, amanhã), em relação aos estados chuvoso e ensolarado. Nesse caso, teremos

		amanhã	
		0 = sol	1 = chuva
hoje	0 = sol	0,8	0,2
	1 = chuva	0,4	0,6



ou, ainda, sem os estados do tempo escritos:

$$P = \begin{vmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{vmatrix}$$

Como se pode observar, essa matriz tem algumas propriedades interessantes: nenhuma de suas entradas pode ser um número negativo (pois probabilidades são sempre positivas ou nulas!) e a soma de cada linha é sempre 1 (corresponde aos 100% que discutimos anteriormente). Interpretando essa matriz, temos que cada linha representa o estado no tempo i (ou seja, hoje foi chuvoso ou ensolarado). Se hoje foi ensolarado, observamos a primeira linha da matriz:

$$\begin{array}{cc} & \begin{array}{cc} 0 = \text{sol} & 1 = \text{chuva} \end{array} \\ \begin{array}{c} 0 = \text{sol} \end{array} & \begin{vmatrix} 0,8 & 0,2 \end{vmatrix} \end{array}$$

Ou seja: a probabilidade de ser ensolarado novamente amanhã é de 0,8. Já a de chover, é o complementar: 0,2. Por outro lado, se hoje choveu, a probabilidade de chover amanhã está na segunda linha e na segunda coluna da matriz (0,6); e a probabilidade de não chover consta da segunda linha e da primeira coluna (0,4).

É comum representar uma cadeia de Markov por meio de um diagrama de estados (figura 3), e o leitor vai perceber as relações presentes na matriz de transição P .

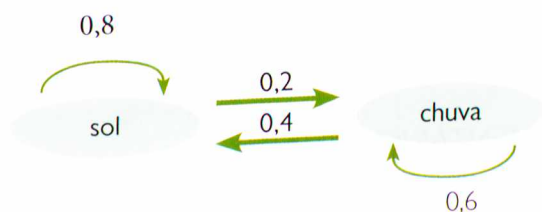


figura 3: Diagrama de estados do exemplo para o tempo (chuva ou sol)

Genericamente, podemos escrever:

$$P = \begin{vmatrix} p & 1-p \\ m & 1-m \end{vmatrix}$$

onde p e m são números não negativos. O usual para as entradas da matriz de transição P é a notação p_{ij} , o que faz a matriz $P_{2 \times 2}$ (como a do exemplo dado) ter a expressão a seguir,

$$P = \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}$$

com as seguintes características:

- i) P é quadrada.
- ii) Todas as entradas p_{ij} são não negativas $i, j = 1, 2$
- iii) A soma dos elementos de cada linha é igual a 1.

É claro que o estudo das cadeias de Markov não para por aqui. A matriz de transição, por exemplo, já foi pesquisada pelos matemáticos e probabilistas em profundidade e traz muitas informações sobre o fenômeno estudado. Sob determinadas condições, a matriz de transição P , dada anteriormente, permite estabelecer probabilidades para a condição do tempo, por exemplo, dois dias à frente.

Como prever **hoje** a condição do tempo em termos probabilísticos para depois de amanhã? Vamos imaginar que hoje está **sol** e vamos fazer o diagrama incluindo **amanhã** e **depois de amanhã** (figura 4 na próxima página).

As probabilidades do evento **sol depois de amanhã** (dado que fez sol hoje) seriam obtidas pelos casos: **sol amanhã** e **sol depois de amanhã** ($0,8 \times 0,8$) ou **chuva amanhã** e **sol depois de amanhã** ($0,2 \times 0,4$). Isso nos remete ao cálculo da probabilidade da união de dois eventos disjuntos (ver []), ou seja,

$$\begin{aligned} P(\text{sol depois de amanhã} \mid \text{sol hoje}) &= \\ 0,8 \times 0,8 + 0,2 \times 0,4 &= 0,64 + 0,08 = 0,72. \end{aligned}$$

Agora, para calcularmos as probabilidades do evento **chuva depois de amanhã** (dado que fez sol hoje), precisamos verificar os caminhos **sol amanhã** e **chuva depois de amanhã** ($0,8 \times 0,2$) ou **chuva amanhã** e **chuva depois de amanhã** ($0,2 \times 0,6$). Temos então $P(\text{chuva depois de amanhã} \mid \text{sol hoje}) =$

$$0,8 \times 0,2 + 0,2 \times 0,6 = 0,16 + 0,12 = 0,28.$$

Nossa sugestão é mostrar a origem dos cálculos aos alunos pelo diagrama de árvore, para facilitar a compreensão.

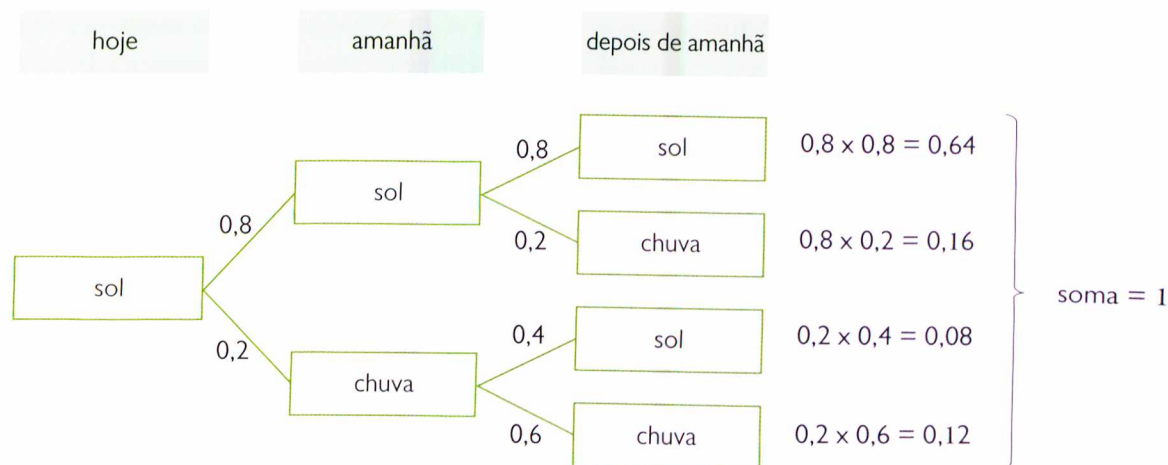


figura 4: Esquema para as probabilidades de chuva ou sol para depois de amanhã, dado que fez sol hoje

É possível construir, de modo análogo, o outro caso, ou seja, caso tenha sido **chuva hoje**, como seriam as probabilidades tanto de sol como de chuva para depois de amanhã? O leitor pode construir o diagrama de árvore e verificar que

$$P(\text{sol depois de amanhã} | \text{chuva hoje}) = 0,4 \times 0,8 + 0,6 \times 0,4 = 0,32 + 0,24 = 0,56, \text{ e}$$

$$P(\text{chuva depois de amanhã} | \text{chuva hoje}) = 0,4 \times 0,2 + 0,6 \times 0,6 = 0,08 + 0,36 = 0,44.$$

Para terminar, vamos associar os resultados obtidos a operações com a matriz de transição P , lembrando que essa matriz tem como elementos as probabilidades de transição de um estado (sol, chuva) para outro (sol, chuva). Vamos reescrever a matriz P para os dados do problema do tempo e vamos propor encontrar $P \times P = P^2$. Assim:

$$P^2 = \begin{bmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{bmatrix} \times \begin{bmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{bmatrix} = \begin{bmatrix} 0,8 \times 0,8 + 0,2 \times 0,4 = 0,72 & 0,8 \times 0,2 + 0,2 \times 0,6 = 0,28 \\ 0,4 \times 0,8 + 0,6 \times 0,4 = 0,56 & 0,4 \times 0,2 + 0,6 \times 0,6 = 0,44 \end{bmatrix}$$

O leitor pode reconhecer nas entradas da matriz P^2 os cálculos feitos para as probabilidades relacionadas a **depois de amanhã**, dada a situação de hoje.

		depois de amanhã	
		0 = sol	1 = chuva
hoje	0 = sol	0,72	0,28
	1 = chuva	0,56	0,44

Se continuássemos com o diagrama de árvore para depois de depois de amanhã, veríamos que os resultados das correspondentes probabilidades de transição referentes a sol e a chuva seriam obtidos por meio do cálculo de P^3 . Isso favorece a exploração de multiplicação de matrizes com os alunos. O leitor interessado pode encontrar na bibliografia material sobre essa área, embora o desenvolvimento para além desses passos não seja trivial para o ensino básico.

COMENTÁRIOS FINAIS

O importante foi enfatizar a existência de estruturas que guardam uma dependência entre as ocorrências, dado que no nível básico pouco se fala de aplicações além da independência entre eventos ou entre variáveis.

Muitas são as áreas que utilizam essa abordagem: *marketing*, para caracterizar fidelidade de clientes a produtos (mudam ou não de marca, dado

que são usuários dessa ou daquela marca); saúde, para estudar mudanças de estado de pacientes internados (de acordo com a gravidade da doença, etc.), sociologia, para observar mobilidade social, dentre outras.

Um exemplo antigo e muito citado na área de Marketing deve-se a Styan&Smith Jr. ([7]), que usaram dados sobre o consumo durante 28 semanas de cem famílias norte-americanas, com relação a compras de produtos de limpeza. Nesse estudo, os estados eram (1) família só compra detergente; (2) família só compra sabão em pó; (3) família compra os dois produtos juntos: sabão em pó e detergente; (4) família não compra nenhum produto de limpeza. Aqui as probabilidades para a construção da matriz de transição da cadeia de Markov foram estimadas a partir da observação de cada família, semana por semana, ao longo de vários meses. O artigo tinha como objetivo responder à pergunta: as porcentagens apresentadas refletem uma certa lealdade dos consumidores ou há mudanças de semana em semana?

A propósito de saúde, um estudo que usou cadeias de Markov é o de Boyle et al. ([2]) cujo objetivo é prever a incidência de diabetes no ano de 2050 nos EUA, onde os estados eram 1 = não ter diabetes, 2 = não ter diabetes diagnosticada, 3 = ter diabetes. É interessante, nesse exemplo, verificar, por exemplo, o valor zero da probabilidade p_{31} (ou seja, passar do estado **ter diabetes** (3) para o estado **não ter diabetes** (1) é impossível, pelo que se conhece até agora).

Greenwall et al. ([4]) relatam um estudo de mobilidade socioeconômica, usando as ideias de sociólogos sobre o tema, que dizem que o mais forte determinante da classe socioeconômica de um indivíduo é a classe à qual pertencem seus pais. Esse estudo mostra também outros passos para tratar com o problema para as gerações seguintes.

Todos os exemplos citados acima são introduzidos através de cadeias de Markov. Outro exemplo a ser citado é a interação dos neurônios, que é tema

de cadeias mais complexas, linha de pesquisa atual do Cepid NeuroMat (IMEUSP), cujos pesquisadores trabalham para construir uma teoria sobre o funcionamento do cérebro.

O horizonte para essas pesquisas é amplo e o desafio é modelar a incerteza com que certos eventos aparecem no mundo real. Embora, como já comentamos, a formalização dessa teoria não seja trivial para a educação básica, estimular os estudantes com as ideias que oferecemos nesse artigo é dar um passo à frente para o futuro aprendizado de probabilidade e estatística, bem como sugerir caminhos de investigação em outras áreas, aplicando essa teoria.

Nota: A segunda autora participou deste artigo como parte das atividades do Centro de Pesquisa, Inovação e Difusão em Neuromatemática (Cepid NeuroMat), financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), processo 2013/07699-0.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Bartholomew, D.J. *Stochastic Models for Social Processes*. London: John Wiley & Sons, 1973.
- [2] Boyle et al. *Population Health Metrics*, 8:29. Biomed Central (online), 2010.
- [3] Durrett, R. *Essentials of stochastic processes*, 2nd. edition. New York: Springer, 2012.
- [4] Greenwell, R.N. et al. *Calculus for the Life Sciences*. Boston: Addison-Wesley, 2003.
- [5] Morettin, P.A. & Bussab, W.O. *Estatística Básica*. 7ª. ed. São Paulo: Saraiva, 2012.
- [6] Stirzaker, D. *Elementary Probability*. Cambridge: Cambridge University Press, 2003.
- [7] Styan, G.P.H. & Smith Jr, H. *Markov Chains applied to Marketing*. *Journal of Marketing Research*, 1; p. 50-55, 1964.
- [8] Tijms, H.C. *A first course in Stochastic Models*. Chichester: John Wiley & Sons, 2003.

