

# Elementos de Estatística Bayesiana

Marcus L. Nascimento

21 de outubro de 2025

1. Introdução
2. Princípios da Aprendizagem Bayesiana
3. Inferência Bayesiana
4. Monte Carlo via Cadeias de Markov (MCMC)

# Introdução

- O arcabouço moderno acerca da teoria de probabilidade foi desenvolvido por Andrey Kolmogorov (1903-1987) quando o mesmo estabeleceu probabilidade em termos de teoria dos conjuntos/teoria da medida (Kolmogorov, 2018).
  - Estrutura matemática coerente para se trabalhar com probabilidades;
  - Não há o desenvolvimento de uma interpretação de probabilidade.
- Em estatística, há duas interpretações majoritárias de probabilidade: **Frequentista** e **Bayesiana**.
  - Ambas interpretações tomam como base a teoria desenvolvida por Kolmogorov.

# Interpretação Frequentista

- Interpretação frequentista:
  - Probabilidade = Frequência (sequência de experimentos idênticos repetidos inúmeras vezes);
  - Visão ontológica (probabilidade “existe” e é idêntica a algo que pode ser observado);
  - Restritiva no sentido de que não é aplicável para descrever eventos que ocorrem apenas uma vez.
    - Aplicada assintoticamente (grandes amostras).

# Interpretação Bayesiana

- Interpretação Bayesiana:
  - “Probability does not exist- Bruno de Finetti (1906–1985).
    - Probabilidade é atribuída e mutável, não uma propriedade inerente a um objeto.
  - Probabilidade é interpretada como a descrição do conhecimento e de incerteza.
  - Interpretação válida independentemente do tamanho da amostra ou do número de repetições de um experimento.

## OBSERVAÇÃO:

- Note que a diferença entre as duas interpretações não está no uso do Teorema de Bayes, mas na contraposição entre uma visão ontológica (frequentista) e uma visão epistemológica (bayesiana) de probabilidade.

# Princípios da Aprendizagem Bayesiana

# Da Priori a Distribuição a Posteriori

## Teorema de Bayes:

A probabilidade de um evento  $A$  dado o evento  $B$  é

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}.$$

- O aprendizado estatístico bayesiano aplica o teorema de Bayes para atualizar o estado de conhecimento acerca de um parâmetro a luz de um conjunto de dados.
- Ingredientes:
  - $\theta$  parâmetro(s) de interesse;
  - Distribuição a priori com densidade  $p(\theta)$  descrevendo a incerteza sobre  $\theta$ ;
  - Processo gerador dos dados  $p(x|\theta)$ .



# Da Priori a Distribuição a Posteriori

- Como tanto a distribuição a priori sobre os parâmetros quanto o processo gerador dos dados devem ser especificador, o modelo subjacente à abordagem bayesiana consiste na distribuição conjunta

$$p(\theta, x) = p(\theta)p(x|\theta).$$

- Como a incerteza sobre  $\theta$  é atualizada a luz de uma nova informação (nova observação  $x$ )?

Através da aplicação do Teorema de Bayes a densidade da priori é atualizada para uma densidade a posteriori.

$$\underbrace{p(\theta|x)}_{\text{posteriori}} = \underbrace{p(\theta)}_{\text{priori}} \times \frac{p(x|\theta)}{p(x)}.$$

# Da Priori a Distribuição a Posteriori

Note que no denominador da fórmula de Bayes é necessário computar  $p(x)$ , onde

$$\begin{aligned} p(x) &= \int_{\Theta} p(x, \theta) d\theta \\ &= \int_{\Theta} p(\theta) p(x|\theta) d\theta, \end{aligned}$$

ou seja, a marginalização de  $\theta$  da distribuição conjunta de  $x$  e  $\theta$ .

A depender do contexto  $p(x)$  é denominada

- **Constante de normalização** já que garante que o valor da integral da densidade a posteriori  $p(\theta|x)$  seja 1;
- **Densidade preditiva a priori** da observação  $x$  dado o modelo  $M$  antes que algum dado seja visto. Para enfatizar o condicionamento implícito no modelo, é possível escrever  $p(x|M)$ .

# Verossimilhança e Atualização Bayesiana

Considerando um conjunto de observações independentes e identicamente distribuídas  $D = \{x_1, x_2, \dots, x_n\}$  tenha sido observado, a posteriori é descrita da seguinte forma:

$$\underbrace{p(\theta|D)}_{\text{posteriori}} = \underbrace{p(\theta)}_{\text{priori}} \times \frac{\mathcal{L}(\theta|D)}{p(D)},$$

onde  $\mathcal{L}(\theta|D) = \prod_{i=1}^n p(x_i|\theta)$  é a verossimilhança e  $p(D) = \int_{\Theta} p(\theta)p(x_1, \dots, x_n|\theta)d\theta$ .

Comparando a verossimilhança com o procedimento bayesiano:

- Conduzir uma análise estatística bayesiano requer a solução de uma integral a fim de encontrar a constante normalizadora;
- Conduzir uma análise via verossimilhança requer resolver um problema de otimização a fim de encontrar a máxima verossimilhança.

# Atualização Sequencial

- O procedimento de atualização bayesiano pode ser repetidamente: a posteriori pode ser utilizada como uma nova priori e, então, atualizada com um novo conjunto de dados.
- É possível atualizar a densidade a posteriori sequencialmente com os dados  $x_1, x_2, \dots, x_n$  sendo observados um após o outro.

$$\begin{aligned} p(\theta|x_1) &= p(\theta) \times \frac{p(x_1|\theta)}{p(x_1)} \\ p(\theta|x_1, x_2) &= p(\theta|x_1) \times \frac{p(x_2|\theta, x_1)}{p(x_2|x_1)} = p(\theta|x_1) \times \frac{p(x_2|\theta)}{p(x_2|x_1)} = p(\theta) \times \frac{p(x_1|\theta)p(x_2|\theta)}{p(x_1)p(x_2|x_1)} \\ &\vdots \\ p(\theta|x_1, \dots, x_n) &= p(\theta) \times \frac{\prod_{i=1}^n p(x_i|\theta)}{p(D)}. \end{aligned} \tag{1}$$

# Atualização Sequencial

Na equação (1),  $p(D)$  é dado por:

$$p(D) = \prod_{i=1}^n p(x_i|x_{<i}), \text{ onde } p(x_i|x_{<i}) = \int_{\Theta} p(x_i|\theta)p(\theta|x_{<i})d\theta.$$

- $p(x_i|x_{<i})$  é a densidade preditiva a posteriori de uma nova observação  $x_i$  após serem observados  $x_1, \dots, x_{i-1}$ ;
- Como a incerteza associada ao parâmetro  $\theta$  depende da quantidade de dados já observados, a probabilidade de uma nova observação  $x_i$  depende dos dados observados previamente.
  - Apenas nos casos em que o parâmetro é completamente conhecido e não há incerteza associada ao mesmo, as observações  $x_i$  são independentes.

# Inferência Bayesiana

# Estimadores de Bayes

- Sob a perspectiva bayesiana, há uma uma forte ligação entre estimação e a Teoria da Decisão.
  - Sendo  $\delta$  o estimador de Bayes, busca-se minimizar o risco a posteriori sob uma função perda  $L(\theta, \delta(x))$ ;
  - Um bom estimador de Bayes é aquele para o qual, com alta probabilidade, o erro  $\delta(x) - \theta$  é o menor possível.
- Definiremos risco a posteriori como

$$r(\delta|x) = E_{\theta|x}[L(\theta, \delta(x))] = \int_{\Theta} L(\theta, \delta(x)) d\mu(\theta|x).$$

- As principais funções perda consideradas são perda quadrática, perda absoluta e perda 0-1.

## Estimadores de Bayes (Perda Quadrática)

Considerando o risco a posteriori com relação a função perda quadrática,  $L(\theta, \delta(x)) = (\theta - \delta)^2$ , temos:

$$\begin{aligned}r(\delta|x) &= E_{\theta|x} [L(\theta, \delta(x))] \\&= E_{\theta|x} [(\theta - \delta)^2] \\&= E_{\theta|x} [\theta^2 - 2\delta\theta + \delta^2] \\&= E_{\theta|x}(\theta^2) - 2\delta E_{\theta|x}(\theta) + \delta^2.\end{aligned}\tag{2}$$

Minimizando (2) com relação a  $\delta$ , obtemos:

$$\frac{dr(\delta|x)}{d\delta} = 0 - 2E_{\theta|x}(\theta) + 2\delta = 0 \Rightarrow \delta = E_{\theta|x}(\theta) \text{ (Média a posteriori)}.$$

Como  $\frac{d^2r(\delta|x)}{d\delta^2} = 2 > 0$ ,  $E_{\theta|x}(\theta)$  é ponto de mínimo.



# Estimadores de Bayes (Perda Absoluta)

Considerando o risco a posteriori com relação a função perda absoluta,  $L(\theta, \delta(x)) = |\theta - \delta|$ , temos:

$$\begin{aligned}r(\delta|x) &= \int_a^b |\theta - \delta| d\mu(\theta|x) \\ &= \int_a^\delta (\delta - \theta) d\mu(\theta|x) + \int_\delta^b (\theta - \delta) d\mu(\theta|x).\end{aligned}\tag{3}$$

Para derivar com respeito a  $\delta$ , aplicaremos a regra de Leibniz, isto é,

$$\begin{aligned}\frac{d}{dx} \left( \int_{a(x)}^{b(x)} f(x, t) dt \right) &= \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x, t) dt \\ &\quad + f(x, b(x)) \frac{d}{dx} b(x) - f(x, a(x)) \frac{d}{dx} a(x).\end{aligned}$$

# Estimadores de Bayes (Perda Absoluta)

Minimizando (3) com relação a  $\delta$ , obtemos:

$$\begin{aligned}\frac{dr(\delta|x)}{d\delta} &= \int_a^\delta 1 \, d\mu(\theta|x) + (\delta - \delta) \times 1 - (\delta - a) \times 0 \\ &\quad - \int_\delta^b 1 \, d\mu(\theta|x) + (b - \delta) \times 0 - (\delta - \delta) \times 1 \\ &= \int_a^\delta 1 \, d\mu(\theta|x) - \int_\delta^b 1 \, d\mu(\theta|x) \\ &= 0 \\ \Rightarrow \int_a^\delta 1 \, d\mu(\theta|x) &= \int_\delta^b 1 \, d\mu(\theta|x).\end{aligned}\tag{4}$$

## Estimadores de Bayes (Perda Absoluta)

A igualdade em (4) é obtida quando  $\delta$  é igual à mediana. Aplicando novamente a regra de Leibniz para verificar se é ponto de mínimo, temos

$$\begin{aligned}\frac{d^2 r(\delta|x)}{d\delta^2} &= \int_a^\delta 0 \, d\mu(\theta|x) + 1 \times \frac{d}{d\delta}\delta - 1 \times \frac{d}{d\delta}a \\ &\quad - \left( \int_\delta^b 0 \, d\mu(\theta|x) + 1 \times \frac{d}{d\delta}b - 1 \times \frac{d}{d\delta}\delta \right) \\ &= 0 + 1 \times 1 - 1 \times 0 - (0 + 1 \times 0 - 1 \times 1) = 2 > 0\end{aligned}$$

Logo, a mediana a posteriori minimiza  $r(\delta|x)$ .

# Estimadores de Bayes (Perda 0-1)

A função perda 0-1 é definida como

$$L(\theta, \delta(x)) = 1 - \Delta(\theta - \delta),$$

onde  $\Delta(\theta - \delta) = 1$  quando  $\theta = \delta(x)$  e  $\Delta(\theta - \delta) = 0$  caso contrário.

Considerando o risco a posteriori com relação a função perda absoluta, temos:

$$\begin{aligned} r(\delta|x) &= \int_{\Theta} (1 - \Delta(\theta - \delta)) d\mu(\theta|x) \\ &= 1 - \int_{\Theta} \Delta(\theta - \delta) d\mu(\theta|x) \\ &= 1 - d\mu(\delta|x). \end{aligned}$$

O risco bayesiano é mínimo quando a densidade a posteriori  $d\mu(\delta|x)$  é máxima e este máximo é atingido quando  $\delta$  é a moda.

- O conjunto  $C \in \tau$  é dito ser um conjunto de credibilidade  $\alpha$  se

$$\mathbb{P}(\theta \in C | X = x) \geq 1 - \alpha.$$

- Existem diversos conjuntos com a mesma probabilidade e a definição de um critério para a escolha entre os diferentes conjuntos é necessária.
- Minimização do volume entre as diferentes regiões de credibilidade  $\alpha$ .
  - Um conjunto  $C \in \tau$  é denominado região HPD (*Highest Posterior Density*) de credibilidade  $\alpha$  se  $C = \{\theta : d\mu_{\Theta|X}(\theta|x) \geq k\}$ , onde  $k$  é o maior valor tal que

$$\mathbb{P}(\Theta \in C | X = x) \geq 1 - \alpha.$$

- $k$  pode ser interpretado como uma linha horizontal posicionada sobre a densidade a posteriori cujas interseções definem uma região com probabilidade  $1 - \alpha$ .

## Monte Carlo via Cadeias de Markov (MCMC)

- Métodos de Monte Carlo fornecem uma abordagem numérica para a solução de funções complicadas.
  - Em vez de resolver analiticamente, aproxima-se a solução tomando amostras das distribuições.
- Não raramente, não é possível amostrar diretamente de uma distribuição.
  - Suponha que desejamos amostrar de  $p(z)$ , onde

$$p(z) = \frac{f(z)}{K}.$$

- Em casos nos quais  $f(z)$  é conhecida, porém  $K$  é difícil de estimar, não conheceremos  $p(z)$  e, conseqüentemente, não conseguiremos amostrar diretamente da distribuição de interesse.

- Algoritmos de Monte Carlo via Cadeias de Markov (MCMC) trata tais casos permitindo que a estimação a partir de  $p(z)$  seja realizada considerando uma função  $f(z)$  proporcional a  $p(z)$ 
  - O algoritmo constrói uma cadeia de Markov com valores de  $z$  tais que a distribuição estacionária da cadeia  $\pi(z)$  seja igual a  $p(z)$ .
- Em inferência bayesiana, a distribuição de interesse consiste na distribuição a posteriori que, por sua vez, é função da constante de normalização  $p(D)$ .
  - Recordando que  $p(D) = \int_{\Theta} p(\theta)p(x_1, \dots, x_n|\theta)d\theta$ , não é difícil observar que tal integral não raramente tal integral não poderá ser resolvida analiticamente;
  - Como a posteriori é proporcional ao produto entre a priori e a verossimilhança, algoritmos MCMC são amplamente utilizados.



# Algoritmos de Metropolis-Hastings

- Suponha  $q(y|z)$  uma densidade condicional e  $p(z)$  a densidade objetivo, o algoritmo de Metropolis-Hasting constrói uma cadeia de Markov  $(Z_n)$  através dos passos abaixo:
  1. Inicialize  $x_k$ ;
  2. Gere  $x^{\text{prop}}$  com distribuição  $q(\cdot|x_k)$ ;
  3. Tome

$$x_{k+1} = \begin{cases} x^{\text{prop}}, & \text{com probabilidade } \alpha \\ x_k, & \text{com probabilidade } 1 - \alpha, \end{cases}$$

onde

$$\alpha = \min \left\{ 1, \frac{p(x^{\text{prop}})q(x_k|x^{\text{prop}})}{p(x_k)q(x^{\text{prop}}|x_k)} \right\}.$$

- A densidade  $q$  é conhecida como densidade proposta e  $\alpha$ , como probabilidade de aceitação.

# Algoritmos de Metropolis-Hastings

- A escolha da densidade proposta é essencial para a implementação do algoritmo.
  - Passeio aleatório;
  - Propostas independentes.
- No caso passeio aleatório, a ideia é explorar a vizinhança no em torno do valor atual da cadeia. Podemos considerar, por exemplo:

$$x^{\text{prop}} = x_k + \varepsilon,$$

onde onde  $\varepsilon$  é uma perturbação aleatória simétrica em torno de 0.

- $q(x^{\text{prop}}|x_k) = g(|x^{\text{prop}} - x_k|)$ , ou seja,  $g$  é simétrica em torno de  $x_k$ ;
- $q(x^{\text{prop}}|x_k) = g(|x^{\text{prop}} - x_k|)$  implica que  $q(x^{\text{prop}}|x_k) = q(x_k|x^{\text{prop}})$ .

# Algoritmos de Metropolis-Hastings

- Na prática, a distribuição normal é bastante aplicada para a perturbação aleatória  $\varepsilon$  e, neste caso, é necessário ajustar a variância da distribuição:
  - Variância pequena implica em uma maior aceitação, porém o domínio dos parâmetros é mais lentamente explorado;
  - Variância pequena implica em uma menor aceitação, porém visitamos mais rapidamente o domínio de interesse.
- No caso passeio aleatório, a proposta não depende do valor da cadeia no passo anterior,  $q(x|x_k) = g(x)$ .
  - Como a probabilidade de aceitação depende do passo anterior, ainda temos uma cadeia de Markov.

$$\alpha = \min \left\{ 1, \frac{p(x^{\text{prop}})g(x_k)}{p(x_k)g(x^{\text{prop}})} \right\}$$

- Algoritmo MCMC onde a densidade proposta é a distribuição condicional completa.
  - Resultados teóricos garantem que podemos recuperar a distribuição conjunta a partir de distribuições condicionais;
  - Em vez de gerarmos valores de uma distribuição conjunta, podemos gerar valores da condicionais.
- No Amostrador de Gibbs, as propostas são sempre aceitas.

- O algoritmo consiste dos seguintes passos:
  1. Inicialize  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})$ .
  2. Amostre de cada condicional completa iterativamente.
    - $\theta_1^{(j)}$  com densidade  $\pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, y)$ ;
    - $\theta_2^{(j)}$  com densidade  $\pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, y)$ ;
    - $\theta_3^{(j)}$  com densidade  $\pi(\theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_p^{(j-1)}, y)$ ;
    - $\vdots$
    - $\theta_p^{(j)}$  com densidade  $\pi(\theta_p | \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_{p-1}^{(j)}, y)$ ;

Kolmogorov, Andrey Nikolaevich. 2018. *Foundations of the Theory of Probability*. 2nd ed. New York: Dover.