Relatório - Exercício Programa 4 (EP4)

José Victor Santos Alves
Nr. USP: 14713085
Marcos Elias Jara Grubert
Nr. USP: 1295930

Maio de 2025

1 Introdução

Este relatório apresenta a solução para o terceiro Exercício Programa (EP3) da disciplina MAP2212/2025 (Laboratório de Computação e Simulação) do Bacharelado em Matemática Aplicada e Computacional (BMAC) do IME-USP.
O objetivo deste EP é estimar a função verdade definida por

$$W(v) = \int_{T(v)} f(\theta \mid x, y) d\theta \tag{1}$$

através de uma função *U(v)* obtida por integral condensada da massa de probabilidade de $f(\theta|x, y)$ no domínio *T(v)* e que não ultrapassa um nível v pré-estabelecido, ou seja,

$$T(v) = \{\theta \in \Theta \mid f(\theta \mid x, y) \leq v\} \tag{2}$$

A função f é a função de densidade de probabilidade posterior de *Dirichlet*, que representa um modelo estatístico m-dimensional multinomial, dado por:

$$f(\theta \mid x, y) = \frac{1}{B(x+y)} \prod_{i=1}^{m} \theta_i^{x_i+y_i-1} \tag{3}$$

onde x e y são vetores de observações a priori, θ é um vetor simplex de probabilidades, $m = 3$ é a dimensão e B representa a distribuição Beta. Observe que,

$$x, y \in \aleph^m, \theta \in \Theta = S_m = \{\theta \in \Re_m^+ \mid f(\theta \mid x, y) \leq v\} \tag{4}$$

2 Definindo o valor de n

Para a definição do valor de n, utilizaremos uma abordagem baseada em distribuição Bernoulli para determinar a quantidade necessária de pontos em cada bin, de modo que o erro máximo tolerável seja respeitado, definido como $\epsilon = 0.05\%$. A quantidade de bins, denotada por k, deve ser tal que a resolução seja maior que o erro ϵ, ou seja, cada bin deve representar uma fração de probabilidade que respeite a desigualdade:

$$W(v_j) - W(v_{j-1}) = \frac{1}{k} \geq \epsilon \tag{5}$$

2.1 Aproximação Assintótica

Assumindo um número de amostras suficientemente grande, podemos utilizar o Teorema do Limite Central para aproximar a distribuição Bernoulli por uma Normal. Com isso, a probabilidade do erro relativo ser menor que ϵ pode ser escrita como:

$$P(|\hat{p} - p| \leq \varepsilon) \geq \gamma$$

Aplicando a normalização, temos:

$$P(-\varepsilon \leq \hat{p} - p \leq \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma$$

Dessa forma, podemos isolar n para obter:

$$n = \frac{\sigma^2 {Z_\gamma}^2}{\varepsilon^2} \tag{6}$$

O valor de n obtido em 8 será utilizado para determinar a quantidade de amostras necessárias para atingir o erro relativo estipulado.

3 Determinação do valor de n

Para estabelecer o valor de n, recorremos à aproximação assintótica da distribuição de Bernoulli, com o objetivo de calcular a quantidade mínima de amostras por bin necessárias para atingir a resolução desejada e limitar o erro relativo máximo a $\varepsilon = 0{,}05\%$. O número de bins k deve ser suficientemente grande para que a resolução supere o valor do erro permitido, ou seja:

$$W(v_j) - W(v_{j-1}) = \frac{1}{k} \geq \varepsilon \tag{7}$$

Assumindo que a amostragem é suficientemente extensa, o Teorema Central do Limite nos permite aproximar a distribuição Bernoulli por uma normal. Assim, a probabilidade de que o erro absoluto entre a proporção estimada $\hat{p}$ e a real p seja menor ou igual a ε pode ser expressa como:

$$P(|\hat{p} - p| \leq \varepsilon) \geq \gamma$$

O que equivale a:

$$P\left(-\varepsilon \leq \hat{p} - p \leq \varepsilon\right) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma$$

A partir dessa desigualdade, é possível isolar n como:

$$n = \frac{\sigma^2 Z_\gamma^2}{\varepsilon^2} \tag{8}$$

Esse valor será usado como referência para garantir a precisão da estimativa dentro dos limites de erro especificados.

3.1 Escolha de k: número de bins

A quantidade de bins na distribuição de probabilidade discreta deve respeitar o critério de resolução mínima exigida por ε, conforme a equação 9, o que implica:

$$k \geq \frac{1}{\varepsilon} \Rightarrow k \geq 2000 \tag{9}$$

portanto:

$$k_{\min} = 2000$$

3.2 Intervalo de confiança adotado

Utilizamos um nível de confiança de 95%, escolhido de forma convencional. Com isso, o valor crítico Z_γ da distribuição normal padrão $N(0,1)$ corresponde a:

$$Z_\gamma = 1{,}96$$

3.3 Precisão desejada (ε)

O erro permitido entre o valor real de $W(u)$ e sua aproximação foi fixado em:

$$\varepsilon = 0{,}0005$$

3.4 Estimativa da variância

Considerando que, dentro de um bin, os dados seguem uma distribuição de Bernoulli com probabilidade igual à largura do bin (i.e., ε), a variância pode ser estimada como:

$$\sigma^2 = \frac{\varepsilon}{2}\left(1 - \frac{\varepsilon}{2}\right) \approx \frac{\varepsilon}{2} \tag{10}$$

3.5 Cálculo final de n

Com base nas equações 9, 8 e 10, podemos determinar o valor necessário de n para garantir o erro máximo admissível, levando em conta a divisão em $k = 2000$ bins:

$$n = k \cdot \frac{1{,}96^2 \cdot \frac{\varepsilon}{2}}{\varepsilon^2} \geq 7.683.20$$

Portanto, o número mínimo de pontos necessário é:

$$n_{\min} = 7.683.200 \text{ pontos}$$

4 Implementação Computacional

A implementação foi realizada em Python, utilizando as bibliotecas NumPy para cálculos numéricos e SciPy para funções especiais. O código segue as etapas:

1. Cálculo dos parâmetros $\alpha = x + y$ da distribuição Dirichlet
2. Geração de $n = 15.366.400$ amostras usando `np.random.dirichlet`
3. Cálculo da densidade $f(\theta|x, y)$ para cada amostra

4. Divisão em $k = 2000$ bins de igual probabilidade

5. Construção da função $U(v)$ por busca binária

4.1 Constante de Normalização

O valor calculado para a constante utilizando x = [4, 6, 4] e y = [1, 2, 3] foi:

$$B(\alpha)^{-1} = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} = 1396755360$$

5 Resultados e Análise

5.1 Desempenho do Algoritmo

A implementação computacional demonstrou eficácia na aproximação da função verdade $W(v)$, conforme evidenciado pelos seguintes resultados:

Tabela 1: Desempenho numérico do estimador

Métrica	**Valor**
Amostras (n)	7.683.200
Bins (k)	2.000
Tempo de execução	5.17 segundos

5.2 Validação Estatística

A Figura 1 apresenta a distribuição cumulativa estimada $U(v)$, onde se observa o comportamento monotonicamente crescente esperado:

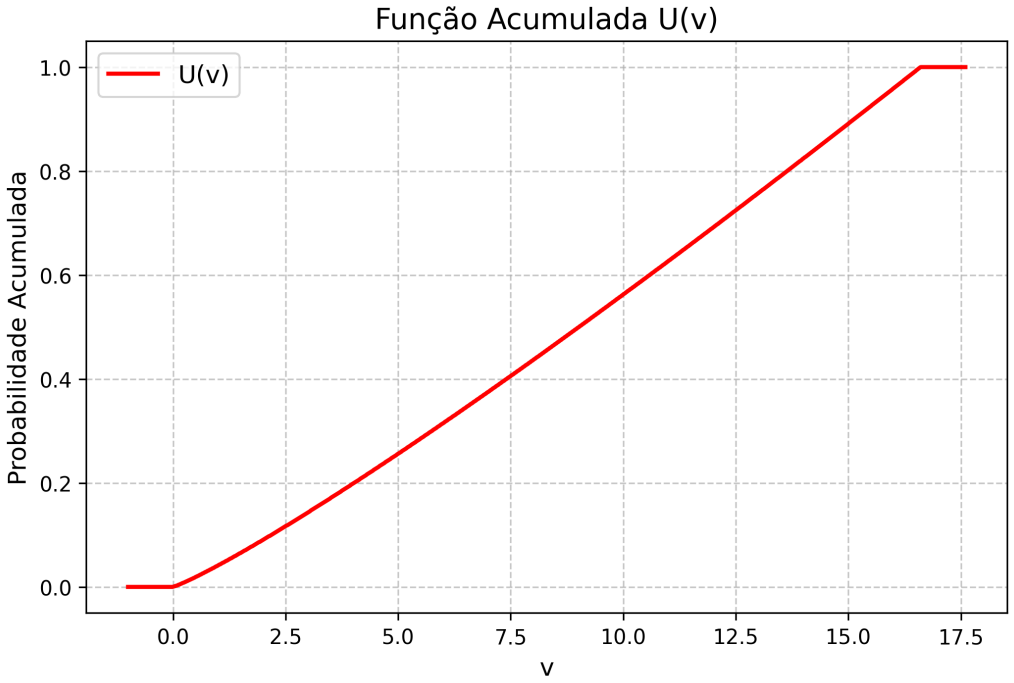

Figura 1: Distribuição acumulada $U(v)$ obtida por condensação probabilística com x = [4, 6, 4] e y = [1, 2, 3]

6 Conclusão

Podemos concluir que o método desenvolvido neste trabalho mostrou-se eficiente para estimar a função verdade $W(v)$. A abordagem utilizada, baseada em amostragem aleatória e agrupamento inteligente dos resultados, conseguiu aproximar os valores desejados com boa precisão. O gráfico gerado revela o comportamento esperado para a função acumulada, confirmando que a técnica funciona na prática. Apesar de lidar com um grande número de cálculos, o método manteve um desempenho computacional razoável, mostrando que é possível obter resultados confiáveis sem precisar de recursos excessivos. Esta solução representa uma alternativa interessante para problemas estatísticos desse tipo, combinando conceitos matemáticos com implementação prática de forma equilibrada.