Exercício Programa 4 MAP-2212

Luiz Guilherme De Padua Sanches NUSP:13686431
Daniel Vertamatti NUSP:2370859
Victor de Faro Quadros NUSP: 12556632

Maio, 2023

1 Objetivos e considerações

Foi solicitada no enunciado a aproximação da seguinte função, denominada função verdade $W(v)$, através de uma função $U(v)$ com erro de até 0,05%:

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

A região de integração T(v) é definida por:

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

Em que v é um nível preestabelecido, θ é um vetor de parâmetros no simplex m-dimensional e $f(\theta|x,y)$ é a uma densidade definida como:

$$f(\theta|x,y) = \frac{1}{B(x+y)} \prod_{i=1}^{m} \theta_i^{x+y-1}$$

Essa função densidade é conhecida como função densidade de probabilidade de Dirichlet, em que x é um vetor de observações e y um vetor de informações a priori, ambos com dimensão m.

Nesse projeto, foi solicitada dimensão igual a três, ou seja, $m = 3$. Os vetores x e y são solicitados ao usuário e o vetor de parâmetros θ será gerado através deste gerador Dirichlet.

A função $f(\theta|x,y)$, durante a maior parte do programa, não utilizará a constante de normalização (recebendo o nome de função potencial). Essa constante será utilizada apenas quando informado o valor v, em que esse valor será multiplicado pela constante. Dessa forma, o valor de v refere-se ao valor da função densidade.

Foi requisitado no exercício que sejam feitas K partições, em que a quantidade de pontos dentro dessas partições seja aproximadamente igual, e que essas partições sejam utilizadas como uma aproximação de uma partição da função verdade.

2 Estimativa do erro

2.1 Partições

Foi solicitado que $W(t_j) - W(t_{j-1})$ seja aproximadamente $\frac{1}{K}$. O erro máximo admitido dessa diferença é 0,05%.

Dessa forma:

$$\frac{1}{K} \leq 0,05\%$$

$$K \geq 2.000$$

Definiu-se $K = 2.500$ a partir desse critério.

2.2 Número de elementos gerados

Para isso, considera-se que a probabilidade de um ponto qualquer recair numa partição específica é uma variável Bernoulli.

A probabilidade de recair na partição é dada por:

$$p = \frac{1}{2500}$$

enquanto a probabilidade de não recair nela é:

$$1 - p = \frac{2500 - 1}{2500} = \frac{2499}{2500} \approx 1$$

A variância de uma variável aleatória Bernoulli é dada pela multiplicação desses dois valores ($\sigma^2 = p(1-p)$). Dessa forma aproximou-se a variância por $\sigma^2 = \frac{1}{2500}$

Utiliza-se o Teorema Central do Limite para aproximar-se a Bernoulli por uma Distribuição Normal.

Sabe-se que o erro em estimação por intervalo, em uma normal, é dado por:

$$\epsilon = z_{\gamma/2} \frac{\sigma}{\sqrt{n}}$$

Que manipulando chega-se a:

$$n = {z_{\gamma/2}}^2 \frac{\sigma^2}{\epsilon^2}$$

Substituindo $z_{\gamma/2}$ por 1,96 (95% de taxa de acerto), o valor da variância e do erro chega-se ao valor para n:

$$n = 6146,56$$

Adotou-se:

$$n = 6150$$

Essa é a quantidade necessária para cada partição. Multiplicando esse valor por K chega-se ao valor total de pontos necessários $N = 15.375.000$.

3 Funcionamento do programa

Inicialmente solicita-se ao usuário a inserção de valores para os vetores X e Y, que serão utilizados para as futuras operações.

Geram-se 15.375.000 amostras utilizando-se o gerador Dirichlet. Esses valores são utilizados no cálculo dos valores da função potencial.

Os valores da função potencial são inseridos num vetor, chamado de vetor *valores*, que é posteriormente ordenado e separado nas 2.500 partes. São colocados em um segundo vetor, chamado *vetor fronteira*, os maiores pontos de cada parte. O último ponto do vetor fronteira (o maior dos números, ou teto) é mostrado ao usuário, conforme solicitado no enunciado.

É solicitado que o usuário insira um valor v maior igual a zero. Esse valor é multiplicado pela constante de normalização e é buscada sua posição no vetor fronteira. Essa posição é multiplicada por 6.150 e o valor dessa multiplicação alocado em uma variável.

A posição encontrada no vetor fronteira refere-se a uma partição específica. Dessa forma, também é realizada a busca da posição do valor v nessa partição. Esta posição é somada ao valor atual da variável.

A aproximação de $W(v)$ é dada pela divisão entre o valor da variável e o valor total de pontos gerados. Essa aproximação é arredonda para 4 casas decimais e apresentada ao usuário.

4 Conclusão

Os resultados do programa, para as entradas padrão fornecidas ($x = [4, 6, 4]$ e $y = [1, 2, 3]$) foram comparadas com sucesso aos resultados esperados. O tempo de execução é principalmente devido à geração dos mais de 15 milhões de pontos, ordenação, e cortes necessários para agrupamento dos dados. Numa máquina mediana, o tempo de execução não leva mais que 30 segundos, o que foi considerado bastante razoável. O método sugerido pelo Professor, em utilizar a Função Potencial e não a FDP, ou seja, trabalhar sem a constante de normalização, aumenta a velocidade de execução, sem prejuízo dos resultados finais.