EXERCICIO PROGRAMA 4 - MAP2212 2024

Victor Rocha Cardoso Cruz 11223757

Larissa Aparecida Marques Pimenta Santos 12558620

1 Enunciado

Considere o modelo estatístico multinomial m-dimensional com obsevações x, informação *a priori* y e parâmetro θ. $x, y \in \mathbb{N}^m, \theta \in \Theta = S_m = \{\theta \in \mathbb{R}_+^m : \theta^T 1 = 1\}$
Esse modelo estatístico é composto por:

- Potencial *a posteriori* $f(\theta|x,y) = \prod_{i=i}^m \theta_i^{x_i+y_i-1}$;
- Conjunto de corte $T(v) = \{\theta \in \Theta : f(\theta|x,y) \leq v\}, v \geq 0$;
- Função verdade $W(v) = \int_{T(v)} f(\theta|x,y)d\theta$

$W(v)$ é a massa de probabilidade a posteriori dentro de $T(v)$, i.e., a massa de probabilidade onde o potencial a posteriori, $f(\theta|x,y)$, não excede a cota v.

Obs.: $Dirichlet(\theta|a) = \frac{1}{B(a)} \prod_{i=i}^m \theta_i^{a_i-1}$, onde $m \geq 2, \theta \in S_m, a \in \mathbb{R}_+^m$ e $B(a) = \frac{\prod_{i=1}^m \Gamma(a_i)}{\Gamma(\sum_{i=1}^m a_i)}$ é a função Beta multivariável.

- Defina k pontos de corte, $0 = v_0 < v_1 < \cdots < v_k = \sup f(\theta)$.
- Use um gerador de números aleatórios Gamma para gerar n pontos em $\Theta, \theta_1, \ldots, \theta_n$, distribuídos de acordo com a função de densidade *a posteriori*.
- Use a fração de pontos simulados θ_t dentro de cada bin, $v_{j-1} < f(\theta_t) < v_j$, como uma aproximação de $W(v_j) - W(v_{j-1})$.
- Ajuste dinamicamente as bordas de cada *bin*, v_j, para obter *bins* com pesos aproximadamente iguais, i.e., $W(v_j) - W(v_{j-1}) \approx \frac{1}{k}$.
- Obtenha como saída uma função $U(v)$ que dê uma boa aproximação de $W(v)$.

Seu programa será avaliado pela acurácia de $U(v)$ (erro $< 0.05\%$), por sua compreensibilidade e pelo seu tempo de execução.

2 Estimativa do erro

2.1 Partições

No enunciado do exercício, é estabelecido que a diferença $W(v_j) - W(v_{j-1})$ seja aproximadamente igual a $1/k$. O erro máximo admitido para essa diferença é de 0,05%. Dessa forma, tem-se:

$$\frac{1}{K} \leq 0,05\%$$
$$K \geq 2000$$

Dessa forma, foi fixado $K = 2500$ para os quantia de partições utilizadas.

2.2 Número de elementos gerados

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

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

Manipulando a expressão acima, é possível isolar n na forma:

$$n = z^2_{\frac{\gamma}{2}} \frac{\sigma^2}{\epsilon^2}$$

Substituindo $z_{\frac{\gamma}{2}}$ por 1,96, quantil da normal correspondente à 95% de taxa de acerto, chega-se a um valor estipulado 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 Implementação do Programa

Inicialmente, são estabelecidos os valores padrão para os vetores X e Y. 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 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.

É solicitado que o usuário insira um valor v maior ou igual a zero. Esse número é multiplicado pela constante de normalização e é buscada sua posição no vetor fronteira. Essa posição é multiplicada por 6.150 e resultado 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 arredondada para 4 casas decimais e apresentada ao usuário.

4 Conclusão

Os resultados do programa, para as entradas padrão fornecidas (x = [1, 2, 3] e y = [4, 5, 6]) foram comparadas com sucesso aos resultados esperados. O longo tempo de execução ocorre principalmente devido à geração dos mais de 15 milhões de pontos, ordenação, e cortes necessários para agrupamento dos dados.