



Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

02 de setembro de 2022

Lista 4: Desafio de velocidade

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 1

Prof. Guilherme Rodrigues

César Augusto Fernandes Galvão (aluno colaborador)

Gabriel Jose dos Reis Carvalho (aluno colaborador)

1. As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Quarto* ou outra ferramenta equivalente.
2. O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, *blogs* e artigos.
3. O trabalho poderá ser feito individualmente ou em dupla. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.
4. Os códigos *R* utilizados devem ser disponibilizados na íntegra, seja no corpo do texto ou como anexo.
5. O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.
6. O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.
7. Escreva seu código com esmero, evitando operações redundantes, visando eficiência computacional, otimizando o uso de memória, comentando os resultados e usando as melhores práticas em programação.

Simulação computacional (https://en.wikipedia.org/wiki/Monte_Carlo_method) é uma poderosa ferramenta amplamente adotada em estudos de sistemas complexos. Aqui, para fins meramente didáticos, simularemos os resultados dos jogos da Copa do Mundo Fifa 2022, sediada no Catar, para responder questões de possível interesse prático.

Consideraremos um modelo probabilístico notavelmente rudimentar e de baixa precisão. Especificamente, assumamos que o resultado do jogo entre os times i e j , com $i \neq j$, segue a distribuição Poisson bivariada definida a seguir.

$$\begin{aligned}(X_i, X_j) &\sim \text{Poisson}(\lambda_{ij}, \lambda_{ji}), \quad \text{com} \\ P(X_i = x_i, X_j = x_j) &= P(X_i = x_i) P(X_j = x_j) \\ &= \frac{\lambda_{ij}^{x_i}}{x_i!} \exp(-\lambda_{ij}) \frac{\lambda_{ji}^{x_j}}{x_j!} \exp(-\lambda_{ji}),\end{aligned}$$

onde X_i e X_j representam o número de gols marcados pelas seleções i e j , respectivamente, $P(X_i, X_j)$ denota a densidade conjunta do vetor (X_i, X_j) e λ_{ij} e λ_{ji} indicam, respectivamente, as médias (esperanças matemáticas) de X_i e X_j . Considere ainda que λ_{ij} é calculado, deterministicamente, como a média entre GF_i e GS_j , onde GF_i e GS_j representam, respectivamente, a média de gols feitos pelo time i nos últimos 15 jogos e a média de gols sofridos pelo time j nos últimos 15 jogos.

As estatísticas dos times classificados para o torneio estão disponíveis em <https://footystats.org/world-cup> e na pasta da tarefa no Teams. A tabela de jogos e o regulamento da Copa estão disponíveis em <https://ge.globo.com/futebol/copa-do-mundo/2022/>.

Questão 1: Simulando a Copa do mundo

Para responder os itens a seguir, use os conhecimentos adquiridos no curso para acelerar o máximo possível os cálculos. Uma lista não exaustiva de opções inclui:

1. Usar uma lógica que evite realizar cálculos desnecessários;
2. Investigar os gargalos do código (*profiling*);
3. Criar parte do código em C++ usando o pacote Rcpp;
4. Executar as operações em paralelo usando um cluster (com múltiplas *cores*) na nuvem.

a) Sob o modelo assumido, qual é a probabilidade do Brasil vencer na estreia por 5x0? Compare o resultado exato com uma aproximação de Monte Carlo baseada em uma amostra de tamanho 1 milhão. Agora, compare com as probabilidades (1/Odds) estimadas pelas casas de aposta (<https://www.bettingodds.com/football/world-cup/brazil-vs-serbia>). O modelo adotado parece ser compatível com os usados pelo “mercado”?

b) Qual é o jogo mais decisivo do Brasil na fase de grupos? Isso é, aquele que, se vencido, leva à maior probabilidade de classificação da seleção para a segunda fase. Responda simulando os resultados do grupo do Brasil.

Observação: Esse tipo de análise é usado para definir questões comercialmente estratégicas como o calendário de competições, preço de comercialização do produto, entre outras.

c) Qual é a probabilidade do Brasil ser campeão, em uma final contra a Argentina, tendo se classificado em primeiro do grupo? Para responder ao item, gere 100 milhões de amostra de Monte Carlo usando um cluster na nuvem!

Atenção: Nas fases eliminatórias, em caso de empate, sorteie o classificado considerando probabilidade de 50% para cada time (como dizem - equivocadamente -, *penalty* é loteria).

Considerações finais

Aqui consideramos um exemplo lúdico, mas o mesmo procedimento é útil para resolver problemas em genética, engenharia, finanças, energia, etc.

Há uma vasta literatura na área de modelagem preditiva de resultados esportivos (via modelos probabilísticos e de aprendizagem de máquina - algorítmicos). Entretanto, por não ser esse o foco do curso, optamos por não modelar o número esperado de gols marcados por equipe. Com base em resultados passados, seria possível ajustar modelos bem mais sofisticados, que levassem em consideração, por exemplo, contra quem os últimos resultados foram alcançados. Decidimos também modelar a incerteza usando distribuições Poisson independentes. Essa é obviamente uma suposição equivocada. Alternativas mais flexíveis podem ser adotadas para melhorar a capacidade preditiva do processo.