Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

19 de fevereiro de 2023

Lista 4: Desafio de velocidade
Resolução - William Rappel - 22/0006032

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 2

Prof. Guilherme Rodrigues

César Augusto Fernandes Galvão (aluno colaborador)

Gabriel José 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 integra, 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 probabilistico notavelmente rudimentar e de baixa precisão. Especificamente, assuma 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 GA_j, onde GF_i e GA_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 múltiplus *cores*) na nuvem.

a) Sob o modelo assumido, qual era 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.

Solução

Primeiro, vamos realizar a leitura dos dados com as estatísticas de todas as seleções para os últimos 15 jogos.

```
stats <- read_excel('data.xlsx')
stats %>% head()
```

```
## # A tibble: 6 x 4
##   Country   Games    GF    GA
##   <chr>     <dbl> <dbl> <dbl>
## 1 Argentina    15    28     4
## 2 Belgium      15    32    19
## 3 Brazil       15    31     5
## 4 Cameroon     15    27    12
## 5 Canada       15    27     8
## 6 Croatia      15    29    15
```

Em seguida, vamos criar um novo objeto, chamado `stats_mean`, contendo a média de gols feitos e sofridos por cada seleção.

```r
stats_mean <- stats %>%
  mutate(GF_mean = GF/Games,
         GA_mean = GA/Games) %>%
  select(Country, GF_mean, GA_mean)
stats_mean %>% head()
```

```
## # A tibble: 6 x 3
##   Country   GF_mean GA_mean
##   <chr>       <dbl>   <dbl>
## 1 Argentina    1.87   0.267
## 2 Belgium      2.13   1.27
## 3 Brazil       2.07   0.333
## 4 Cameroon     1.8    0.8
## 5 Canada       1.8    0.533
## 6 Croatia      1.93   1
```

Em seguida, vamos obter λ_{ij} e λ_{ji}, para Brasil fazer gol na Sérvia e vice-versa.

```r
# funcao para obter medias
extract <- function(country='Brazil', G='GF_mean', data=stats_mean) {
  stats_mean %>%
    filter(Country == country) %>%
    select(any_of(G)) %>%
    pull()
}

# lambdas
lambda_ij <- (extract() + extract('Serbia', 'GA_mean'))/2
lambda_ji <- (extract('Serbia') + extract(G='GA_mean'))/2
```

Em seguida, calculamos a probabilidade do Brasil vencer na estreia por 5x0, a partir do modelo probabilístico assumido.

```r
placar <- c(5, 0)
lambdas <- c(lambda_ij, lambda_ji)
(prob <- placar %>% dpois(lambdas) %>% prod())
```

```
## [1] 0.004342969
```

Logo, a probabilidade é de 0,43%. Agora, vamos comparar o resultado exato com uma aproximação de Monte Carlo baseada em uma amostra de tamanho 1 milhão.

```r
# simulacoes
nsim <- 1e6
goals <- rpois(2*nsim, rep(lambdas, each=nsim))
scores <- tibble(Brazil=goals[1:nsim], Serbia=goals[(nsim+1):(2*nsim)])

# probabilidade
(prob_mc <- scores %>% filter(Brazil == 5 & Serbia == 0) %>% nrow() / nrow(scores))
```

```
## [1] 0.004247
```

A aproximação de Monte Carlo obtida é próxima do resultado teórico.

b) Qual era o jogo mais decisivo do Brasil na fase de grupos? Isso é, aquele que, se vencido, levaria 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.

Solução

Primeiro, vamos criar um vetor com os nomes das seleções do grupo do Brasil. Em seguida, criamos um objeto que conterá os 6 jogos desse grupo, além de colunas relativas aos λs de cada confronto.

```
countries <- c('Brazil', 'Serbia', 'Switzerland', 'Cameroon')
games <- combn(x=countries,  m=2) %>%
  t() %>%
  as_tibble() %>%
  setNames(c('home', 'visitor'))
lambdas <- games %>%
  left_join(stats_mean, by=join_by(home == Country)) %>%
  left_join(stats_mean, by=join_by(visitor == Country), suffix = c('_i', '_j')) %>%
  mutate(lambda_ij = (GF_mean_i + GA_mean_j)/2,
         lambda_ji = (GF_mean_j + GA_mean_i)/2)
```

Agora, criamos uma função para realizar as simulações.

```
simulation <- function(i){
  board <- lambdas
  board$GF_i <- rpois(n=nrow(lambdas), lambda=lambdas$lambda_ij)
  board$GF_j <- rpois(n=nrow(lambdas), lambda=lambdas$lambda_ji)
  board <- board %>%
    mutate(result_i = case_when(GF_i > GF_j ~ '+',
                                GF_i < GF_j ~ '-',
                                TRUE ~ '='),
           result_j = case_when(GF_j > GF_i ~ '+',
                                GF_j < GF_i ~ '-',
                                TRUE ~ '='))
  outcome_i <- board %>% mutate(GA_i = GF_j) %>% select(home, GF_i, GA_i, result_i)
  outcome_j <- board %>% mutate(GA_j = GF_i) %>% select(visitor, GF_j, GA_j, result_j)
  colnames(outcome_j) <- colnames(outcome_i)
  outcome <- outcome_i %>% bind_rows(outcome_j)
  outcome$points <- ifelse(outcome$result_i == '+', 3, ifelse(outcome$result_i == '=', 1, 0))
  outcome$cards <- rpois(n=12, lambda=2)
  class <- outcome %>%
    group_by(home) %>%
    summarise(points=sum(points),
              GF=sum(GF_i),
              GA=sum(GA_i),
              DIFF=GF-GA,
              CARDS=sum(cards)) %>%
    arrange(desc(points), desc(DIFF), desc(GF), CARDS) %>%
    mutate(pos = 1:n())
  output <- board %>%
    slice(1:3) %>%
    select(result_i) %>%
    t() %>%
    cbind(class %>% filter(home == 'Brazil') %>% select(pos) %>% pull() <= 2) %>%
    as_tibble() %>%
    setNames(c('serbia', 'switzerland', 'cameroon', 'classified'))
  output$classified <- as.logical(output$classified)
  row.names(output) <- i
  return(output)
}
```

E rodamos 10.000 simulações com paralelismo de 4 cores.

```
tic()
plan(multisession, workers=4)
result <- future_map(1:1e4, ~ simulation(.x))
toc()
```

```
## 431.58 sec elapsed
```

Agora, calculamos as probabilidades.

```
final <- do.call(bind_rows, result)
final %>% filter(serbia == '+') %>% summarise(mean(classified))
```

```
## # A tibble: 1 x 1
##   `mean(classified)`
##                <dbl>
## 1              0.879
```

```
final %>% filter(switzerland == '+') %>% summarise(mean(classified))
```

```
## # A tibble: 1 x 1
##   `mean(classified)`
##                <dbl>
## 1              0.850
```

```
final %>% filter(cameroon == '+') %>% summarise(mean(classified))
```

```
## # A tibble: 1 x 1
##   `mean(classified)`
##                <dbl>
## 1              0.876
```

Logo, o jogo mais decisivo é o da Sérvia, com probabilidade de classificação de mais de 88%, caso vencido pelo Brasil.

c) Qual era a probabilidade do Brasil ser campeão, em uma final contra a Argentina, tendo se classificado em primeiro do grupo? Para responder ao item, gere 10 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.