Toán rời rạc và thuật toán
Đại học Khoa học Tự nhiên
Khoa Toán - Cơ - Tin học
Khoa học dữ liệu K4
Tháng 8 năm 2022

Bài tập số 1

Nguyễn Mạnh Linh, Nguyễn Thị Đông, Triệu Hồng Thúy

1 Bài 1

Xin chào, đây là Linh

2 Bài 2

2.1 Chia để trị

2.1.1 Ý tưởng

Phương pháp chia để trị dựa trên 2 thao tác chính:

- Chia (*devide*): phân rã bài toán ban đầu thành các bài toán con có kích thước nhỏ hơn, có cùng cách giải.
- Trị (*conque*): giải từng bài toán con (theo cách tương tự bài toán đầu - đệ qui) rồi tổng hợp các lời giải để nhận kết quả của bài toán ban đầu.

Việc "Phân rã": thực hiện trên miền dữ liệu (chia miền dữ liệu thành các miền nhỏ hơn tương đương 1 bài toán con)

2.1.2 Mô hình và lược đồ

Xét bài toàn P trên miền dữ liệu R.
Gọi $D\_C(R)$ là thuật giải P trên miền dữ liệu R.
Nếu R có thể phân rã thành n miền con: $R = R_1 \cup R_2 \cup ... \cup R_n$
Với R_0 là miền đủ nhỏ để $D\_C(R)$ có lời giải, ta có lược đồ giải thuật chia để trị như sau:

```
Divide_Conque(R):
    if(R = R_0):
        solve Divide_Conque(R_0)
    else
        divide R to R_1, R_2, ..., R_n
        for (i = 1, 2, ..., n):
            Divide_Conque(R_i)
```

```
            Combine and get result
    end
```

2.1.3 Phân tích và đánh giá

Để phân tích và đánh giá độ phức tạp của thuật toán, ta thực hiện 2 công đoạn

- Xây dựng công thức truy hồi đánh giá độ phức tạp thuật toán
- Giải công thức truy hồi xác định độ phức tạp thuật toán.
  - Phép thế liên tiếp
  - Sử dụng định lí chính

2.1.4 Ví dụ

Ta xét bài toán *tìm kiếm nhị phân trên một mảng được sắp xếp*.

- Cho dãy n phần tử được sắp theo thứ tự (*tăng dần*) và một giá trị x bất kỳ. Kiểm tra xem phần tử x có trong dãy không?
- Phân tích ý tưởng: so sánh giá trị x với phần tử giữa của dãy tìm kiếm. Dựa vào giá trị này sẽ quyết định giới hạn tìm kiếm ở bước kế tiếp là nửa trước hay nửa sau dãy.
- Lược đồ của thuật toán như sau:

```
BinarySearch(a, x, L, R):
    // Search element x in array a from position L to R
    if (L = R):
        return (x = a_L ? L : -1)
    else
        M = (L + R)/2
        if (x = a_M)
            return (M)
        else
            if (x < a_M)
                BinarySearch(a, x, L, R)
            else
                BinarySearch(a, x, M + 1, R)
            endif
        endif
    endif
end
```

Tính đúng của thuật toán
Ta chứng mình bằng quy nạp như sau

- Cơ sở quy nạp: $n = R - L + 1 = 1$ (dãy có 1 phần tử)

  - Câu lệnh `return (x = a_L ? L : -1)` trả về giá trị L hoặc -1

- Giả thiết quy nạp: Thuật toán đúng với mọi dãy có độ dài $n = R - L + 1$. Hay hàm `BinarySearch(a, x, L, R)` trả về đúng kết quả tìm kiếm x với mọi dãy có đội dài $1 \leq n' \leq n = R - L + 1$

- Tổng quát: Chứng minh thuật toán đúng với $n + 1 = R - L + 2$

  - Đặt $M = (L + R + 1)/2$, ta có $L \leq M \leq R$
  - Nếu $x = a_M$ thì kết quả trả về là M: đúng
  - Nếu $x < a_M$ thì kết quả là của bài toán tìm x trong tập $a_L, ..., a_M$. Theo giả thiết quy nạp thì `BinarySearch(a, x, L, R)` đúng vì $1 \leq M - L + 1 = (R - L + 1)/2 + 1 \leq R - L + 1$
  - Tương tự với $x > a_M$

Độ phức tạp của thuật toán

$$T(n) = \begin{cases} 1 & \text{when } n = 1 \\ T(n/2) + 1 & \text{when } n > 1 \end{cases}$$

Do đó $T(n) = O(\log n)$
Source code: `https://github.com/batman0911/dma_homework/blob/master/hw_01/src/main.ipynb`

This homework answers the problem set sequentially.

1. *Download the US Presidential Elections data set* `uspresidentialelections.dta` *from the course ILIAS site. Load the data set in* `R`.

   ```
   Copy your R Code to answer the question here.
   ```

2. *Describe the dataset. What variables does it contain? How many observations are there? What time span does it cover?*

   Please type your answer here.

   ```
   Put the right R command here.
   ```

3. *Compute measures of central tendency and variability of the variables* `vote` *and* `growth` *using* `R`*. Use the numerical measures of central tendency and variability discussed in class. Describe them in your own words and make a nice table. Plot the distribution of both variables using a boxplot and histogram. Make sure to make your plots as nice-looking as possible. Especially, include a title and label the axes.*

   Your answer goes here

   ```
   3 R commands here
   ```

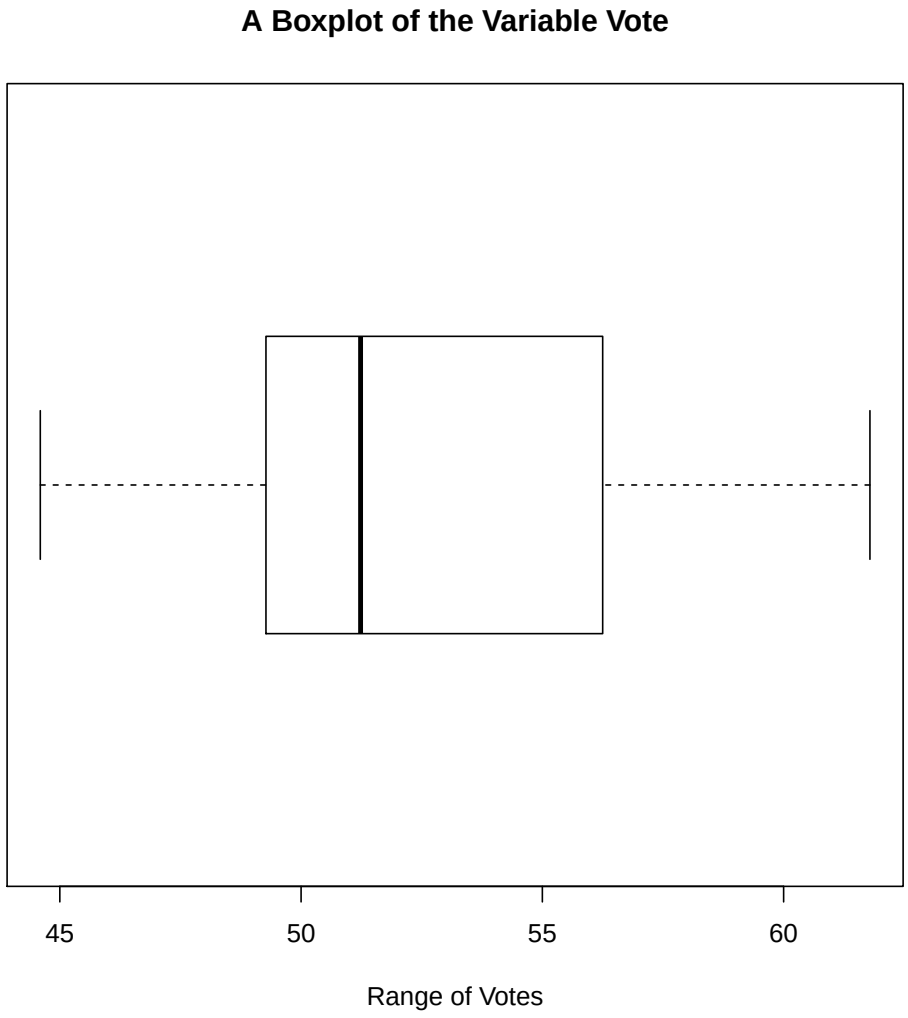
   Potentially, your answer continues here.

   ```
   4 R commands here
   ```

   And more of your answer here.

   ```
   And more space for your R commands.

   # This is the code to produce the first boxplot.
   pdf(file = "box1.pdf")
   boxplot(us_data$vote, horizontal = T,
           main  = "A Boxplot of the Variable Vote",
           names = "Vote",
           xlab  = "Range of Votes")
   dev.off()
   ```



Hình 1: Boxplot of Incumbent Vote share

Variable	*Mean*	*Median*	*Mode*	*Var*	*SD*	*Range*	*IQR*
Vote	x	x	x	x	x	x	x
Growth	x	x	x	x	x	x	x

Bảng 1: Measures of central tendency and variability.

4. *Make a bar plot of the party affiliation of incumbent presidential candidates.* Include the code for the bar plot and the plot here.

5. *During the presidential campaign in 1992, Bill Clinton's campaign coined the phrase "It's the economy, stupid!" Let's investigate the relationship between the economy and electoral success. Generate a nice-looking scatterplot of economic growth and vote share. Label the data points with the year of the election. Describe the pattern that you see in your own words.*

   Include the code for the scatterplot as well as the plot here.

   Then, describe the pattern you see. In the scatterplot we can see that...

R-Code

Finally, copy and paste the entire script here.