ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

CẤU TRÚC RỜI RẠC CHO KHMT (CO1007)

Ứng dụng thống kê
khảo sát kết quả của bài tập online cho phép nộp bài nhiều lần

GVHD: Huỳnh Tường Nguyên
Trần Tuấn Anh
Nguyễn Ngọc Lễ
SV thực hiện: Nguyễn Văn A – 22102134
Trần Văn B – 88471475
Lê Thị C – 36811334
Phạm Ngọc D – 97501334
Kiều Thị E – 12341334

Tp. Hồ Chí Minh, Tháng 05/2020

Mục lục

1 Động cơ nghiên cứu

Trong mùa dịch Covid-19, trường Đại học Bách Khoa, DDHQF-HCM đã triển khai giảng dạy trực tuyến và yêu cầu sinh viên thực hiện các bài tập nhỏ để thu nhận phản hồi về việc học tập và hiểu biết của các bạn thông qua các tài nguyên online được cung cấp.

Phân tích & thống kê dữ liệu qua các lần nộp bài của sinh viên không những giúp giáo viên có những hướng đúng trong việc phát hiện ra những kiến thức mà sinh viên chưa chắc chắn, cũng như có hướng để cải thiện bổ sung phần học liệu trong tương lai để phù hợp với hơn người học.

2 Mục tiêu

Khai phá dữ liệu từ hệ thống nộp bài online có ý nghĩa quan trọng trong việc đánh giá chất lượng của sinh viên. Ngoài ra, những đánh giá kết quả nộp bài của từng sinh viên, hay từng bài tập sẽ góp phần xác định những điểm mạnh, điểm yếu của sinh viên để giáo viên có phương pháp phù hợp trong việc cải thiện kỹ năng của sinh viên.

Trong bài tập lớn này, các sinh viên sẽ bắt đầu với các bài toán thống kê đơn giản từ những dữ liệu được cung cấp. Qua đó, các em sẽ tìm ra những con số thú vị, có ý nghĩa đối với các dữ liệu thực tế trong quá khứ của hệ thống chấm bài online. Những kết quả mà các em tìm ra sẽ là bước khởi đầu cho việc khai phá nguồn dữ liệu của hệ thống sau này, nhằm đạt tới mục tiêu nâng cao kỹ năng lập trình, kỹ năng giải quyết vấn đề cho người học cũng như hướng tới mục tiêu cao hơn khi tích hợp với các hệ thống quản lý và cải thiện chất lượng dạy và học.

3 Mô tả dữ liệu

Đính kèm đề bài tập lớn là file score.csv trong đó chứa thông tin về điểm qua các lần nộp bài của hai bài tập lớn của các sinh viên. Trong đó:

1. *sid* là số định danh của lần nộp bài đó, không có 2 số *sid* trùng nhau
2. *tid* là mã số bài tập (file dữ liệu gồm có 2 mã bài tập là 1 và 2)
3. *uid* là mã số định danh sinh viên nộp bài, mỗi sinh viên có một uid duy nhất và không trùng với *uid* của các sinh viên khác
4. *stime* thời gian theo dạng "yyyy-MM-dd HH:mm:ss" là thời gian mà sinh viên nộp bài vào hệ thống để chấm, chú ý rằng "MM" là số chỉ tháng (01..12) và "mm" là số chỉ số phút (00..59), "HH" có giá trị 00..23
5. *score* là điểm số của bài nộp tính theo thang điểm 10 và làm tròn tới 2 chữ số lẻ

4 Nhiệm vụ

Gọi MD là mã đề riêng cho mỗi nhóm, nhóm sinh viên sẽ thực hiện các yêu cầu dưới đây với các giá trị xác định như sau:

- mã *tid* nhóm cần thực hiện là $tid_n = MD \bmod 2 + 1$
- với $MD \bmod 3$, dựa vào số dư này các nhóm sẽ làm bài tập như sau: nhóm 0 làm các bài 3,4; nhóm 1 làm các bài 3,5 và nhóm 2 làm các bài 4,5. Ngoài các bài theo từng nhóm, thì các bài 1,2,6 là các bài tập chung, tất cả các nhóm đều phải làm. Các bài sau bài 6 là các bài tự chọn, nhóm sinh viên sẽ được điểm thưởng khi làm thêm các bài này.
- trích xuất các dòng dữ liệu trong tid_n bắt đầu từ dòng dữ liệu $start_n = 1 + \lfloor \frac{MD}{6} \rfloor * 30$

1. Hãy trích xuất ra các điểm (score) của 400 lần nộp bài đầu tiên trong tập dữ liệu của bài tập ứng với tid_n bắt đầu từ dòng dữ liệu $start_n$ theo MÃ ĐỀ riêng của từng nhóm. Đây sẽ là mẫu dùng cho các câu tiếp theo ở dưới.

2. Xác định số lượng sinh viên trong tập mẫu
3. Nhóm câu hỏi liên quan đến điểm số của các sinh viên
   a) Xác định điểm số là điểm tổng của các bài làm với mỗi câu hỏi đơn vị đều có điểm tối đa là 1 điểm.
   b) Xác định điểm số thấp nhất
   c) Xác định danh sách các sinh viên có ít nhất một bài có số điểm thấp nhất
   d) Xác định phổ theo số lần nộp bài của các sinh viên có ít nhất một bài có số điểm thấp nhất
   e) Xác định điểm số tổng kết thấp nhất
   f) Xác định danh sách các sinh viên có điểm số tổng kết thấp nhất
   g) Xác định phổ theo số lần nộp bài của các sinh viên có điểm số tổng kết thấp nhất
   h) Xác định điểm số cao nhất
   i) Xác định danh sách các sinh viên có tối thiểu một bài nộp có số điểm số cao nhất
   j) Xác định phổ theo số lần nộp bài của các sinh viên có tối thiểu một bài nộp có điểm số cao nhất
   k) Xác định điểm số tổng kết cao nhất
   l) Xác định danh sách các sinh viên có điểm số tổng kết cao nhất
   m) Xác định phổ theo số lần nộp bài của các sinh viên có điểm số tổng kết cao nhất
   n) Xác định điểm số trung bình của của các sinh viên trong mẫu
   o) Xác định số lượng sinh viên có điểm số trung bình
   p) Tính trung vị mẫu, cực đại mẫu, cực tiểu mẫu của trên.
   q) Hãy đo mức độ phân tán của điểm số (xung quanh giá trị trung bình) của mẫu.
   r) Tính độ méo lệch (skewness), và độ nhọn (kurtosis) của dữ liệu trong mẫu trên.
   s) Tính tứ phân vị (quartile) thứ nhất (Q_1) và thứ ba (Q_3) của mẫu.
   t) Xác định số lượng sinh viên có điểm số nằm trong 2 mức điểm cao nhất
   u) Xác định phổ theo số lần nộp bài của các sinh viên có điểm số tổng kết ở 2 mức điểm cao nhất
   v) Xác định phổ theo số lượng sinh viên có điểm số tổng kết ở mức điểm cao thứ k với k cho trước
   w) Xác định phổ theo số lần nộp bài của các sinh viên có điểm số tổng kết ở mức điểm cao thứ k với k cho trước
4. Nhóm câu hỏi liên quan đến số lần nộp bài
   a) Xác định số lần nộp bài ít nhất
   b) Xác định danh sách các sinh viên có số lần nộp bài ít nhất
   c) Xác định phổ điểm của các sinh viên có số lần nộp bài ít nhất
   d) Xác định số lần nộp bài nhiều nhất
   e) Xác định các sinh viên có số lần nộp bài nhiều nhất
   f) Xác định phổ điểm của các sinh viên có số lần nộp bài nhiều nhất
   g) Xác định số lần nộp bài trung bình của của các sinh viên
   h) Xác định số lượng sinh viên có số lần nộp trung bình
   i) Xác định phổ theo điêm số của các sinh viên có lần nộp bài trung bình
   j) Tính trung vị mẫu, cực đại mẫu, cực tiểu mẫu của trên.
   k) Hãy đo mức độ phân tán của điểm số (xung quanh giá trị trung bình) của mẫu.
   l) Tính độ méo lệch (skewness), và độ nhọn (kurtosis) của dữ liệu trong mẫu trên.
   m) Tính tứ phân vị (quartile) thứ nhất (Q_1) và thứ ba (Q_3) của mẫu.
   n) Xác định danh sách các sinh viên nằm trong nhóm có số lần nộp bài nhiều nhì
   o) Xác định danh sách các sinh viên nằm trong nhóm có số lần nộp bài nhiều nhất hoặc nhiều nhì

p) Xác định số lượng sinh viên nằm trong nhóm có số lần nộp bài nhiều nhất hoặc nhiều nhì

q) Xác định phổ theo điểm số của các sinh viên có lần nộp bài nhiều nhất hoặc nhiều nhì

r) Xác định danh sách các sinh viên nằm trong nhóm một phần ba đầu theo thứ tự số lần nộp bài giảm dần

s) Xác định số lượng các sinh viên nằm trong nhóm một phần ba đầu theo thứ tự số lần nộp bài giảm dần

t) Xác định phổ theo điểm số của các sinh viên nằm trong nhóm một phần ba đầu theo thứ tự số lần nộp bài giảm dần

u) Xác định phổ theo điểm số của các sinh viên nằm trong k nhóm đầu mà mỗi nhóm chứa các sinh viên có cùng số lần nộp bài và các nhóm được sắp xếp theo thứ tự giảm dần của có số lần nộp bài (với k cho trước).

5. Nhóm câu hỏi liên quan đến thời gian, tần suất nộp bài của các sinh viên.

a) Với mỗi sinh viên, xác định thời gian dài nhất tính từ lần nộp bài đầu tiên đến lần nộp cuối.

b) Xác định phổ thời gian làm việc (được tính từ lần nộp bài đầu tiên đến lần nộp cuối) của các sinh viên.

c) Tần suất nộp bài được tính bằng phân số giữa khoảng thời gian tính từ lần nộp bài đầu tiên đến lần nộp cuối và số lần nộp bài.

d) Xác định danh sách các sinh viên có tần suất nộp bài ít nhất

e) Xác định phổ điểm của các sinh viên có tần suất nộp bài ít nhất

f) Xác định số lượng sinh viên có tần suất nộp bài nhiều nhất

g) Xác định các sinh viên có tần suất nộp bài nhiều nhất.

h) Xác định phổ điểm của các sinh viên có tần suất nộp bài nhiều nhất.

i) Xác định các sinh viên nằm trong nhóm có tần suất nộp bài nhiều nhì.

j) Xác định các sinh viên nằm trong nhóm có tần suất nộp bài nhiều nhất hoặc nhiều nhì.

k) Hãy tính thời gian trung bình (tính bằng giây) giữa hai lần nộp bài liền nhau của cùng một sinh viên trong mẫu đã chọn.

l) Tính tần số, tần suất và tần suất tích lũy của mẫu trên.

m) Vẽ biểu đồ tần số của mẫu trên. Hãy nhận xét về biểu đồ.

n) Vẽ biểu đồ tần suất của mẫu trên. Hãy nhận xét về biểu đồ.

o) Vẽ biểu đồ tần suất tích lũy của mẫu trên. Hãy nhận xét về biểu đồ.

p) Tính trung vị mẫu, cực đại mẫu, cực tiểu mẫu của trên.

q) Hãy đo mức độ phân tán của điểm số (xung quanh giá trị trung bình) của mẫu.

r) Tính độ méo lệch (skewness), và độ nhọn (kurtosis) của dữ liệu trong mẫu trên.

s) Tính tứ phân vị (quartile) thứ nhất (Q_1) và thứ ba (Q_3) của mẫu.

6. *Đối với câu hỏi này (và các bài liên quan ở sau câu này), nhóm sinh viên thực hiện trên tổng thể của bài tid_n, không thực hiện trên mẫu*
Gọi điểm số lần nộp bài thứ k của sinh viên i cho bài tập j là s_{ijk} với i là uid, $j \in (1, 2)$ và $k \in (1, 2, 3, ...)$. Điểm tổng hợp của sinh viên tính tới lần nộp thứ k là điểm lớn nhất cho bài tập đó mà sinh viên đạt được cho tới lần nộp thứ k, tức là:

$$score_{ijk} = max(s_{ij1}, s_{ij2}, ..., s_{ijk})$$

Đối với sinh viên nộp ít hơn k lần thì vẫn tính theo công thức với giá trị khuyết xem như là 0. Gọi TB_k là điểm trung bình của các sinh viên tính tới lần nộp thứ k.

a) Hãy tính và vẽ biểu đồ sự phân bố về điểm đạt được của sinh viên sau $k = 6$ lần nộp bài.

b) Áp dụng câu 6a với k được tính theo công thức sau:

$$((MD * 37 + 59) \ mod \ 5 + 1) * (tid_n * 2 - 3) + 6$$

c) Hãy tính các giá trị TB_k và vẽ biểu đồ thể hiện sự thay đổi của các giá trị trung bình này với sự thay đổi của k. Hãy nhận xét về biểu đồ mà các em vừa vẽ được.

d) Hãy cho biết trung bình điểm số mà các sinh viên đạt được qua bài tập tid_n này là bao nhiêu.

7. Sinh viên **siêng năng** là sinh viên có nộp bài sớm hơn thời điểm t_1.
   a) Hãy xác định thời điểm t_1 phù hợp.
   b) Xác định số lượng sinh viên siêng năng.
   c) Xác định phổ điểm của các sinh viên siêng năng.
8. Sinh viên học **đối phó** là sinh viên có nộp bài lần đầu tiên trễ hơn thời điểm t_2.
   a) Hãy xác định thời điểm t_2 phù hợp.
   b) Xác định số lượng sinh viên học đối phó.
   c) Xác định phổ điểm của các sinh viên học đối phó.
9. Sinh viên học **giỏi** là sinh viên có kết quả nằm trong top x % điểm cao nhất.
   a) Hãy xác định giá trị x phù hợp.
   b) Xác định số lượng sinh viên học giỏi.
   c) Xác định phổ điểm của các sinh viên giỏi.
10. Sinh viên **thông minh** là sinh viên có kết quả tốt (điểm lớn hơn k) ngay từ n lần nộp đầu tiên.
    a) Hãy xác định giá trị k và n phù hợp.
    b) Xác định số lượng sinh viên thông minh.
    c) Xác định phổ điểm của các sinh viên thông minh.
11. Sinh viên học **chủ động** là sinh viên thông minh hoặc sinh viên siêng năng mà có nộp bài nhiều lần để cải thiện điểm.
    a) Hãy xác định các thông số phù hợp.
    b) Xác định số lượng sinh viên biết cách học chủ động.
    c) Xác định phổ điểm của các sinh viên biết cách học chủ động.
12. Xác định phần giao của các loại sinh viên đánh giá ở trên (từ câu 7 đến câu 11) và vẽ biểu đồ thống kê minh họa.
13. (*Điểm thưởng*) Nhóm có thể tự đề xuất và bổ sung thêm những giá trị thống kê hữu ích đối với tập dữ liệu điểm này.

5 Hướng dẫn và yêu cầu

5.1 Hướng dẫn

- Cài đặt đồng thời cả R và Rstudio.
- Đọc kĩ và xử lý lại tất cả những thí dụ đã có trong file mẫu.
- Tìm hiểu kĩ cách soạn thảo văn bản bằng LaTeX và cách sử dụng phần mềm R trong các file hướng dẫn và tìm hiểu thêm trong các tài liệu khác.
- Tạo một folder chung chứa mọi thứ cần thiết để share giữa các thành viên trong nhóm trên các cloud services như Google Drive hay Dropbox,...
- Dùng Doodle để lên kế hoạch họp nhóm.
- Dùng Trello để quản lý project.

5.2 Yêu cầu

Mỗi nhóm, từ 3 đến 4 sinh viên, đề xuất giải pháp. Nhóm cần nộp báo cáo trình bày về lời giải cho các câu hỏi và kết quả thực nghiệm. Đồng thời, nhóm cũng cần nộp source code, và trình bày các kết quả của nhóm trong khoảng 5 minutes.

Báo cáo và slide trình bày cần được viết dưới dạng LaTeX.

- Thời gian làm bài: Từ 22/05/2019 – 12/06/2020.

  Đối với mỗi bài toán, yêu cầu sinh viên trình bày lời giải theo lối truyền thống, sử dụng các công thức, kết quả lý thuyết trong phần kiến thức chuẩn bị. Đồng thời, sau đó trình bày kết quả tính toán và biểu đồ minh họa bằng R.

- Trình bày cả code R và kết quả tính toán trong R giống như file mẫu.

- Viết báo cáo theo đúng **bố cục như trong file mẫu** bằng LaTeX.

- Mỗi nhóm khi nộp bài **cần phải nộp theo file log (nhật ký)** ghi rõ: tiến độ công việc, phân công nhiệm vụ, trao đổi của các thành viên,...

5.3 Nộp bài

- SV chỉ nộp bài qua hệ thống Sakai: nén tất cả các file cần thiết (file .tex, file .R, ...) thành một file tên là "*BTL-CO1007-MT161-Nhom-n.zip*" và nộp trong mục Assignment.

- Lưu ý: mỗi nhóm **chỉ cần một thành viên là nhóm trưởng nộp bài.**

6 Cách đánh giá và xử lý gian lận

6.1 Đánh giá

Mỗi bài làm sẽ được đánh giá như sau.

Nội dung	Tỉ lệ điểm (%)
Giải đúng các bài toán bằng công thức và lập luận	30%
Các lệnh (hàm) R được sử dụng đúng đắn và hợp lý	30%
Trình bày kiến thức chuẩn bị rõ ràng, phù hợp	20%
Trình bày văn bản đẹp, đúng chuẩn	20%

6.2 Xử lý gian lận

Bài tập lớn phải được sinh viên (nhóm) TỰ LÀM. Sinh viên (nhóm) sẽ bị coi là gian lận nếu:

- Có sự giống nhau bất thường giữa các bài thu hoạch (nhất là phần kiến thức chuẩn bị). Trong trường hợp này, TẤT CẢ các bài nộp có sự giống nhau đều bị coi là gian lận. Do vậy sinh viên (nhóm) phải bảo vệ bài làm của mình.

- Sinh viên (nhóm) không hiểu bài làm do chính mình viết. Sinh viên (nhóm) có thể tham khảo từ bất kỳ nguồn tài liệu nào, tuy nhiên phải đảm bảo rằng mình hiểu rõ ý nghĩa của tất cả những gì mình viết.

Bài bị phát hiện gian lận thì sinh viên sẽ bị xử ý theo quy định của nhà trường.

Tài liệu

[Dal] Dalgaard, P. *Introductory Statistics with R.* Springer 2008.

[K-Z] Kenett, R. S. and Zacks, S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP,* 2nd ed., John Wiley and Sons, 2014.

[Ker] Kerns, G. J. *Introduction to Probability and Statistics Using R,* 2nd ed., CRC 2015.