Thi giữa kỳ học phần KHDL

Lớp HP: 1911

Thời gian thi: 7h15 thứ tư ngày 13/4/2022

Đề thi

Câu 1 (3 điểm)

Viết chương trình cài đặt phương pháp Bootstrap để xuất ra phân bố lấy mẫu và dải tin cậy của một thông số thống kê của một biến (đặc trưng) nào đó của dataset với:

- Biến (đặc trưng): do SV chọn từ dataset
- Thông số thống kê: độ lệch chuẩn
- Viết hàm có:
  - Input parameters: n (sample size), M (number of bootstrap samplings), x (confidence interval in %)
  - Return values: phân bố lấy mẫu (bằng histogram) và dải tin cậy của độ lệch chuẩn của đặc trưng đã chọn

Câu 2 (7 điểm)

Dựa trên cùng dataset với Câu 1, viết chương trình xuất ra và thuyết trình các thống kê mô tả quan trọng của các biến, các mối quan hệ giữa các biến đã có và biến tạo thêm (nếu có), quy luật tiềm ẩn có thể suy diễn từ dataset nhờ các công cụ trực quan hoá dữ liệu (ví dụ: distribution plot, heat/cluster map, linear model plot,…).

Yêu cầu

1. Dataset dùng chung cho cả 2 câu:

- Nguồn: SV tự tìm (https://www.kaggle.com/datasets, https://archive.ics.uci.edu/ml/datasets.php,...)
- **Kích thước: số lượng mẫu > 500**
- SV nhập tên dataset và link to dataset vào link GV chia sẻ sau buổi học. SV đăng ký sau ko đc đăng ký trùng lặp dataset với SV đăng ký trước.

2. Các yêu cầu khác:

- Mã nguồn:
  - Viết code với định dạng Jupyter notebook (.ipynb)
  - Tự viết code chỉ dùng các hàm built-in của Python và các thư viện Numpy, Pandas, Matplotlib và Seaborn. Các SV copy code của nhau thì sẽ nhận 0 điểm.
  - Sạch sẽ (clean), chú thích đầy đủ (1 dòng chú thích cho 1 cell)
- Thuyết trình:
  - Dùng Markdown để note ý tưởng trình bày và các kết luận trên file .ipynb
  - **Mỗi SV trình bày và demo bài làm trong tối đa 07 phút. Quá giờ sẽ bị cắt phần trình bày.**
  - Dùng laptop của GV trình bày để tiết kiệm thời gian

Hình thức nộp bài

- **Deadline to submit: 6h00 thứ tư ngày 13/4/2022** (Hết hạn SV ko nộp bài thì nhận 0 điểm)
- SV submit thư mục bài làm (ko nén) qua link GV gửi trên MS Teams 24 giờ trước deadline.
- Đặt tên thư mục bài làm theo quy tắc: **Mã SV_Họ và tên**.

  Thư mục này chứa:
  - 01 file mã nguồn .ipynb duy nhất cho cả 2 câu (Link to data source đặt ở dòng đầu tiên của notebook)
  - Các file dữ liệu (total file size < 10 MB)