Machine Learning

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Bài 1:
Cho dữ liệu ở đường link, hãy thực hiện các yêu cầu sau:

https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f

1. Đọc dữ liệu và hiển thị ra 20 mẫu đầu tiên
2. Cho biết Kích thước dữ liệu
3. Cho biết Kiểu dữ liệu
4. Cho biết số lượng mỗi lớp có trong dữ liệu
5. Thực hiện các thống kê cơ bản về các trường trong bộ dữ liệu
6. Tính mối tương quan Pearson cho từng cặp thuộc tính và trực quan hóa kết quả
7. Tính mối tương quan Spearman cho từng cặp thuộc tính và trực quan hóa kết quả
8. Tính Độ lệch (skewness) cho mỗi thuộc tính
9. Hãy sửa những dữ liệu có giá trị bằng 0 là dữ liệu thiếu (NaN)
10. Đếm số lượng giá trị NaN trên mỗi thuộc tính
11. Thay thế giá trị còn thiếu bằng giá trị trung bình của mỗi thuộc tính và lưu vào 1 file
12. Sử dụng mô hình hồi quy để dự đoán với file dữ liệu trên
13. Làm lại ý 11 áp dụng công thức sau:
    X_std = (X - Xmin)/ (Xmax - X.min)
    X_scaled = X_std * (max - min) + min
14. Thực hiện lại ý 12 với dữ liệu ở ý 13
15. So sánh kết quả đạt được ở ý 12 và 14

Bài 2:
2.1 Hãy khởi tạo điểm số là số nguyên của 100 học sinh học môn Toán (0-10) một cách ngẫu nhiên.
2.2 Cho biết khoảng tin cậy với mức ý nghĩa 5%
2.3 Hãy tính khoảng tin cậy của dữ liệu trên

Bài 3:
Trả lời cho câu hỏi Học online hay học qua sách giáo khoa cái nào tốt hơn, chúng ta tiến hành thực hiện chọn ngẫu nhiên 100 học sinh để học online và chọn ngẫu nhiên 100 học sinh để học qua sách giáo khoa. Cho 2 nhóm này học cùng một kiến thức toán nhưng từ hai nguồn khác nhau (online, sách giáo khoa) sau đó cho làm bài kiểm

tra toán trắc nghiệm để đánh giá 2 nhóm này. Hãy thực hiện các yêu cầu sau, cho biết công thức tính theo mean như sau:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
$$UCL = (\hat{\mu}_1 - \hat{\mu}_1) + (Zscore * SE)$$
$$LCL = (\hat{\mu}_1 - \hat{\mu}_1) - (Zscore * SE)$$

3.1 Khởi tạo mỗi nhóm 100 sinh viên.
3.2 Cho nhóm học Online làm 30 câu kiểm tra với mỗi câu 1 điểm, sinh điểm ngẫu nhiên và in ra kết quả của nhóm này.
3.3 Cho nhóm học qua sách giáo khoa làm 30 câu kiểm tra với mỗi câu 1 điểm, sinh điểm ngẫu nhiên in ra kết quả của nhóm này.
3.4 Tính điểm trung bình đạt được của mỗi nhóm
3.5 So sánh sự khác biệt về điểm số trung bình của 2 nhóm này
3.6 Tính độ lệch chuẩn std của mỗi nhóm
3.7 Tính standard error giữa 2 nhóm lấy mẫu ở trên
3.8 Tính khoảng tin cậy với mức ý nghĩa 1%
Hãy thực hiện kiểm định các giả thuyết sau:
3.9 Học Online tốt hơn học qua sách giáo khoa
3.10 Học Online không có gì khác học qua sách giáo khoa

Bài 4:
Làm lại bài 3 nếu áp dụng theo tỷ lệ với công thức sau:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$
$$UCL = (\hat{p}_1 - \hat{p}_1) + (Zscore * SE)$$
$$LCL = (\hat{p}_1 - \hat{p}_1) - (Zscore * SE)$$

Bài 5:
A/B testing (hay còn được gọi là split testing) là một quy trình mà trong đó hai phiên bản (A và B) sẽ được cùng so sánh trong một môi trường / tình huống được xác định và qua đó đánh giá xem phiên bản nào hiệu quả hơn.
Quy trình khi chạy A/B testing:
-> Thu thập mẫu (gather samples)
-> Chia nhóm (assign buckets)

-> Áp dụng thử nghiệm (apply treatments)
-> Đo lường kết quả (measure outcomes)
-> So sánh kết quả (make comparisons)
Hãy thực hiện lại bài 3 nếu tổng số mẫu là 300 theo cách học theo sách giáo khoa theo các tỷ lệ 1:3 và 1:2

Bài 6:
Chi-square: còn gọi là goodness-of-fit test: dùng để test các thống kê có phân hoạch theo chu kỳ hay nhóm theo công thức:

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

6.1 Hãy giả lập dữ liệu Báo cáo lượt view trong tuần từ phía Facebook, trực quan hóa dữ liệu này
6.2 Hãy giả lập dữ liệu Báo cáo lượt view trong tuần từ phía Google, trực quan hóa dữ liệu này.
6.3 Sử dụng phân tích Chi-square để kiểm tra xem Báo cáo lượt view trong tuần từ phía Facebook có khác Google không với mức ý nghĩa 2%?

Bài 7:
Phân tích phương sai (analysis of variance-ANOVA) là phương pháp thống kê để phân tích tổng quy mô biến thiên của biến số phụ thuộc (tổng đó tổng quy mô biến thiên được định nghĩa là tổng các độ lệch bình phương so với số bình quân của nó) thành nhiều phần và mỗi phần được quy cho sự biến thiên của một biến giải thích cá biệt hay một nhóm các biến giải thích. Phần còn lại không thể quy cho biến nào được gọi là sự biến thiên không giải thích được hay phần dư. Phương pháp này được dùng để kiểm định giả thuyết không nhằm xác định xem các mẫu thu được có được rút ra từ cùng một tổng thể không. Kết quả kiểm định cho chúng ta biết các mẫu thu được có tương quan với nhau hay không.
Phân tích phương sai một yếu tố (còn gọi là oneway anova) dùng để kiểm định giả thuyết trung bình bằng nhau của các nhóm mẫu với khả năng phạm sai lầm chỉ là 5%.
Một số giả định khi phân tích ANOVA:
– Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.
– Các nhóm so sánh phải có phân phối chuẩn or cỡ mẫu phải đủ lớn để được xem như tiệm cận phân phối chuẩn.
– Phương sai của các nhóm so sánh phải đồng nhất.
Một số công thức:

- SST: Total sum of squares

- SSW: Sum of squares within
- SSB: Sum of squares between
- f-statistics

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

$$SSB = n\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2$$

$$SST = SSW + SSB$$

$$F - stat = \frac{\frac{SSB}{m-1}}{\frac{SSW}{n_i - m}}$$

Hãy thực hiện các yêu cầu sau:
7.1 Xây dựng dữ liệu đánh giá phim cho 10 bộ phim với rate từ 1-5 của 20 người một cách tự động
7.2 Tính SST
7.3 Tính SSW
7.4 Tính SSB
7.5 f-statistic
7.6 Hãy cho biết Đánh giá các bộ phim có giống nhau hay không với mức ý nghĩa 5%?