TRUNG TÂM TIN HỌC ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HỒ CHÍ MINH

Đề thi cuối khóa: (gồm có 2 trang)

MATHEMATICS AND STATISTICS FOR DATA SCIENCE

Ngày thi: 03/11/2019

Thời gian : 120 phút

*** Học viên (HV) tạo một thư mục có tên là *HoVaTen*, lưu tất cả bài làm để nộp chấm điểm ***

*** HV được sử dụng tài liệu ***

*** HV sẽ bị trừ điểm nếu làm bài giống nhau ***

Lưu ý:

- Bài làm của mỗi câu được lưu trong 1 file (đặt tên: *Cau1.ipynb*, …), viết bằng ngôn ngữ Python trên *jupyter notebook*, và các nhận xét về kết quả được viết trong cell với định dạng Markdown.
- Cần hiển thị thông tin chung của dữ liệu bằng cách dùng shape, head(), tail(), info()… để có cái nhìn ban đầu về dữ liệu trước khi thực hiện những yêu cầu.

Câu 1. Giảm chiều dữ liệu (2 điểm)

Tập tin ***Phan_lop.csv*** chứa những mẫu dữ liệu phân lớp các đối tượng thuộc về một trong 6 loại (class): 0..5, dựa trên 12 thuộc tính f1..f12 của đối tượng.

1.1) Áp dụng phương pháp giảm chiều của dữ liệu để số chiều thấp hơn so với dữ liệu gốc.

1.2) Số chiều được giảm còn lại bao nhiêu ? Giải thích nguyên nhân/cơ sở về số chiều được giảm?

Câu 2. Thống kê – Xác suất (6 điểm)

Tập tin ***IQ.xls*** chứa những mẫu dữ liệu được thu thập về mối quan hệ giữa chỉ số IQ và điểm thi môn Toán (diemToan) của sinh viên. Người ta muốn biết liệu rằng điểm thi môn Toán có thể được sử dụng để dự đoán chỉ số IQ của sinh viên hay không.

2.1) Đọc và xem thông tin của dữ liệu.

2.2) Vẽ biểu đồ phân phối tần suất của diemToan. Nhận xét kết quả.

2.3) Thực hiện các thống kê cơ bản cho diemToan và IQ (mean, median, mode, max, min, range).

2.4) Cho biết các giá trị ở phân vị thứ 20, 35, 65 và 90 của IQ. Biểu diễn phân vị và giá trị tương ứng trên biểu đồ.

2.5) Vẽ boxplot cho diemToan và cho IQ.

2.6) diemToan có outlier(s) hay không ? IQ có outlier(s) hay không ?

2.7) Tìm phương sai (variance) của diemToan.

2.8) Tìm độ lệch chuẩn (standard deviation) của IQ.

2.9) Tìm độ xiên (skewness) của diemToan. Nhận xét kết quả.

2.10) Tìm độ nhọn (kurtosis) của diemToan. Nhận xét kết quả.

2.11) Cho biết số lượng mẫu có giá trị IQ > 130. Xác suất các mẫu có IQ lớn hơn 130 là bao nhiêu ?

2.12) Tìm xác suất của $P(85 \leq IQ \leq 130)$.

2.13) Vẽ biểu đồ thể hiện mối quan hệ giữa diemToan và IQ. Nhận xét kết quả.

2.14) Tính giá trị tương quan giữa diemToan và IQ.

2.15) Dựa vào thể hiện dữ liệu trực quan và giá trị tương quan ở trên, hãy cho biết có thể dựa trên diemToan để dự đoán giá trị của IQ hay không ? Giải thích nguyên nhân.

2.16) Giả sử có thể dựa trên diemToan để dự đoán giá trị của chỉ số IQ. Xây dựng hệ phương trình $y = mx + b$ (với y là IQ và x là diemToan).

2.17) Tìm m và b.

2.18) Từ m và b, hãy tính toán lại các chỉ số IQ trong mẫu dữ liệu. Trực quan hóa dữ liệu.

2.19) Tính các giá trị IQ tương ứng với diemToan lần lượt là 2.0, 5.0, 8.0, 9.5.

Câu 3. Kiểm định giả thuyết (2 điểm)

Hai mẫu dữ liệu độc lập được thu thập từ các quần thể, *không biết trước phương sai*, và lưu trữ trong các tập tin Mau_1.txt và Mau_2.txt.

3.1) Đọc và xem thông tin của dữ liệu.

3.2) Với $\alpha = 0.05$, hãy cho kết luận về giả thuyết vô hiệu H_0: "Hai mẫu có cùng giá trị trung bình" bằng 2 phương pháp:

a) Tính toán truyền thống,

b) Dùng các hàm thống kê có sẵn.

--- Chúc các HV làm bài tốt ☺ ---