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: 08/12/2019

Thời gian : 120 phút

Lưu ý:

- Lưu bài làm của mỗi câu trong 1 file riêng (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.
- Nén tất cả bài làm vào 1 file .RAR (hay .ZIP) với cách đặt tên: <tên>, <Họ>.RAR
  VD: **Anh, TranTuan.RAR**

Câu 1. Giảm chiều dữ liệu (3 điểm)

Tập tin ***Phan_lop.csv*** chứa những mẫu dữ liệu phân lớp cho các đối tượng thuộc về một trong 6 lớp (class): 0..5, dựa trên các thuộc tính f1, f2, …, f12 của đối tượng.

1.1) Áp dụng phương pháp PCA để giảm xuống k chiều so với dữ liệu gốc ($2 < k < 12$).
Giải thích nguyên nhân hay cơ sở về số chiều được giảm?

1.2) So sánh độ biến thiên của dữ liệu TRƯỚC và SAU khi giảm chiều.

1.3) Giảm chiều xuống còn $k = 2$ và trực quan hóa dữ liệu. Nhận xét kết quả.

Câu 2. Thống kê – Xác suất (5 điểm)

Tập tin ***IQ2.xls*** chứa những mẫu dữ liệu được thu thập về mối quan hệ giữa chỉ số IQ với các điểm thi môn Toán (diemToan) và môn Anh văn (diemAV) của sinh viên.

2.1) Thể hiện những thông tin của dữ liệu. Vẽ biểu đồ phân phối tần số của diemToan và diemAV.

2.2) Tính các giá trị mean, median và variance của diemToan, diemAV và IQ.

2.3) Vẽ các biểu đồ boxplot cho diemToan, diemAV và IQ.

2.4) Xác định outlier(s), nếu có, của diemToan, diemAV và IQ dựa trên các z-scores.

2.5) Vẽ biểu đồ thể hiện mối quan hệ giữa diemToan và IQ. Tính giá trị tương quan giữa diemToan và IQ. Nhận xét kết quả.

2.6) Vẽ biểu đồ thể hiện mối quan hệ giữa diemAV và IQ. Tính giá trị tương quan giữa diemAV và IQ. Nhận xét kết quả.

2.7) Dựa vào các kết quả của 2.5) và 2.6), hãy chọn diemToan hoặc diemAV để dự đoán giá trị của IQ. Giải thích nguyên nhân.

2.8) Gọi x là diemToan hoặc diemAV đã chọn ở câu 2.7). Xây dựng hệ phương trình $y = mx + b$ (với y là IQ). Tìm m và b.

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

2.10) Tính các giá trị IQ tương ứng với $x \in \{ 2.0, 5.0, 8.0, 9.5 \}$.

Câu 3. Kiểm định giả thuyết (2 điểm)

Hai tập tin Mau_1.txt và Mau_2.txt lưu trữ hai mẫu dữ liệu PHỤ THUỘC.

3.1) Đọc và xem thông tin của dữ liệu.

3.2) Với $\alpha = 0.05$, hãy cho kết luận về giả thuyết H_0: "*Hai quần thể có cùng giá trị trung bình*" bằng 2 phương pháp:
a) Tính toán truyền thống, và
b) Dùng các hàm thống kê có sẵn.

--- Chúc các HV làm bài tốt ☺ ---