Ôn tập cuối khóa:

MATHEMATICS AND STATISTICS FOR DATA SCIENCE

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, Tran Tuan.RAR**

Câu 1. Giảm chiều dữ liệu

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 (**2 < k < 12**) so với dữ liệu gốc. Giải thích nguyên nhân (hay cơ sở) về số chiều được giảm.

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

Câu 2. Hồi quy tuyến tính

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) Tính các giá trị thống kê cơ bản của chỉ số IQ và điểm thi của các môn.

2.2) Xác định outlier(s), nếu có, của chỉ số IQ và điểm thi của các môn dựa trên các z-scores.

2.3) Chọn điểm thi của 1 trong các môn làm cơ sở để dự đoán chỉ số IQ theo phương pháp hồi quy tuyến tính. Trực quan hóa dữ liệu và giải thích nguyên nhân của sự lựa chọn.

2.4) Dự đoán các giá trị IQ tương ứng với x ∈ { 2.0, 5.0, 8.0, 9.5 }.

Câu 3. Kiểm định giả thuyết

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 H_0: "*Hai quần thể có cùng giá trị trung bình*".

Câu 4. Kiểm định ANOVA

Tập tin '***One way ANOVA.txt***' lưu trữ bốn mẫu dữ liệu A, B, C và D.

4.1) Đọc và xem thông tin của dữ liệu.

4.2) Với $\alpha = 0.05$, hãy cho kết luận về giả thuyết H_0: "*Các quần thể có cùng giá trị trung bình.*".

--- Hết ---