Đề thi cuối khóa: **K262** (gồm có 2 trang)

MATHEMATICS AND STATISTICS FOR DATA SCIENCE

Ngày thi: 10/01/2021

Thời gian : 180 phút

Lưu ý:

- Lưu bài làm của mỗi câu trong 1 file riêng (đặt tên: *Cau1.ipynb*, …), viết bằng ngôn ngữ Python 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**
- Bài làm sẽ bị trừ điểm nếu không thực hiện đầy đủ những yêu cầu nêu trên.

Câu 1. Giảm chiều dữ liệu (2 điểm)

Tập tin '***Breast Cancer.csv***' chứa dữ liệu phân lớp bệnh nhân ung thư (*Classification* $\in \{1, 2\}$) dựa trên các thuộc tính: *Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin* và *MCP.1*.

1.1) Áp dụng phương pháp PCA để giảm xuống còn k chiều (**k > 2**) so với dữ liệu gốc. Giải thích nguyên nhân (hay cơ sở) của 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 (4 điểm)

Tập tin '***IQ4.xls***' chứa những mẫu dữ liệu được thu thập về mối quan hệ giữa chỉ số IQ với điểm thi các môn học của sinh viên.

2.1) Thể hiện những thông tin của dữ liệu. Vẽ biểu đồ phân phối tần số điểm thi của các môn.

2.2) Dùng các hàm để tính các giá trị thống kê cơ bản của chỉ số IQ và điểm thi của các môn.

2.3) Xác định outlier(s), nếu có, của chỉ số IQ và điểm thi của các môn dựa trên quy tắc 3-Sigma.

2.4) Chọn điểm thi của **một** trong các môn để làm cơ sở cho việc dự đoán chỉ số IQ theo phương pháp hồi quy tuyến tính bằng Gradient Descent và bằng ma trận giả nghịch đảo. Trực quan hóa dữ liệu và giải thích nguyên nhân của sự lựa chọn.

2.5) Dự đoán các chỉ số IQ tương ứng với x $\in$ { 0.5, 1.0, 1.5, 2.0, …, 9.0, 9.5, 10 }.

Câu 3. Thống kê mô tả (1 điểm)

Một vận động viên bơi lội 200m hỗn hợp có thành tích như sau:

50m bơi bướm với vận tốc 1.92m/s,

50m bơi ngửa với vận tốc 1.67m/s,

50m bơi ếch với vận tốc 1.56m/s, và

50m bơi tự do với vận tốc 1.85m/s.

Hãy cho biết vận tốc bơi trung bình của vận động viên.

Câu 4. Kiểm định trung bình 2 mẫu (1 điểm)

Hai tập tin Duong_huyet_TRUOC.txt và Duong_huyet_SAU.txt lưu trữ hai mẫu dữ liệu về chỉ số đường huyết (mg/dL) của các bệnh nhân được đo trước và sau khi sử dụng thử nghiệm một loại thuốc mới T của hãng dược phẩm 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: "*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.

Câu 5. Kiểm định ANOVA (2 điểm)

Tập tin '***One way ANOVA.txt***' lưu trữ bốn mẫu dữ liệu A, B, C và D được lấy từ các quần thể đều có phân phối chuẩn.

5.1) Với $\alpha = 0.05$, hãy kiểm định giả thuyết H_0: "*Các quần thể có cùng phương sai.*"

5.2) Với $\alpha = 0.05$, hãy cho kết luận về giả thuyết H_0: "*Các 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 ---