Đề thi cuối khóa: **K268** (gồm có 2 trang)

MATHEMATICS AND STATISTICS FOR DATA SCIENCE

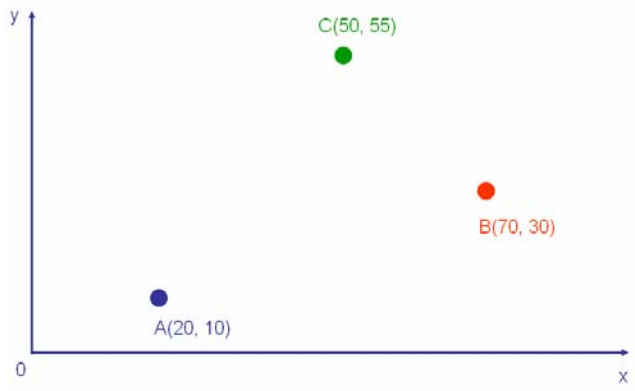
Ngày thi: 06/09/2021

Thời gian : 180 phút

Lưu ý:

- Sử dụng ngôn ngữ Python.
- Loại bỏ các lệnh hay hàm các không cần thiết trong bài làm, ngoại trừ print(), show(), head(), …
- Lưu bài làm của mỗi câu trong 1 file riêng (đặt tên: *Cau1*, *Cau2*, …).
- Nén tất cả bài làm vào 1 file .RAR (hay .ZIP) với cách đặt tên: <Tên>, <Họ>.RAR
  VD: **Anh, TranTuan.RAR** (*không tạo thư mục*)
- Bài làm sẽ bị trừ điểm nếu không thực hiện đầy đủ những yêu cầu nêu trên.

Câu 1. Vectơ (1 điểm)



Cho 3 điểm A, B, C có tọa độ trong mặt phẳng xOy như trong hình vẽ.

1.1) Trong 2 điểm B và C, điểm nào gần với điểm A hơn ? Tại sao ?

1.2) Giả sử cho phép di chuyển theo các hướng song song với các trục Ox, Oy và song song với đường phân giác của xOy (45^o). Tính quãng đường di chuyển từ điểm A đến điểm B.

Câu 2. Giảm chiều dữ liệu (2 điểm)

Tập tin '***Breast Cancer WI.csv***' chứa dữ liệu phân lớp bệnh nhân ung thư (*Class* ∈ {Benign, Malignant}) dựa trên 30 thuộc tính.

2.1) Vẽ biểu đồ phân tích phương sai tích lũy theo sự biến thiên của số chiều **k**. Dựa vào biểu đồ, chọn giá trị cho **k** để giảm chiều, với **k > 2**. Giải thích lý do đã chọn giá trị cho số chiều **k**.

2.2) Chọn giá trị **k** nhỏ nhất để phương sai tích lũy đạt tối thiểu là 99.98%.

2.3) Trực quan hóa dữ liệu với số chiều **k = 2**.

2.4) Nhận xét kết quả phân lớp. Có thể cải thiện kết quả phân lớp ?

Câu 3. Hồi quy tuyến tính (3 điểm)

Tập tin '***IQ6.xls***' chứa dữ liệu huấn luyện về mối quan hệ giữa chỉ số IQ với điểm thi của các môn học.

3.1) Vẽ biểu đồ phân phối tần số và cho biết những giá trị thống kê cơ bản của điểm thi các môn. Xác định outlier(s), nếu có, của chỉ số IQ và điểm thi của các môn dựa trên quy tắc **3-Sigma**.

3.2) Tính hệ số tương quan Pearson giữa IQ và các điểm thi. Trên cơ sở đó, chọn điểm thi của 1 trong các môn để dự đoán chỉ số IQ theo phương pháp hồi quy tuyến tính bằng a) *Gradient Descent* **VÀ** b) *Ma trận giả nghịch đảo*. Trực quan hóa dữ liệu.

3.3) Dự đoán chỉ số IQ cho tập dữ liệu thử nghiệm (test set): { 0.5, 1.0, 1.5, 2.0, …, 9.0, 9.5, 10 }.

3.4) Nhận xét kết quả khi dùng diem_5 để dự đoán chỉ số IQ.

Câu 4. 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,

50m bơi tự do với vận tốc 1.85m/s,

Hãy dùng hàm tính giá trị trung bình của Python để tính vận tốc trung bình của vận động viên.

Câu 5. Kiểm định trung bình 2 mẫu (1 điểm)

Hai tập tin Duong_huyet_TRUOC.txt và Duong_huyet_SAU.txt lưu trữ 2 mẫu dữ liệu về chỉ số đường huyết (mg/dL) của các bệnh nhân được đo trước và sau khi sử dụng loại thuốc T.

5.1) Đọc và xem thông tin của dữ liệu.

5.2) 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 6. Kiểm định ANOVA (2 điểm)

Tập tin '***Samples.txt***' lưu trữ 4 mẫu dữ liệu được lấy từ các quần thể đều có phân phối chuẩn.

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

6.2) 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*.

6.3) Nếu bác bỏ giả thuyết H_0 trong câu 6.2), hãy cho biết những quần thể nào có sự khác biệt về giá trị trung bình.

--- HẾT ---