
 <p>TRƯỜNG ĐH GTVT KHOA CNTT BỘ MÔN: MẠNG VÀ CÁC HTTT</p>	<p>ĐỀ THI KẾT THÚC HỌC PHẦN</p> <p>HỌC PHẦN: KHAI PHÁ DỮ LIỆU</p> <p>THỜI GIAN: 60 PHÚT</p>	<p>Trưởng Bộ Môn</p> 
--	--	---

Câu 1 (6 điểm). Cho tập dữ liệu thu thập được của 2 tập giá trị của **X** và **Y** như sau:

X	7	4.5	5.5	6.5	6.1	5.3	5	5.1	4.9	4.6	5.7	5.4
Y	3.2	2.3	3.5	2.8	2.8	3.7	3.3	3.5	3	3.1	4.4	3.9

- Hãy xác định các giá trị trung bình, trung vị, mode của **X** và **Y**.
- Vẽ biểu đồ Boxplot của **X** và **Y**.
- Sử dụng phương pháp chuẩn hóa dữ liệu z-score sử dụng độ lệch chuẩn tuyệt đối để chuẩn hóa dữ liệu quan sát của **X** và **Y** (nêu cách tính của cặp giá trị thứ *i* với *i* là số cuối cuối cùng trong mã sinh viên của bạn +1).
- Hãy làm trơn dữ liệu ban đầu của **X** và **Y** bằng phương pháp làm trơn biên (bin boundaries), trong đó việc phân chia thùng theo chiều sâu (Equal-depth) với số bin là 3. Mô tả các bước thực hiện.
- Xác định hệ số tương quan giữa **X** và **Y**.

Câu 2 (3 điểm).

Sử dụng thuật toán k-means và khoảng cách Euclide để phân cụm 10 quan sát dưới đây thành 2 cụm: $A_1(5,3)$, $A_2(15,12)$, $A_3(10,15)$, $A_4(24,10)$, $A_5(30,45)$, $A_6(85,70)$, $A_7(71,80)$, $A_8(60,78)$, $A_9(55,52)$, $A_{10}(80,91)$.

Khởi tạo tâm của 2 cụm là A_1 , A_3 .

Câu 3 (1 điểm).

Trình bày ngắn gọn nội dung chính của bài tập lớn của học phần Khai phá dữ liệu mà bạn đã thực hiện.

Ghi chú:

- Thí sinh được sử dụng tài liệu trong khi làm bài.
- Thí sinh không được trao đổi trong khi làm bài.
- Cán bộ coi thi không giải thích gì thêm.