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

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

X	6.4	4.5	5.1	4.9	4.6	5.7	5.4	5.1	5.1	5.3	5	7
Y	3.2	2.3	3.5	3	3.1	4.4	3.9	3.5	3.8	3.7	3.3	3.2

- Hãy xác định các giá trị trung bình, trung vị, mode của **X** và **Y**.
- Vẽ biểu đồ Boxplot của **X** và **Y**.
- Hãy chuẩn hóa giá trị của **X** và **Y** về dạng chuẩn z-score (nêu cách tính của cặp giá trị thứ  $i$  với  $i$  là số 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 trung bình (bin means), trong đó việc phân chia thùng theo chiều rộng (Equal-width) 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 Manhattan để phân cụm 7 quan sát dưới đây thành 3 cụm.  $A_1(1,2)$ ,  $A_2(2,5)$ ,  $A_3(8,4)$ ,  $A_4(5,8)$ ,  $A_5(7,5)$ ,  $A_6(2,10)$ ,  $A_7(6,4)$ .

Khởi tạo tâm của 3 cụm là  $A_1$ ,  $A_4$  và  $A_6$ .

**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.