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

- 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 tỷ lệ thập phân (decimal scale) 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 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). Cho tập dữ liệu giao dịch sau:

ID	Items
T1	F, M, B, E
T2	O, A, D, H, S
T3	A, H, S, O
T4	D, B, F, E
T5	D, A, H, S
T6	E, O, M
T7	O, F, E

- Giả sử ngưỡng của độ hỗ trợ (minimum support) là 40%. Sử dụng thuật toán Apriori tìm tất cả các tập đối tượng thường xuyên.
- Giả sử ngưỡng của độ hỗ trợ là 40%, ngưỡng của độ tin cậy (minimum confidence) là 75%. Từ tập dữ liệu trên hãy xác định các luật kết hợp mạnh.

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.