
 <p>TRƯỜNG ĐẠI HỌC 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	5.1	4.9	4.6	5.7	5.4	5.3	5	7	4.5	5.5	6.5	6.1
Y	3.5	3	3.1	4.4	3.9	3.7	3.3	3.2	2.3	3.5	2.8	2.8

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 min-max để chuẩn hóa dữ liệu quan sát của **X** và **Y** về đoạn $[1, 10]$ (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 trung bình (bin means), 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). Cho tập dữ liệu giao dịch sau:

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

- Giả sử ngưỡng của độ hỗ trợ (minimum support) là 30%. Sử dụng thuật toán Apriori tìm tất cả các tập đối tượng thường xuyên.
- Giả sử ngưỡng của độ hỗ trợ là 30%, ngưỡng của độ tin cậy (minimum confidence) là 70%. 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.