1. (2 credits) What is the confidence for the rules A 🡪 ∅ ?  Herein, ∅ indicates an empty set.

100%

Confidence = σ(A, ∅) / σ(A). Since, σ(A, ∅) = σ(A), σ(A)/ σ(A) = 1

Associate rule: X → Y

Confidence (c): Measures how often items in Y appear in transactions that contain X

(Milk, Diaper) 🡪 (Beer)



2. (2 credits) Let c1, c2, and c3 be the confidence values of the rules {p} 🡪{q}, {p} 🡪{q, r}, and {p, r} 🡪{q}, respectively. If we assume that c1, c2, and c3 have different values, which rule has the lowest confidence? Explain the reason.

C1 = , C2 = , C3 =

According to Anti-monotone Property, σ (p, q) ≥ σ (p, q, r) and σ (p) ≥ σ (p, r). The lowest one means the numerator is the smallest, and denominator is the largest. Therefore, C2 is the lowest, because σ (p, q, r) is the smallest and σ (p) is the largest.

3. (3 credits) Consider the following traffic accident data. To discover the associate patterns, we need to binarize each categorical attribute, e.g., introducing two variables representing ‘good’ or ‘bad’ weather condition, or using two variables representing ‘Yes’ or ‘No’ for Seat Belt. Please store the binarized data in MATLAB and write a script to generate the frequent itemsets. You can use any threshold for the minimal support.



You will need to submit both codes and the discovered frequent itemsets.

4. (4 credits) Based on the previous results, write an Apriori algorithm to generate the association rules with high confidences. Please set a proper confidence threshold to get 1-50 rules.

**Submission instructions: what to hand in**

* Prepare a single PDF to include all required figures, analysis and results.
* Submit your PDF file and source codes through the blackboard on or before the due time.
* No hardcopy is required.