CS5228: Knowledge Discovery and Data Mining

Midterm Exam — Common Mistakes & Feedback

Responsibilities

	Q17	Q18	Q19	Q20	Q21
Liu Chenyan chenyan@u.nus.edu			✔		
Gao Ruize e1023891@u.nus.edu		✔			
Rajdeep Singh Hundal rajdeep@nus.edu.sg	✔				
Luo Yang yang_luo@u.nus.edu					✔
Ayush Goyal e1124577@u.nus.edu				✔	

MCQ/MRQ

- Q1-16: Partial marking using the Canvas scheme
  - Divide the number of points for the question by the number of correct responses. That's how much each selected response is worth. For example, if there are 2 points in total and 3 correct answers, each answer is worth 0.67.
  - For every correct response that is selected, the student gets that many points; for every incorrect response that is selected, the student loses that many points. For example, if a student selected 2 out of 3 correct answers, the student gets 1.33 points. If the student also selected 1 wrong answer, the final points for this question will be 0.67,
  - The final number of points may not go negative even if they mark more incorrect responses than they mark correct responses.
  - Non-selected responses do not figure into the calculation. The student doesn't get anything for not answering.

Clustering

- Q17: AGNES with different Linkage Methods (7 Points)
  - a and b)
    - 3 Points: Correct answer
    - 2 Points or 1 Point: Incorrect answer (depending on severity of mistake)
      - Example: Forgot to specify a merge step but the previous and subsequent merge steps, as well as final clustering, is correct
      - Example: Order of merges were ambiguous and not explicitly specified
      - Example: Did not specify final 3 clusters and only specified merges
      - Example: Did not specify merges and only specified final 3 clusters
      - Example: Merged incorrectly
    - 0 Point: Blank or random answers
  - c)
    - 1 Point: Complete linkage with a **valid explanation**
      - Example: Complete linkage is better because it is less susceptible to noise and thus, “chaining”
    - 0 Point: Complete linkage but explanation is vague/wrong or single linkage
      - Example: Complete linkage because it is better
      - Example: Complete linkage because I use the “eyeball” method

Association Rule Mining

- Q18: Identify candidate itemsets (Apriori algorithm)

This question emphasizes understanding and mastering the Apriori algorithm, particularly in generating candidate 3-itemsets and performing subset pruning. The second part of the question holds more weight in principle.

Initial Candidate 3-Itemsets (2 Points)

- Correct identification of initial candidate 3-itemsets: {A, B, C}, {A, C, D}, {A, B, D}, {B, C, D} earns 2 points.
- If incorrect, 0-1 points will be awarded based on the attempt and logic used.

Surviving Subset Pruning (2 Points)

- Correct identification of the remaining itemsets after pruning: {A, B, D}, {B, C, D} earns 2 points.

Notes:

- If only the second part (subset pruning) is correct but no explanation of the Apriori algorithm's process is provided, 3 points will be awarded overall.
- Full marks (4 points) can still be awarded if the explanation demonstrates a solid understanding of the Apriori algorithm, even if the initial 3-itemsets are not explicitly listed.

Association Rule Mining

- Q19: Improving performance through sampling
  - The scoring points are as follows:
    i. Sampling size (specific size or percentage)
    ii. Adopt lower support & confidence
    iii. Mined rules should go through complete database
    iv. Get true support & confidence
    v. Prune rules that violate
  - Other methods, such as parallelism between groups or clustering, can also score few point.

Evaluation of Classifiers

- Q20: Calculation of FP and TN (6 points)
  - The correct answer is:
    - FP = 10
    - TN = 70
  - Both correct : 6 points
  - One correct : 3 points
  - Both wrong : 0 point
  - No partial marks given.
  - No penalty for showing no calculation.
  - In case answers obtained were right, but the student reversed FP and TN, 1 point is deducted

Tree-Based Models

- Q21 a: Annotate Decision Tree
  - To be specific, the correct answer is
    D1:Altitude; D2:Size; D3:Size; B1:low; B2:high; B3:large; B4:small; B5:small; B6:large.

    The two branches of D1 can be swapped. (6 points)
    Partially correct: some nodes are set incorrectly. (e.g., D3 is set as “Color”).
    Blank or random answers (0 point)
- Q21 b: Re-use of features
  - All attributes have only two different values each (i.e., binary features). (1 point) Corresponding explanation. (1 point)
  - The answer with no explanation will get 1 point.