Midterm

(Advanced) Data Mining - Winter 2022

Feb 16-17, 2022 - 120mins

FLINT

Instructions

- You can use notes, books, and course slides. You have two hours to complete to give you additional time for the logistics.
- You can submit each question separately with different formats, i.e., txt file, script.
- You need show your work step by step for some questions.
- You can use a calculator or any programming environment (R, Python, Wolfram) for your calculations, but you should show the details of your calculation
- Suppose you want to compute $(6/10)*log_2(6/10)+(4/10)*log_2(4/10)$, you can either submit this as a text file with its results or use R/Python and submit your script with results in the comments

  (If you want, you can write a little code: $sqrt((x1-y1)^2+(x2-y2)^2)$ where x1=4, x2=9 and y1=2, y2=5. Every time you want to calculate a distance, you can just change the x and y values.)
- Any sign of group work will not be tolerated.

1. Suppose you have these data points: 29, 75, 13, 20, 168, 163, 140, 52, 4, 37, 36, 123, 120, 31, 111. Then,

(a) If you draw a histogram with a bin size of 25, how many bars will you have in your chart? Please justify your answer. (7 points)

(b) What's the value at the 60^{th} percentile. You can use R and submit the statement with the answer. (7 points)

(c) Use z-score normalization to transform the value 36. You can use R and submit the statement with the answer. (7 points)

(d) Use min-max normalization to transform the value 13 onto the range [1, 10]. You can use R and submit the statement with the answer. (7 points)

(e) Suppose you have a bin depth 3 and use smoothing by bin median to smooth the first bin. You can use R and submit the statement with the answer. (7 points)

2. Compute the distance between objects 3 and 4 in the table below. Solution can be typed or scanned before submission and you can use calculator or R etc. (15 points)

Object	test-1 (nominal)	test-2 (ordinal)
1	A	excellent
2	B	fair
3	A	good
4	A	excellent

3. Use chi-square for the data below to find out whether there's a relation between playing basketball and eating cereal. Based on your result describe the relation. Solution can be typed or scanned before submission and you can use calculator or R etc. (15 points)

	Basketball	Not basketball	Sum (row)
Cereal	213	203	416
Not cereal	138	110	248
Sum(col.)	351	313	664

4. Based on the table below, please solve either 4a or 4b depending on your standing, i.e., undergrad, grad. Solution can be typed or scanned before submission and you can use calculator or R etc. (35 points)

gender	**age**	**income**	**play golf?**	**count**
male	young	medium	yes	30
male	young	low	no	20
female	young	low	no	30
female	teenager	medium	no	20
male	young	high	yes	15
female	young	medium	no	30
female	elder	high	yes	13
male	middle age	medium	yes	10
female	elder	medium	yes	4

(a) (**Undergraduates**) Using the data table above, calculate the information gain for gender and age. Please show your calculations.

(b) (**Graduates**) Using the data table above, calculate the gain ratio for age and income. Notice that you will be performing the gain calculation and dividing it by the split info. Please show your calculations.