Data Mining Fall 2018 22:544/198:650:30

Homework 4 Due: Dec 05 2018

Problem 1 (**40 points**)

Consider the data set shown in the Table below.

Age	Number of Hours Online per Week (B)				
(A)	$0-5$	$5-10$	$10-20$	$20-30$	$30-40$
$10-15$	2	3	5	3	2
$15-25$	2	5	10	10	3
$25-35$	10	15	5	3	2
$35-50$	4	6	5	3	2

(a) For each combination of rules given below, calculate the confidence and specify the rule that has the highest confidence.

    a. 15 <A< 25 −→ 10 <B< 20

    b. 10 <A< 25 −→ 10 <B< 20

    c. 15 <A< 35 −→ 10 <B< 30

(b) Suppose we are interested in finding the average number of hours spent online per week by Internet users between the age of 10 and 25. Write the corresponding statistics-based association rule to characterize the segment of users. To compute the average number of hours spent online, approximate each interval by its midpoint value (e.g., use B = 7.5 to represent the interval 5 <B< 10).

(c) Test whether the quantitative association rule given in part (b) is statistically significant by comparing its mean against the average number of hours spent online by other users who do not belong to the age group.

Problem 2 (**30 points**)
(a) List all the 4-subsequences contained in the following data sequence:
$< \{1, 3\} \{2\} \{2, 3\} \{4\} >$.

(b) List all the 3-element subsequences contained in the data sequence for part (a) assuming that no timing constraints are imposed.

Problem 3 (30 points). The Scikit-learn provides 3 robust regression estimators: RANSAC, Theil Sen, and HuberRegressor. Please list the advantages and disadvantages of these estimators.