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 (A)	Number of Hours Online per Week (B)				
	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. $15 < A < 25 \longrightarrow 10 < B < 20$ confidence=10/30=0.33

b. $10 < A < 25 \longrightarrow 10 < B < 20$ confidence=15/(30+14)=0.33

c. $15 < A < 35 \longrightarrow 10 < B < 30$ confidence=(10+10+5+3)/(30+35)=28/65

(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 $5 < B < 10$).

Please note: this is different from the online solution!!!!!!

μ=(4*2.5+8*7.5+15*15+13*25+5*35)/45=17.67

Association rule: $10<A<25 \rightarrow B: \mu = 17.67$

(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.

Null Hypothesis: $\mu = \mu'$

$\mu = 17.67$ s=9.492

$\mu' = 11.5$ s'=9.433

$Z=\frac{u-u'}{\sqrt{\frac{s^2}{n1}+\frac{s'^2}{n2}}} = 3.243$

If you choose 95% confidence interval, Z>1.96 reject null hypothesis, the association rule in part b is an interesting rule

Problem 2 (30 points)

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

< {1, 3} {2} {2} > < {1, 3} {2} {3} > < {1, 3} {2} {4} > < {1, 3}{2, 3} > < {1, 3} {3} {4} > < {1} {2} {2, 3} > < {1} {2} {2} {4} > < {1} {2} {3} {4} > < {1} {2, 3} {4} > < {3} {2} {2, 3} > < {3} {2} {2} {4} > < {3} {2} {3} {4} > < {3} {2, 3} {4} > < {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.

{1, 3} {2} {2, 3} > < {1, 3} {2} {4} > < {1, 3} {3} {4} > < {1, 3} {2} {2} > < {1, 3} {2} {3} > < {1, 3} {2, 3} {4} > < {1} {2} {2, 3} > < {1} {2} {4} > < {1} {3} {4} > < {1} {2} {2} > < {1} {2} {3} > < {1} {2, 3} {4} > < {3} {2} {2, 3} > < {3} {2} {4} > < {3} {3} {4} > < {3} {2} {2} > < {3} {2} {3} > < {3} {2, 3} {4} >

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.

- HuberRegressor should be faster than RANSAC and Theil Sen unless the number of samples are very large, i.e `n_samples` >> `n_features`. This is because RANSAC and Theil Sen fit on smaller subsets of the data. However, both Theil Sen and RANSAC are unlikely to be as robust as HuberRegressor for the default parameters.
- RANSAC is faster than Theil Sen and scales much better with the number of samples
- RANSAC will deal better with large outliers in the y direction (most common situation)
- Theil Sen will cope better with medium-size outliers in the X direction, but this property will disappear in large dimensional settings.