Feedback — Week 2 Quiz

Help Center

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Wed 18 Feb 2015 8:25 AM PST**. You got a score of **10.50** out of **11.00**. You can attempt again, if you'd like.

Question 1

Suppose a school collected some data on students' preference for hot dogs(HD) vs. hamburgers(HM). We have the following 2×2 contingency table summarizing the statistics. If lift is used to measure the correlation between HD and HM, what is the value for lift(HD, HM)?

	HD	¬HD	Σ_{row}
HM	40	24	64
¬HM	210	126	336
Σ_{col}	250	150	400

Your Answer		Score	Explanation
-1			
-∞			
0			
1	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "1".

The lift can be calculated by

$$\text{Lift} = \frac{\text{supp}(A \cup B)}{\text{supp}(A) \cdot \text{supp}(B)},$$

where supp(A) and supp(B) refer to the relative support of A and B respectively. Thus,

$$\text{Lift} = \frac{40/400}{250/400 \times 64/400} = 1.$$

Question 2

Suppose Coursera collected statistics on the number of students who take courses on data mining (DM) and machine learning (ML). We have the following 2×2 contingency table summarizing the statistics. If χ^2 is used to measure the correlation between DM and ML, what is the χ^2 score?

	DM	¬DM	Σ_{row}
ML	700	300	1000
¬ML	500	1500	2000
Σ_{col}	1200	1800	3000

Your Answer		Score	Explanation
-562.5			
-225			
225			
562.5	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "562.5".

The contingency table with expected values is following table

	DM	¬DM	Σ_{row}
ML	700 (400)	300 (600)	1000
¬ML	500 (800)	1500 (1200)	2000

Σ_{col}	1200	1800	3000

χ^2 can be evaluated as follows

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency, and E_i is the expected frequency, thus

$$\chi^2 = \frac{(700-400)^2}{400} + \frac{(300-600)^2}{600} + \frac{(500-800)^2}{800} + \frac{(1500-1200)^2}{1200} = 562.5.$$

Question 3

What is the value range of the Lift measure?

Your Answer		Score	Explanation
[-1, 1]			
[0, +∞)	✔	1.00	
[0, 1]			
(-∞, +∞)			
Total		1.00 / 1.00	

Question Explanation

By definition, the correct answer is: "[0, +∞)".

Question 4

Which of the following measures is NOT null invariant?

Your Answer		Score	Explanation
Cosine			
χ^2	✔	1.00	
Kulcyzynski			
All confidence			

Total	1.00 / 1.00

Question Explanation

The correct answer is: "χ^2".

Null transactions are considered in χ^2.

Question 5

Suppose we are interested in analyzing the transaction history of several supermarkets with respect to purchase of apples(A) and bananas(B). We have the following table summarizing the transactions.

Supermarket	AB	¬AB	A ¬B	¬A ¬B
S1	100,000	7,000	3,000	300
S2	100,000	7,000	3,000	90,000

Denote l_i as the lift measure and ki as the Kulcyzynski measure for supermarket S_i (i = 1, 2). Which of the following is correct?

Your Answer		Score	Explanation
◉ $l_1 \neq l_2$, $k_1 = k_2$	✔	1.00	
○ $l_1 = l_2$, $k_1 \neq k_2$			
○ $l_1 \neq l_2$, $k_1 \neq k_2$			
○ $l_1 = l_2$, $k_1 = k_2$			
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "$l_1 \neq l_2$, $k_1 = k_2$".

Lift is not null invariant and therefore sensitive to the number of null transactions, while Kulcyzynski is null invariant.

Question 6

Consider the support-based and null-invariant definitions for negative patterns. For negative pattern threshold ε = 0.011, which of the following patterns would be considered a negative pattern by the null-invariant definition but not the support-based definition?

Your Answer		Score	Explanation
◉ A media content provider has 1,000,000 users. Movie A and Movie B were viewed by 1000 users each in the last month, but only 10 users viewed both.	✔	1.00	
○ There are 5 million registered students on an online education website. 5000 students registered for Music 101, and 50,000 students registered for Data Mining, but only 500 students registered for both.			
○ Both of the above are correct.			
○ None of the above are correct.			
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "A media content provider has 1,000,000 users. Movie A and Movie B were viewed by 1000 users each in the last month, but only 10 users viewed both."

Null-invariant:

$$\frac{P(A|B)+P(B|A)}{2} = \frac{10/1000+10/1000}{2} = 0.01 \leq \epsilon$$

Support based:

$$\sup(\text{Movie A}) = \sup(\text{Movie B}) = \frac{1000}{1000000} = 10^{-3}$$
$$\sup(\text{Movie A} \cup \text{Movie B}) = \frac{10}{1000000} = 10^{-5}$$
$$\sup(\text{Movie A}) * \sup(\text{Movie B}) = 10^{-3} * 10^{-3} = 10^{-6}$$
$$< \sup(\text{Movie A} \cup \text{Movie B})$$

Thus, {Movie A, Movie B} is a negative pattern by the null-invariant definition but not the support based definition.

Null-invariant:

$$\frac{P(DM|Music)+P(Music|DM)}{2} = \frac{500/5000+500/50,000}{2} = 0.055 > \epsilon$$

Support based:

$\sup(\text{DM}) = \frac{50000}{5000000} = 0.01$
$\sup(\text{Music}) = \frac{5000}{5000000} = 0.001$
$\sup(\text{DM} \cup \text{Music}) = \frac{500}{5000000} = 10^{-4}$
$\sup(\text{DM}) * \sup(\text{Music}) = 0.01 * 0.001 = 10^{-5} < \sup(\text{DM} \cup \text{Music})$

Thus, {DM, Music} is not a negative pattern by either definition.

Question 7

Pat-ID	Item-Sets	Support
P1	{A, C, E, S}	205227
P2	{F, A, C, E, S}	205211
P3	{F, A, C, E, T, S}	101758
P4	{F, A, C, T, S}	161563
P5	{A, C, T, S }	161576

Table 1: Support for frequent itemsets

Consider two patterns P_1 and P_2 such that $O(P_1) \subseteq O(P_2)$, where $O(P_i)$ is the corresponding itemset of pattern P_i. Take a second to convince yourself that the following is true:

$$Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|} = 1 - \frac{|T(P_2)|}{|T(P_1)|}$$

Which of the following patterns in Table 1 is δ-covered by {F, A, C, E, T, S} for δ=0.4? Select all that apply.

Your Answer		Score	Explanation
☑ {F, A, C, T, S}	✔	0.25	
☑ {A, C, T, S}	✔	0.25	
☑ {F, A, C, E, S}	✖	0.00	
☑ {A, C, E, S}	✖	0.00	
Total		0.50 / 1.00	

Question Explanation

The correct answers are: "{F, A, C, T, S}" and "{A, C, T, S}".

Dist({F, A, C, E, T, S}, {A, C, E, S}) = 1- 101758/205227 = 0.504
Dist({F, A, C, E, T, S}, {F, A, C, E, S}) = 1- 101758/205211 = 0.504
Dist({F, A, C, E, T, S}, {F, A, C, T, S}) = 1- 101758/161563 = 0.370
Dist({F, A, C, E, T, S}, {A, C, T, S}) = 1- 101758/161576 = 0.370

Thus, {F, A, C, T, S}, {A, C, T, S} are 0.4-covered by {F, A, C, E, T, S}

Question 8

Transactions	# of Transactions
(abe)	100
(bcf)	100
(acf)	100
(abcef)	100

Table 2: # transactions in a database.

Given the transactions in Table 2, which of the following is a 0.5-core pattern of (abcef)? Select all that apply

Your Answer		Score	Explanation
(a)	✔	0.20	
(e)	✔	0.20	
(abe)	✔	0.20	
(acfe)	✔	0.20	
(abcef)	✔	0.20	
Total		1.00 / 1.00	

Question Explanation

The correct answers are: "(acfe)", "(abcef)", "(e)", and "(abe)".

Every pattern except (a) is contained in at most one other pattern in the database, making it a 0.5-core pattern of (abcef).

Question 9

A constraint is anti-monotone if an itemset S violates the constraint, so do all of its supersets.

Which of following constraints is anti-monotone?

Your Answer		Score	Explanation
◉ Relative support of S > 0.3	✔	1.00	
○ range(S.profit) > 10			
○ avg(S.price) < 10			
○ sum(S.price) > 25			
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "Relative support of S > 0.3"

Suppose X is the superset of S; if so, the following inequalities always hold for all X:

- Relative support of X ≤ Relative support of S
- sum(X.price) ≥ sum(S.price)
- avg(X.price) <> avg(S.price)
- range(X.profit) ≥ range(S.profit)

From these inequalities, we can easily verify that relative support of S > 0.3 is anti-monotone.

Question 10

A constraint is monotone if an itemset S satisfies the constraint, so do all of its supersets. Which of following constraints is monotone?

Your Answer		Score	Explanation
◉ sum(S.price) > 20	✔	1.00	
○ avg(S.price) < 10			
○ min(S.price) > 10			
○ Support of S > 100			
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "sum(S.price) > 20"

Suppose X is the superset of S; if so, the following inequalities always hold for all X:

- support of X ≤ support of S
- min(X.price) ≤ min(S.price)
- avg(X.price) <> avg(S.price)
- sum(X.price) ≥ sum(S.price)

From these inequalities, we can easily verify that sum(S.price) > 20 is monotone.

Question 11

A constraint is succinct if the constraint c can be enforced by directly manipulating the data.

Which of following constraints is NOT succinct?

Your Answer		Score	Explanation
min(S.profit) < 40			
range(S.price) > 2	✔	1.00	It cannot be determined beforehand since range of the price of itemset S keeps increasing.
v ∈ S			
min(S.profit) > 40			
Total		1.00 / 1.00	

Question Explanation

The correct answer is: "range(S.price) > 2"