Distributed Information Systems: Spring Semester 2015

Quiz 6: Association Rules + Clustering & Classification

`Student Name: _________________________________________ Date: 21 May 2015`
`Student ID: ___________________________________________ Time: 11:15AM to 11:30AM`

Total number of questions: 8

Each question has a single answer!

Consider the following *D(Transaction ID, Item List)* database.

TID	*ItemIDs*
T100	I1,I2,I3,I5
T200	I2,I4
T300	I2,I3
T400	I1, I2, I4
T500	I1, I3, I5
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3
T900	I1, I2, I3, I5

I1I2 = 4
I1I3 = 4
I1I4 = 1
I1I5 = 3
I2I3 = 4
I2I4 = 2
I2I5 = 2
I3I4 = 0
I3I5 = 3
I4I5 =

Let L_k be the set of frequent $k-$itemsets and the *minimum support count* for the apriori algorithm be 3.

1. Which of the following itemsets has a support count 3?

☐ *a)* {I1,I2} ☐ *b)* {I1,I3} ☒ *c)* {I1,I5} ☐ *d)* {I2,I3}

2. What is the size of set L_2?

☐ *a)* 3 ☐ *b)* 4 ☒ *c)* 5 ☐ *d)* 6

---

3. Which statement about Association Rules Mining is **correct**?

☐ *a)* If we take the union of two frequent (k-1)-itemsets which differ only by one item, we will always get a frequent k-itemset as a result.

☐ *b)* The JOIN step of the apriori algorithm requires heavy database access.

☒ *c)* A subset of a frequent itemset is always a frequent itemset.

☐ *d)* Using the apriori property results in a larger search space but faster generation of frequent item sets.

4. Which statement about Association Rules Mining is **wrong**?

☒ *a)* The computation of frequent itemsets needs to have a minimum confidence level defined. only need support

☐ *b)* Dynamic discretization allows transforming quantitative values into categorical ones, based on the distribution of the data.

☐ *c)* Confidence metrics determines for a frequent itemset whether a rule is implied.

☐ *d)* A very low support value for a certain rule indicates that the body and the head rarely occur together in the same transaction.

5. Which one is in general considered as a main advantage of the k-means clustering algorithm?

   □ *a)* It often terminates at a local optimum.
   □ *b)* It detects exclusively convex clusters.
   □ *c)* It is necessary to specify k in advance.
   ☒ *d)* It is efficient.

6. Consider the 4 clusters C_1 to C_4 and the following initial assignment of points:

   $C1 = (3, 10), (4, 11), (5, 12)$
   $C2 = (10, 10), (11, 10), (12, 10)$
   $C3 = (3, 3), (4.5, 8), (5, 9), (5.5, 9)$
   $C4 = (7, 4), (8, 7), (8, 8), (9, 6)$

   To which cluster would the k-means algorithm assign the point $(7, 9)$ initially (i.e., after the first iteration)?

   □ *a)* C_1 □ *b)* C_2 □ *c)* C_3 ☒ *d)* C_4

7. While building a decision tree using C4.5, we **cannot** split a leaf further when…

   ☒ *a)* …all samples belong to the same class.
   □ *b)* …all attributes have already been used once in the whole tree.
   □ *c)* …every sample belongs to a different class.
   □ *d)* …all remaining attributes have a different entropy.

8. Which property is common to clustering and classification?

   ☒ *a)* They can be applied to multi-dimensional data with both numerical and categorical attributes.
   □ *b)* They need a training set with the classes assigned.
   □ *c)* They are unsupervised.
   □ *d)* The number of classes is always known beforehand.