Distributed Information Systems: Spring Semester 2016

Quiz 3

Student Name: _____ Date: 7 Apr 2016

Student ID: _____ Time: 11:15AM to 11:30AM

Total number of questions: 8

Each question has a single answer!

1. Which of following is **wrong** about data guide?

   ☐ *a*) The data guide summarizes the data in a concise way (i.e., every path occurs only once)

   ☐ *b*) The nodes in a data guide define classes of nodes in the data graph

   ☐ *c*) The data guide is a deterministic schema graph

   ☒ *d*) The dataguide can never have cycles

2. Given the transactions in the following table, which of the following statements is **true**?

Transaction ID	Purchased Items
1	A,B,C
2	A,C
3	A,D
4	B,E

   ☐ *a*) $A \implies C$ with unknown support and $\approx 66.67\%$ confidence

   ☐ *b*) $C \implies A$ with 100% support and 50% confidence

   ☒ *c*) $A \implies C$ with 50% support and $\approx 66.67\%$ confidence

   ☐ *d*) $C \implies A$ with unknown support and 50% confidence

3. Given a frequent itemset T of size $k \geq 2$, computed from a database of shopping transaction with a given minimum support, which of the following is **true**:

   ☒ *a*) There exist at least k frequent itemsets of size $k-1$.

   ☐ *b*) Using the apriori algorithm, the database has been scanned $k+1$ times to find T.

   ☐ *c*) We can build at least $k-1$ association rules with confidence 100%.

   ☐ *d*) If another frequent itemset T' differs from T by exactly one element, then $T \cup T'$ is a $k+1$ frequent itemset.

4. For schema integration we constructed a Naive Bayes classifier that determines with which probability a data instance i with features T_i belongs to a class A.

   Which of the following probabilities is **not** used to train the classifier

   ☐ *a*) $P(A)$, the probability that an instance belongs to class A

   ☐ *b*) $P(t|A)$, the probability that a feature $t \in T_i$ occurs for an instance of class A

   ☒ *c*) $P(A|T_i)$, the probability that an instance belongs to class A given its features

   ☐ *d*) all the three probabilities are used

5. When integrating heterogeneous databases (e.g. in healthcare environments), the constituents of different schemas need to be related to each other according to semantic similarity. This activity is called:

   ☐ *a*) Schema analysis

   ☐ *b*) Schema extraction

   ☒ *c*) Schema matching

   ☐ *d*) Schema subsumption

6. Which of the following is **false** in the context of the Apriori algorithm for association rule mining:

- ☐ *a*) The PRUNE step removes all k-itemsets that contain a non frequent $(k-1)$-itemset.
- ☒ *b*) After the JOIN and PRUNE step, all remaining k-itemsets are frequent k-itemsets.
- ☐ *c*) The Apriori algorithm reduces the number of database accesses compared to a brute-force approach.
- ☐ *d*) Identifying frequent itemsets in partitions of the database can improve the algorithm's performance in large datasets.

7. Given sets $A = \{a, b, c, d, f\}$ and $B = \{a, b, c, d, e\}$, the Jaccard similarity between A and B is:

- ☐ *a*) 5
- ☐ *b*) 4 / 25
- ☒ *c*) 2 / 3
- ☐ *d*) 4 / 5

8. Given an association rule $I \implies J$. Confidence is the probability

- ☐ *a*) $P(I, J)$
- ☒ *b*) $P(J|I)$
- ☐ *c*) $P(I|J)$
- ☐ *d*) $P(I|J) - P(J)$