Databases and Data Mining 2015
Final Exam

LIACS Room 174
Friday December 18th 2015
10.00 – 13.00

- State your name and student number and affiliation (e.g. LU (Leiden University), TUD (TU Delft), etc.) on every page of your answers.
- Every assignment has the same weight. There are 10 assignments.
- **Always fully explain your answers!** (Answering a question with only 'yes' or 'no' will never be counted as a correct answer.)
- Please note that you have a total of 3 hours to answer the questions.
- It is an open book exam: you are allowed to use your book and course notes (slides).
- All electronic equipment should be off the table and switched off.

1. Assume the following data is given: { 22, 12, 61, 57, 30, 1, 32, 37, 37, 68, 42, 11, 25, 7, 8, 16 }.
   a) Apply data discretization by binning the data into 4 bins using equal-depth and equi-width binning, respectively.
   b) Describe the differences between the two binning methods. Give for each of the binning methods an example application for which that binning method is the most appropriate.
   c) If you know that the data actually represent ages of persons, what kind of binning method would you then use? (You may propose a third binning method.)

2. Suppose that a data warehouse *DW* for a Sales Company consists of five dimensions, *time, location*, *supplier*, *brand*, and *product*, and two measures, *count*, and *price*.
   a) Draw a *snowflake schema* diagram for the data warehouse.
   b) Starting with the base cuboid [*time*, *location*, *supplier*, *brand*, *product*], what specific OLAP operations should one perform in order to list the total *count* for a certain *brand* for each *state* per *year* (assume *location* has three levels: *country, state, city;* and assume *time* has three levels: *year*, *month, day*)?
   c) Assume we already summarized the measures over all *time* and all *suppliers*, we want to compute the remaining 3-dimensional data cube. If there are 100 *locations*, 200 *brands*, and 10000 *products,* in what order would you traverse the cube cells when you use multi-way array aggregation for the data cube computation?
   d) If each dimension has three levels (including *all*), how many cuboids will the data cube contain (including the base and apex cuboids)? (Note: the correct expression for the actual number of cuboids is sufficient, you do not have to evaluate the expression into a single number.)

3. Consider the data warehouse *DW* as described in *Question* 2.
   a. Assume we have the following constraint on the product items: *min(price)* is less or equal than $100. What kind of constraint is this, *antimonotonic*, *monotonic*, or *succinct*? Is this constraint strongly convertible? (As always: explain your answer.)
   b. Give an example of a succinct constraint for the data warehouse *DW.*
   c. What kind of constraints can be 'pushed deep' in constraint-based data mining when using the FP-Growth Algorithm?

4. An optimization in frequent item set mining is mining closed patterns, or mining max patterns instead.
   a. Describe two main differences between mining closed patterns and mining max patterns.
   b. Would the FPGrowth Algorithm be useful for mining closed patterns in a 10^3-dimensional data set of 10^9 elements?

5. Facebook is a social network where links are established between friends (with approval, hence we call this a connection) and followers (without approval, here we do not call this a connection).
   a. Businesses have often tightly connected followers, whereas politicians have mildly connected followers. Explain a possible reason for this.
   b. Assume a Facebook user is not following any politician yet, but will do so very soon. Describe how the collaboration based function prediction method can be modified and used to predict the politician he/she will follow. Give an example of a reinforcement based function predictor for this application.

6. A data cube may be (i) partially materialized or (ii) materialized as an iceberg cube. Describe the advantages and disadvantages of both methods for data cube materialization. Also give for each of the methods a typical application example.

7. A database has five transactions. Let min_sup = 60%, and min_conf = 80%.

TID	*Items_bought*
T100	{A,B,K,S,T,W}
T200	{B,K,N,S,W}
T300	{A,B,N,P,R,S}
T400	{A,B,K,S,T,W}
T500	{ A,K,N,T}

   a. Find all frequent item-sets using the Apriori Algorithm (give the intermediate results for the different steps of the algorithm).
   b. How does the space complexity of this method grow in terms of the dimension and size of the data set? What would be the space complexity for the FP-Growth Algorithm?
   c. Given a set of N transactions of dimension D, what is the maximal number of database scans of the Apriori algorithm? Give two frequent item-set mining methods that will perform better in terms of the number of database scans.

8. Some data sets used for data mining have many thousands of dimensions.
   a. Would the Apriori algorithm be a useful algorithm when mining these kind of data sets?
   b. What kind of data mining method would you propose for mining these kind of data sets?
   c. What kind of method could be used to reduce the dimension of this kind of data set?

9. Assume a database *DB* of transactions is given with items/events {A, B, C, …}. After data mining the database the pattern *A* => *B* is found. (Note: *A* can be something like *"has a car",* and *B "walks less than 5 km a day"*.)
   a. Give an example of a null-invariant interestingness measure for the correlation/dependence of the two items/events *A* and *B*.
   b. What value of your measure would you expect, if the item sets in *DB* are neutral with respect to items/events *A* and *B*?
   c. What value of your measure would you expect, if the item sets in *DB* are imbalanced and neutral with respect to items/events A and B?

10. Around us we can observe several examples of so called *natural* networks.
   a. Give three examples of a *natural* network.
   b. There are several important graph theoretic measures that can be used to characterize these kind of networks. Give three of them and describe their typical behavior/values when measured on one of your examples given in a).
   c. Which network generation model could you use to generate synthetic versions of a natural network? Will the three characteristics, you mentioned earlier, be similar for a synthetic network generated by your generation model?