Data Warehousing and Data Mining
Written Exam

20.06.2011

First name		Last name	
Student number		Signature	

Instructions for Students

- Write your name, student number, and signature on the exam sheet.
- This is an open book exam: you can use your lecture notes.
- You have 2 hours for the exam.

Good luck!

Reserved for the Teacher

Questions	Max. points	Points
DM part	30	
DW part	30	
Total	60	

Data Mining Part

In this part, the guidelines for answering questions are the following:

- an answer to a given question can be composed of more than one choice
- you get **+1** point for each correct answer
- you get **-1** point for each wrong answer
- you get **0** point if you abstain

Advise: *if you are not sure about an answer, it is better to abstain.*

1. In a data mining task where it is not clear what type of patterns could be interesting, the data mining system should:
   - perform all possible data mining tasks
   - handle different granularities of data and patterns
   - perform both descriptive and predictive tasks
   - allow interaction with the user to guide the mining process
2. To detect fraudulent usage of credit cards, the following data mining task should be used:
   - feature selection
   - prediction
   - outlier analysis
3. In high dimensional spaces, the distance between data points becomes meaningless because:
   - it becomes difficult to distinguish between the nearest and farthest neighbors
   - the nearest neighbor becomes unreachable
   - the data becomes sparse
   - there are many uncorrelated features
4. The difference between supervised learning and unsupervised learning is given by:
   - unlike unsupervised learning, supervised learning needs labeled data
   - unlike supervised leaning, unsupervised learning can form new classes
   - unlike unsupervised learning, supervised learning can be used to detect outliers
5. To find inherent regularities in data, we should perform:
   - clustering
   - frequent pattern analysis
   - regression analysis
   - outlier analysis

6. Correlation analysis is used to:
    - extract association rules
    - define support and confidence values
    - eliminate misleading rules
7. Which of the following sentences is correct?
    - moments describe the nature of a distribution
    - entropy and uncertainty are positively correlated
    - hypothesis testing is used to identify highly skewed distributions
    - Bayesian classifiers assume dependency between variables
8. Which of the following is a data mining task?
    - detect the kinds of DNA sensitive to a given drug
    - compute for each Web document the number of its outgoing links
    - identify costumers with the highest purchase rate
    - sale campaign analysis
9. The goal of clustering analysis is to:
    - maximize the inter-class similarity
    - maximize the intra-class similarity
    - maximize the number of clusters
10. To facilitate implementations and provide high system performance, it is desirable to use:
    - no coupling between data mining and database systems
    - loose coupling between data mining and database systems
    - semi-tight coupling between data mining and database systems
    - tight-coupling between data mining and database systems
11. In non-parametric models:
    - there are no parameters
    - the distribution is not specified a priori
    - the parameters are fixed in advance
    - the parameters are flexible
    - a type of probability distribution is assumed, then its parameters are inferred
12. In decision tree algorithms, attribute selection measures are used to
    - reduce the dimensionality
    - select the splitting criteria that best separate the data
    - reduce the error rate
    - rank attributes

13. Pruning a decision tree always:
    - increase the error rate
    - reduce the size of the tree
    - provide partitions with lower entropy
14. In decision tree algorithms applied to datasets with a large number of classes and numerical attributes, the attribute selection method to avoid is:
    - Information gain
    - Gain ratio
    - Gini index
15. Which of the following sentences are correct:
    - partitions that have similar size have a low SplitInfo value
    - a skewed distribution of tuples over partitions has a high SplitInfo value
    - the Gini Index considers binary splits
    - Information gain is biased towards tests with many outcomes
16. Which of the following classifiers fall in the category of lazy learners:
    - Decision trees
    - Bayesian classifies
    - K-NN classifiers
    - Rule-based classifiers
17. In KNN classification, choosing high values of K results in
    - predicting the most frequent class label
    - increasing the risk of overfitting due to noise in the training data
    - avoiding over-smoothing
18. In rule-based classification, the learned rules are always:
    - mutually exclusive
    - mutually exhaustive
    - mutually exclusive and exhaustive only if they are extracted from unpruned trees
19. Prediction differs from classification in:
    - not requiring a training phase
    - the type of the outcome value
    - using unlabeled data instead of labeled data

20. In binary classification, the following measure should be used when the accuracy is not acceptable:
    - precision
    - relative absolute error
    - sensitivity
    - specificity
21. How does K-means differ from DBSCAN
    - the number of iterations on the data
    - the shape of the clusters
    - the sensitivity to noise
    - the training phase
22. Which of the following clustering algorithm handle categorical data
    - k-means
    - DB-scan
    - k-medoids
    - **CLARA**
23. At which level changes should be made to make the Birch algorithm suitable for categorical data
    - the number of entries in a leaf
    - the Clustering Feature (CF)
    - the distance measures
    - the branching factor
24. In clustering, when the property of being an outlier is not a binary property, this means
    - binary classification cannot be used to detect outliers
    - a data object can be an outlier for a set of data points and a non outlier for others
    - a data object has a degree to which it is an outlier
    - a data object can be a local or a global outlier
25. Which application requires hierarchical clustering
    - clustering items purchased by customers of the same region
    - clustering tourists having the same profiles
    - clustering images that are visually similar
    - clustering Web documents based on their topics

26. In a transactional database, the following association rule *computer* $\rightarrow$ *webcam*$(60\%, 100\%)$ means:

    - 100% of costumers bought both a computer and a webcam
    - 60% of costumers bought both a computer and a webcam
    - 100% of costumers who bought a computer bought also a webcam
    - 60% of costumers who bought a computer bought also a webcam

27. In frequent pattern analysis, the number of frequent itemsets to be generated is:

    - in the worse case, equal to $M \times N$ where M is the number distinct items and N is the max length of transactions
    - sensitive to minconf threshold
    - exponential when the minsup threshold is high
    - exponential when the minsup threshold is low

28. The bottleneck of the Apriori algorithm is caused by:

    - the number of association rules
    - the number of scans required
    - the computation of support counting for candidates
    - the number of generated candidates

29. In a transactional database, the lift measure of the items **bred** and **rice** is equal to 0.5. This means that

    - if consumers buy bred they are more likely to buy rice
    - if costumers buy bred they are less likely to buy rice
    - if costumers buy bred they can buy rice or not with the same probability

30. Sampling is a widely used optimization technique for improving the efficiency of data mining algorithms. What are the criteria to take into account when generating a representative sample:

    - the type of features used to describe the dataset
    - the main memory usage
    - the distribution of the original dataset
    - the similarity between data objects
    - the noise

Data Warhousing Part

In this part, the guidelines for answering questions are the following:

- each question has exactly **one** correct answer
- +1 for each correct answer
- -1 for each wrong answer
- 0 if you abstain

Advise: *if you are not sure about an answer, it is better to abstain.*

1. What is Business Intelligence?
   - A combination of processes, technologies, and applications used to support decision making
   - A system that makes intelligent decisions for the user
   - An method to store huge amounts of data in a central repository
2. Which of the following statements is true for warehouse-driven data integration:
   - Data are not duplicated
   - There is a delay in query processing
   - High query performance is achieved
3. What is the first step in the design of a DW?
   - Choose the business processes
   - Choose the dimensions
   - Choose the measures for the fact table
4. The multidimensional model
   - Is more flexible and general than the ER model
   - Serves one purpose and describes what is important and what describes the important things
   - Contains facts that describe important things and dimensions that are the important things
5. At which granularity level should facts be stored in the multidimensional model?
   - lowest (finest) granularity
   - average granularity
   - highest (coarsest) granularity

6. Which of the following statements is not correct?
   - Surrogate keys produce smaller fact tables
   - Surrogate keys make the DW independent from operational changes
   - Surrogate keys contain "intelligence" which is helpful for data analysis
7. What are conformed dimensions?
   - Dimensions with the same attributes
   - If a dimension is identical or a strict mathematical subset of the most granular dimension
   - Dimensions that keep information about changes over time
8. A measure *quantity* that stores the number of items of a specific product in an inventory DW is
   - additive
   - semi-additive
   - non-additive
9. Which measures are easiest to handle in a DW?
   - additive
   - semi-additive
   - non-additive
10. Compared to the star schema, the snowflake schema
    - has de-normalized dimension tables
    - has a better performance
    - is harder to use due to many joins
11. The use of shared dimensions is important for
    - the design of data marts that can be integrated
    - increasing the query performance
    - to break down the development process into small chunks
12. How many result groups are produced by the following GROUP BY clause, if a has 2, b has 3, c has 1 and d has 4 different values?

    ```
    GROUP BY a, ROLLUP(b, c, d)
    ```

    - 24
    - 38
    - 39

13. We learned three SQL GROUP BY extensions: CUBE, ROLLUP, and GROUPING SETS. Which one can be used to express the two others?

    - CUBE
    - ROLLUP
    - GROUPING SETS

14. The use of composite columns in the SQL GROUP BY extensions allows to

    - skip aggregation across certain levels
    - a more efficient query evaluation
    - a concise formulation of queries

15. What is a correct processing order of an SQL statement?

    - FROM, WHERE, GROUP BY, HAVING, NTILE(4) OVER ()
    - FROM, WHERE, HAVING, GROUP BY, NTILE(4) OVER ()
    - NTILE(4) OVER (), FROM, WHERE, HAVING, GROUP BY

16. How many result groups are produced by the following SQL statement, if a, b and c all have 4 different values?

```
SELECT a, b, SUM(c)
RANK() OVER (PARTITION BY a ORDER BY SUM(c) DESC)
GROUP BY a, b
```

    - 64
    - 16
    - 12

17. What is one of the core features of the Generalized MD-Join (compared to SQL)?

    - Decouples grouping from agregation
    - Performs automatically a join of the base table and detail table
    - Allows to compute several aggregate functions in parallel

18. Which kind of aggregates cannot be computed by SQL window functions?

    - Distributive aggregates
    - 1D cumulative aggregates
    - 2D cumulative aggregates

19. How are algebraic aggregate functions evaluated with the Generalized MD-Join?

    - Are natively supported
    - Reduction to distributive aggregates in combination with a pre- and post-processing step
    - Reduction to holistic aggregates

20. Pre-aggregation in DW is done to

    - reduce the space requirements
    - improve query performance
    - both to reduce space requirements and to improve query performance

21. How many pre-aggregates can be computed in an n-dimensional data cube?

    - 2^n
    - n^2
    - n

22. In the greedy algorithm for pre-aggregate selection, the benefit of a view v depends

    - only on the views w that depend on v, i.e., $w \leq v$
    - on the set of all views
    - on the set of already selected views and the views that depend on v

23. When is the greedy algorithm for pre-aggregate selection optimal?

    - If all the benefits are equal
    - Never
    - If the benefit of the first view is much larger than the other benefits

24. Incremental maintenance of aggregation views require to store additional bookkeeping information. For instance, for the MIN aggregate functions, tuples of the form $(g, \mathit{minimum}, \mathit{count})$ are stored (g is the group, *minimum* is the MIN value, and *count* is a counter). Assume an entry $(g, 1000, 1)$ in a view. How is the new MIN value determined when the tuple $(g, 1000)$ is deleted?

    - Entire base table must be searched
    - Take the previous element in the view in sort order
    - Search the dimension tables

25. What is an efficient index structure for attributes with low cardinality?

    - Bitmap index
    - Hash index
    - B-tree index

26. The compressed bitmap index of 00000100010 is

    - 11001101
    - 1110101111011
    - 110011010

27. The data staging area is used for

    - querying the DW
    - data transformations and cleansing
    - indexing dimensions

28. In the ETL process, what must be updated first?

    - Dimensions
    - Fact table
    - Indices

29. What is a good strategy for ETL?

    - Implement all transformation in one single programm
    - Implement the transformations in a sequence of small operation
    - Implement the transformations in the source database

30. Which of the following solutions for slowly changing dimensions captures the correct information in the DW?

    - Solution 1
    - Solution 2
    - Solution 3