Distributed Information Systems
Class questions

Marc Bourqui

May 21, 2015

Part I
Introduction

An Overview

Information Systems (Week 1)

Functions in models

- ○ Are always computable
- ○ Can always be represented as data
- ○ Can be constrained by axioms

2. Interpretation relationships
   - ○ Are always computable
   - ○ Relate constants to real-world entities
   - ○ Are uniquely defined

Data Management

3. What is not specified in the data definition language ?
   - ○ The structure of a relational table
   - ○ The query of user
   - ○ A constraint on a relational table
4. Logical data independence means
   - ○ An abstract data type is implemented using different data structures
   - ○ A new view is computed without changing an existing database schema
   - ○ A model can be represented in different data modelling formalisms

Data Management Tasks

5. Which is wrong ? An index structure
   - ○ Is created as part of physical database design
   - ○ Is selected during query optimization
   - ○ Accelerates search queries
   - ○ Accelerates tuple insertion
6. Persistence means that
   - ○ A change of a transaction on a database is never lost after it is completed
   - ○ The state of a database is independent of the lifetime of a program
   - ○ The same logical database can be stored in different ways on a storage medium

Information Management

7. Grouping Twitter users according to their interest by analyzing the content of their tweets is
   - ○ A retrieval task
   - ○ A data mining task
   - ○ An evaluation task
   - ○ A monitoring task

Distributed Information Systems

8. Creating a web portal for comparing product prices is (primarily) a problem of
   - ○ Distributed data management
   - ○ Heterogeneous data integration
   - ○ Collaboration among autonomous systems

Distributed Data Management

9. When you open a Web page with an embedded Twitter stream, the communication model used by Twitter is
   - ○ Push, unicast and conditional
   - ○ Pull, multicast and ad-hoc
   - ○ Push, multicast and ad-hoc
   - ○ Pull, unicast and conditional

Heterogeneity

10. Creating a web portal for comparing product prices requires to address
    - ○ Syntactic heterogeneity
    - ○ Semantic heterogeneity
    - ○ Both
11. An ontology is a
    - ○ Sdatabase
    - ○ database schema
    - ○ data model
    - ○ data modeling formalism
    - ○ model

Autonomy

12. Trust is

- ◯ A quality of information
- ◯ A quality of a user
- ◯ A quality of the relationship among user and information
- ◯ A quality of the relationship among users

Part II
Storage

Distributed Data Management

Schema Fragmentation

Relational Databases

13. At which phase of the database lifecycle is fragmentation performed ?

- √ **At database design time**
- ◯ During distributed query processing
- ◯ During updates to a distributed database

14. The reconstruction property expresses that

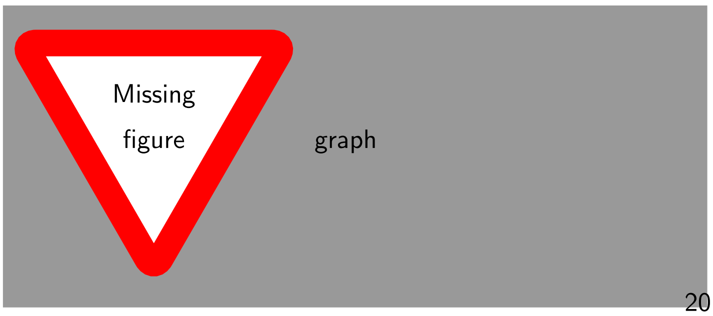
- ◯ In case of a node failure the data can be recovered from a fragment from another node
- √ **The original data can be fully recovered from the fragments**
- ◯ Every data value of the original data can be found in at least one fragment

Primary Horizontal Fragmentation (Week 2)

15. Example: application A1 accesses

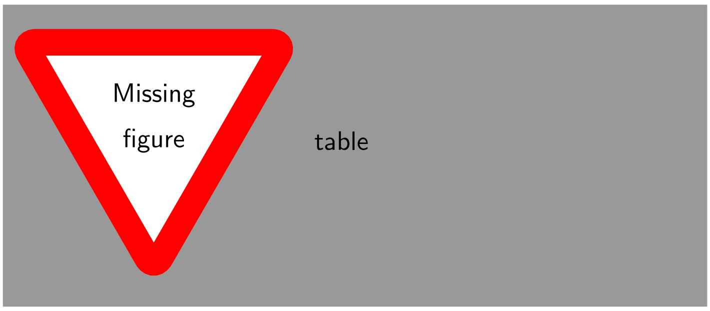
    1. Fragment F1: with frequency 3
    2. Fragment F2: with frequency 1

    A1 accesses the whole relation with frequency



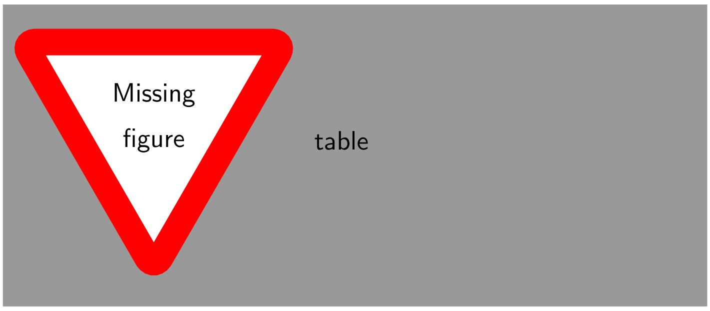
- √ $^{13}/_{7}$
- ◯ $^{4}/_{7}$
- ◯ $^{14}/_{7}$

16. Consider the access frequencies below: How many horizontal fragments would a minimal and complete fragmentation have?



- √ **3**
- ◯ 4
- ◯ 6

17. Which of the following sets of simple predicates is complete?



- ◯ Location = "Munich", Budget > 200000
- ◯ Location = "Munich", Location = "Bangalore"
- ◯ Location = "Paris", Budget $\leq$ 200000
- √ **None of those**

18. Which is true for MinFrag algorithm?

- ◯ The output is independent of the order of the input
- ◯ It produces a monotonically increasing set of predicates
- √ **It always terminates**
- ◯ All of the above statements are true

19. When deriving a horizontal fragmentation for relation S from a horizontally fragmented relation R

- √ **Some primary key attribute in R must be a foreign key in S**
- ◯ Some primary key attribute in S must be a foreign key in R
- ◯ Both are required

Graph Databases

Semi-structured Data (Week 3)

20. Semi-structured data

- ◯ Is always schema-less
- √ **Always embeds schema information into the data**
- ◯ Must always be hierarchically structured
- ◯ Can never be indexed

21. Why is XML a document model?
    - ◯ It supports application-specific markup
    - ◯ It supports domain-specific schemas
    - √ **It has a serialized representation**
    - ◯ It uses HTML tags

Graph Data Model

22. In a graph database
    - ◯ There is a unique root node
    - √ **Each node has a unique identifier**
    - ◯ Data values in leaf nodes are unique
    - ◯ The labels of edges leaving a node are different
    - ◯ There is a unique path from the root to each leaf
23. The simulation relationship is a relation
    - √ **Among nodes in the data and schema graph**
    - ◯ Among edges in the data and schema graph
    - ◯ Among sets of nodes in the data and schema graph
    - ◯ Among sets of edges in the data and schema graph
24. Which is true?
    - ◯ For each labelled edge in S a corresponding edge in D can be identified
    - ◯ For each root node in S a corresponding root node D can be identified
    - √ **For each leaf node in D a corresponding typed node in S can be identified**
    - ◯ For each node in S a unique path reaching it from a root node can be identified
25. If there exists a uniquely defined simulation relationship among a graph database D and a schema graph S
    - ◯ The data and schema graph are simulation equivalent
    - √ **Ambiguous classification cannot occur**
    - ◯ Multiple classification cannot occur
26. If schema graph S_1 subsumes S_2
    - ◯ Every graph database corresponding to S_1 corresponds also to S_2
    - √ **S_2 simulates S_1**
    - ◯ S_1 has fewer nodes than S_2

Schema Extraction

27. Which is wrong? In a dataguide
    - ◯ Every path in the data graph occurs only once
    - √ **Every node in the data graph occurs only in one data guide node**
    - ◯ Every data guide node has a unique set of nodes
    - ◯ A leaf node in the data graph corresponds always to a leaf node in the data guide
28. In a non-deterministic schema graph
    - √ **Every node of the data graph occurs exactly once**
    - ◯ Every path of the data graph occurs at most once
    - ◯ Every label of an outgoing edge of a node in the schema graph is unique

Part III
Search

Information Retrieval and Data Mining

Information Retrieval

Information Retrieval (Week 4)

29. A retrieval model attempts to model
    - ◯ The interface by which a user is accessing information
    - √ **The importance a user gives to a piece of information**
    - ◯ The formal correctness of a query formulation by user
    - ◯ All of the above
30. If the top 100 documents contain 50 relevant documents
    - ◯ The precision of the system at 50 is 0.5
    - √ **The precision of the system at 100 is 0.5**
    - ◯ The recall of the system is 0.5
    - ◯ None of the above
31. If retrieval system A has a higher precision than system B
    - ◯ The top k documents of A will have higher similarity values than the top k documents of B
    - √ **The top k documents of A will contain more relevant documents than the top k documents of B**
    - ◯ A will recall more documents above a given similarity threshold than B
    - ◯ Relevant documents in A will have higher similarity values than in B

Text-based Information Retrieval

32. Full-text retrieval means that
    - ◯ The document text is grammatically deeply analyzed for indexing
    - ◯ The complete vocabulary of a language is used to extract index terms
    - √ **All words of a text are considered as potential index terms**
    - ◯ All grammatical variations of a word are indexed
33. The term-document matrix indicates
    - √ **How many relevant terms a document contains**
    - ◯ How relevant a term is for a given document
    - √ **How often a relevant term occurs in a document collection**
    - √ **Which relevant terms are occurring in a document collection**

34. Let the query be represented by the following vectors: (1, 0, -1) (0, -1, 1); the document by the vector (1, 0, 1)
    - ◯ Matches the query because it matches the first query vector
    - √ **Matches the query because it matches the second query vector**
    - ◯ Does not match the query because it does not match the first query vector
    - ◯ Does not match the query because it does not match the second query vector

35. Which is right? The term frequency is normalized
    - √ **By the maximal frequency of a term in the document**
    - ◯ By the maximal frequency of a term in the document collection
    - ◯ By the maximal frequency of a term in the vocabulary
    - ◯ By the maximal term frequency of any document in the collection

36. The inverse document frequency of a term can increase
    - ◯ By adding the term to a document that contains the term
    - √ **By adding a document to a document collection that does not contain the term**
    - ◯ By removing a document from the document collection that does not contain the term
    - ◯ By adding a document to a document collection that contains the term

Advanced Retrieval Models

Latent Semantic Indexing (Week 5)

37. In vector space retrieval each row of the matrix $\mathbf{M}^T$ corresponds to
    - √ **A document**
    - ◯ A concept
    - ◯ A query
    - ◯ A query result

38. Applying SVD to a term-document matrix **M**. Each concept is represented
    - ◯ As a singular value
    - √ **As a linear combination of terms of the vocabulary**
    - ◯ As a linear combination of documents in the document collection
    - ◯ As a least square approximation of the matrix **M**

39. The number of term vectors in the SVD for LSI
    - ◯ Is smaller than the number of rows in the matrix **M**
    - √ **Is the same as the number of rows in the matrix M**
    - ◯ Is larger than the number of rows in the matrix **M**

40. A query transformed into the concept space for LSI has
    - √ s **components (number of singular values)**
    - ◯ m components (size of vocabulary)
    - ◯ n components (number of documents)

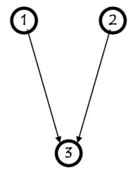
User Relevance Feedback

41. Can documents which do not contain any keywords of the original query receive a positive similarity coefficient after relevance feedback ?
    - ◯ No
    - ◯ Yes, independent of the values β and γ
    - √ **Yes, but only if** $\beta > 0$
    - ◯ Yes, but only if $\gamma > 0$

Link-based Ranking

42. A positive random jump value for exactly one node implies that
    - √ **a random walker can leave the node even without outgoing edges**
    - ◯ a random walker can reach the node multiple times even without outgoing edges
    - √ **a random walker can reach the node even without incoming edges**
    - ◯ none of the above

43. Given the graph below and an initial hub vector of $(1, 1, 1)$. The hub-authority ranking will result in the following



    - ◯ authority vector $(0, 0, 1)$; hub vector $(1, 1, 0)$
    - ◯ authority vector $(0, 0, 2)$; hub vector $(2, 2, 0)$
    - √ **authority vector** $(0, 0, 1)$ **; hub vector** $(\frac{1}{2}, \frac{1}{2}, 0)$
    - ◯ authority vector $(0, 0, 2)$; hub vector $(1, 1, 0)$

Inverted Files (Week 6)

44. A posting indicates
    - ◯ The frequency of a term in the vocabulary
    - ◯ The frequency of a term in a document
    - √ **The occurrence of a term in a document**
    - ◯ The list of terms occurring in a document

45. When indexing a document collection using an inverted file, the main space requirement is implied by
    - ◯ The access structure
    - ◯ The vocabulary
    - ◯ The index file
    - √ **The postings file**

46. Using a trie in index construction
    - ◯ Helps to quickly find words that have been seen before
    - ◯ Helps to quickly decide whether a word has not been seen before
    - ◯ Helps to maintain the lexicographic order of words seen in the documents
    - √ **All of the above**

47. Maintaining the order of document identifiers when partitioning the document collection is important
    - √ **In the index merging approach for single node machines**
    - ○ In the map-reduce approach for parallel clusters
    - ○ In both
    - ○ In neither of the two

Distributed Retrieval

48. When applying Fagin's algorithm for a query with three different terms for finding the k top documents, the algorithm will scan
    - ○ 2 different lists
    - √ **3 different lists**
    - ○ k different lists
    - ○ it depends how many rounds are taken

49. Once k documents have been identified that occur in all of the lists
    - ○ These are the top-k documents
    - √ **The top-k documents are among the documents seen so far**
    - ○ The search has to continue in round-robin till the top-k documents are identified
    - ○ Other documents have to be searched to complete the top-k list

Peer-2-Peer Search

Peer-2-Peer Systems

P2P Systems and Resource Location (Week 7)

50. Which resource is in Napster not shared in a P2P approach ?
    - ○ File storage
    - √ **File metadata storage**
    - ○ Network bandwidth
    - ○ Content rights

51. "Churn" refers to the fact that in a peer-to-peer system :
    - √ **Peers constantly join and leave the network**
    - ○ Peers constantly add and remove resources
    - ○ Peers constantly search for resources

52. An "overlay network" supports :
    - ○ Efficient routing to a given IP address
    - √ **Efficient routing to the location of a resource identifier**
    - ○ Efficient exchange of large files
    - ○ Efficient messaging in centralized social network

Unstructured P2P Overlay Networks

53. In an unstructured overlay network (such as Gnutella) a peer receiving a "peer discovery" message (ping)
    - ○ Responds by sending a message to the originator of the message
    - √ **Responds by replying to the last forwarder of the message**
    - ○ Responds by sending a message to all its neighbors

54. If the largest city in the world has 16 Mio inhabitants, the second largest 11.3 Mio inhabitants, the third largest 9.2 Mio, the fourth largest 8.0 Mio, and so on, then this is
    - ○ A Powerlaw distribution
    - √ **A Zipf distribution**
    - ○ None of the two

55. Assume that in a country the size of cities follows a powerlaw distribution with exponent 2. A city of 16 Mio inhabitants has probability of $1/256$ to occur. Then a city of 8 Mio inhabitants is
    - ○ Twice as probable
    - √ **Four times as probable**
    - ○ Eight times as probable

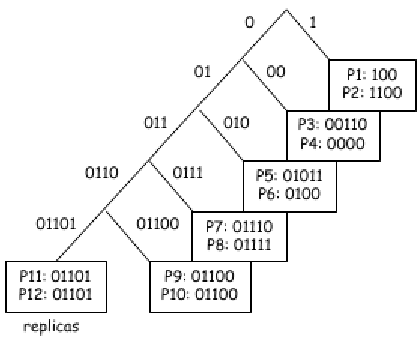
56. Expanding ring search is particularly suitable to locate
    - √ **Frequent items**
    - ○ Rare items
    - ○ Does not matter

57. With the square root rule for replica allocation : given two items that are accessed with probabilities $p_1 > p_2$ that are replicated r_1 and r_2 times. Which is always true ?
    - ○ $r_1 < r_2$
    - √ $r1/p1 < r2/p2$
    - ○ $r_1 - p_1 < r_2 - p_2$

Hierarchical P2P Overlay Networks (Week 8)

58. The index information in a structured overlay network
    - √ **Provides references to route a search request within the overlay network**
    - ○ Provides for a given key the reference to the peer that stores the resource
    - ○ Is replicated in routing tables to support redundant search paths

59. For the given routing table, the search request for the key 0101 is routed

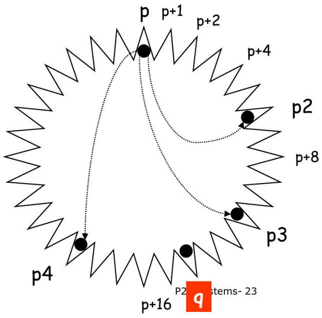


- √ **Always to peer P_5**
- ○ Either to peer P_5 or P_6
- ○ Either to peer P_3, P_4, P_5 or P_6

60. When routing in Chord
- √ **The next hop is always uniquely determined**
- ○ The next hop can be chosen among a constant number of possible candidates
- ○ The next hop can be chosen among $\log n$ possible candidates

61. When adding q to the Chord ring : in the routing table of p

i	s_i
1	p_2
2	p_2
3	p_2
4	p_3
5	p_4



- ○ Entries for $i = 1, 2, 3, 4$ change
- ○ The entry for $i = 4$ changes
- ○ The entry for $i = 5$ changes
- √ **No entry changes**

62. When adding n peers to CAN the number of new zones
- √ **Is exactly n**
- ○ It depends what the keys of the peers were
- ○ It depends on the dimensionality of the key space

Solution: One zone per new peer.

63. In CAN, for a fixed dimensionality $d > 2$, when moving from 1 to 2 realities
- ○ The number of entries in the routing table increases by 2
- ○ The number of entries in the routing table increases by d
- √ **The number of entries in the routing table doubles**

64. In FreeNet the routing table is updated
- ○ When a search request message arrives
- √ **When a query answer message arrives**
- ○ When an insert file message arrives

65. For which of the following structured overlay networks the length of a search path is always guaranteed to be shorter than the length of the longest key
- √ **P-Grid**
- ○ CAN
- ○ FreeNet

66. The local clustering coefficient is the probability that two of my friends are also friends. If I have 10 friends and among them 15 friendships exist, my local clustering coefficient is
- ○ $1/6$
- √ $1/3$
- ○ $2/3$
- ○ $3/2$

Solution: Look at the formula in the slides notes.

67. A random graph has
- ○ High clustering and low diameter
- ○ High clustering and high diameter
- √ **Low clustering and low diameter**
- ○ Low clustering and high diameter

68. In a three-dimensional Kleinberg small world network with $\log n$ long range links the search cost is
- √ $\log n$
- ○ $\log^2 n$
- ○ $\log^3 n$

Part IV

Dissemination

Data Broadcasting in Mobile Networks (Week 9)

69. Latency is
- ○ The time a client is connected to a broadcast channel
- ○ The time a client listens actively on a broadcast channel
- √ **The time a client waits for receiving a data item on a broadcast channel**

70. Data Broadcast is beneficial when
- ○ Clients have a high upstream bandwidth
- √ **Many clients are interested in the same information**
- ○ Clients have many different requests

71. Assume the broadcast channel has one item accessed with frequency 9 and three others accessed with frequency 1. The expected delay for accessing the first item in an optimal broadcast organization will be
- √ **1**
- ○ 2
- ○ 3

72. Assume the broadcast channel has one item accessed with frequency 9 and three others accessed with frequency 1. The expected delay for accessing the second type of items will be

○ 1

√ **3**

○ 6

73. When organizing a broadcast disk a "chunk"

○ Contains always all elements of the broadcast disk

√ **Contains sometimes all elements of the broadcast disk**

○ Contains never all elements of the broadcast disk

74. When organizing a broadcast disk, which is true ?

√ **The number of copies of different chunks in a broadcast disk is constant**

○ The number of copies of different data items in a broadcast disk is constant

√ **The number of data items in the chunks of one disk is constant**

○ The data items in the chunks of one disk are always the same

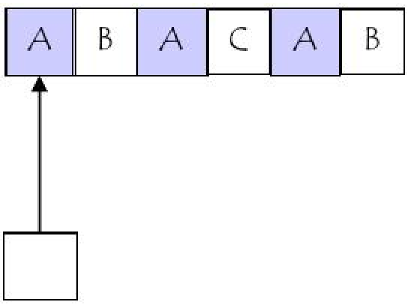
75. Which is true ?

√ **LRU (least recently used) is not optimal because it does not consider the frequency of data items in a data broadcast**

√ **MPA (most probable accessed) is not optimal because it does not consider the frequency of data items in a data broadcast**

√ **Only PIX considers the frequency of data items in a data broadcast**

76. Assume the broadcast and access pattern below. Assuming that $c = 1/2$ what is the access frequency estimate for B at time 6 ?



○ ⅓

√ **¼**

○ ⅙

○ 1⁄12

Solution:
At t_2, B has value $\frac{1/2}{2-0} + 0 = \frac{1}{4}$.
At t_6, B has value $\frac{1/2}{6-2} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{4}$

77. The minimal latency of a broadcast channel can be achieved

√ **By not indexing the broadcast**

○ By indexing the broadcast only once

○ By indexing the broadcast according to the (1,m) rule

78. The term "probe wait" refers to

○ The time for waiting for a data page

√ **The time for waiting for an index segment**

○ The time for waiting for a data segment

Part V

Big Data Analytics

Association Rules (Week 10)

79. Based on the analysis of search terms and subsequent link clicks, a search engine provider places ads on search results that are most likely to be clicked by the users. This task is an example of :

○ Local rule discovery

√ **Predictive modelling**

○ Descriptive modelling

○ Exploratory data analysis

Pattern structure

80. Let's assume that the transactions are stored in a relation $T(x, A1, \ldots, A5)$, where x is the customer and each attribute $A1, \ldots, A5$ can have 3 different values. How many different items exist after reduction to a single dimension ?

○ 5

○ 243

○ 125

√ **15**

Scoring function

81. 10 itemsets out of 100 contain item A, of which 5 also contain B. The rule A→ B has :

○ 5% support and 10% confidence

○ 10% support and 50% confidence

√ **5% support and 50% confidence**

○ 10% support and 10% confidence

Solution: 5/100 transactions which have A and B support, confidence half of the time we buy B when we buy A

82. 10 itemsets out of 100 contain item A, of which 5 also contain B. The rule B→A has :

○ unknown support and 50% confidence

○ unknown support and unknown confidence

○ 5% support and 50% confidence

√ **5% support and unknown confidence**

83. Given the frequent 2-itemsets {1,2}, {1,4}, {2,3} and {3,4}, how many 3-itemsets are generated and how many are pruned ?

    ◯ 2, 2

    ◯ 1, 0

    √ **1, 1**

    ◯ 2, 1

84. After the join step, the number of $k+1$-itemsets ...

    ◯ is equal to the number of frequent k-itemsets

    √ **can be equal, lower or higher than the number of frequent k-itemsets**

    ◯ is always higher than the number of frequent k-itemsets

    ◯ is always lower than the number of frequent k-itemsets

    Solution: {1,2,3}, {1,2,4} → {1,2,3,4} {1,2,5} → {1,2,3,5},{1,2,4,5}

85. If rule {A,B}→{C} has confidence c 1 and rule {A}→{C} has confidence c 2 , then ...

    √ $c_2 \geq c_1$

    ◯ $c_1 > c_2$ and $c_2 > c_1$ are both possible

    ◯ $c_1 \geq c_2$

Clustering & Classification (Week 11)

Clustering

86. Suppose we have a dataset of pictures and we want to cluster them. Which partitioning algorithm seems more appropriate?

    ◯ k-medoids

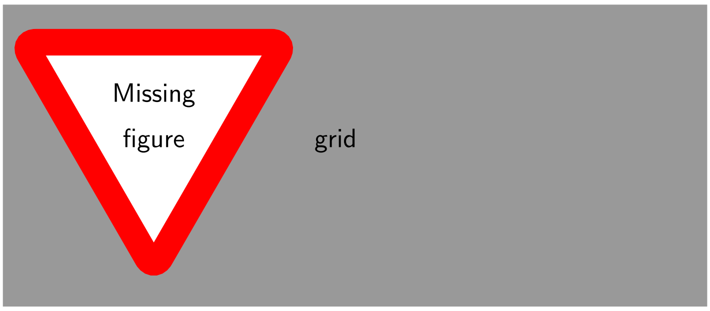
    ◯ k-medians

    ◯ k-means

    √ **none of the above**

Classification

87. What will be the color of the middle points after convergence?



    ◯ Green

    ◯ Yellow

    ◯ Blue

    ◯ k-means does not converge

88. If a classifier has 75% accuracy, it means that ...

    ◯ correctly classifies 75% of the data items in the training set

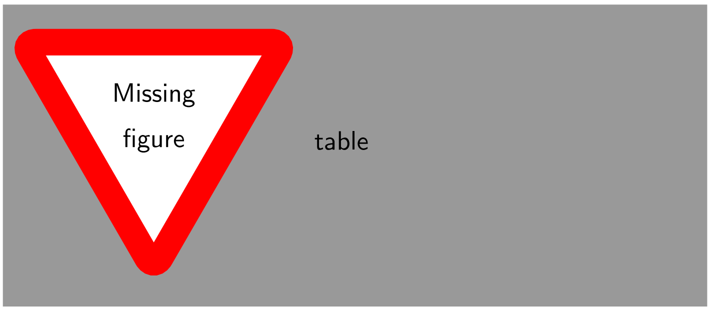
    ◯ It correctly classifies 100% of the data items in the training set but only 75% in the test set

    √ **It correctly classifies 75% of the data items in the test set**

    ◯ It correctly classifies 75% of the unknown data items

    Solution: A model that fits 100% of the training data might be too complex and give poor results on the test set.

89. Given the distribution of positive and negative samples for attributes A_1 and A_2, which is the best attribute for splitting ?



    √ A_1

    ◯ A_2

    ◯ They are the same

    ◯ There is not enough information to answer the question

    Solution: Entropy of A_1 =
    Entropy of A_2 = 0.8 * 0.5 + 0.8 * 0.5

Credits

Quiz questions were taken from the lecture notes of Prof. K. Aberer. Answers are provided with no guarantee.