About the exam

Høgskolen i Østfold

EXAMINATION

Course Code and name:
ITF301416 Big Data - Processing and Analysis

Date and duration:
29.11.2018, 4 hours

No support materials permitted.

Instructors:
Edgar Bostrøm
Cathrine Linnes

Please note:

- You are able to view only one question at the time.
- You may go back to view each question before you submit your exam.
- You can have the exam question and attchment link open at the same time.

Part 1 (120 minutes)
Consists of essays questions (3 tasks, 50 points).

- Task 1 (6 questions)
- Task 2 (2 questions)
- Task 3 (2 questions)

Part 2 (120 minutes)
Please refer to Appendix A while answering some of these questions.
Question 1-50

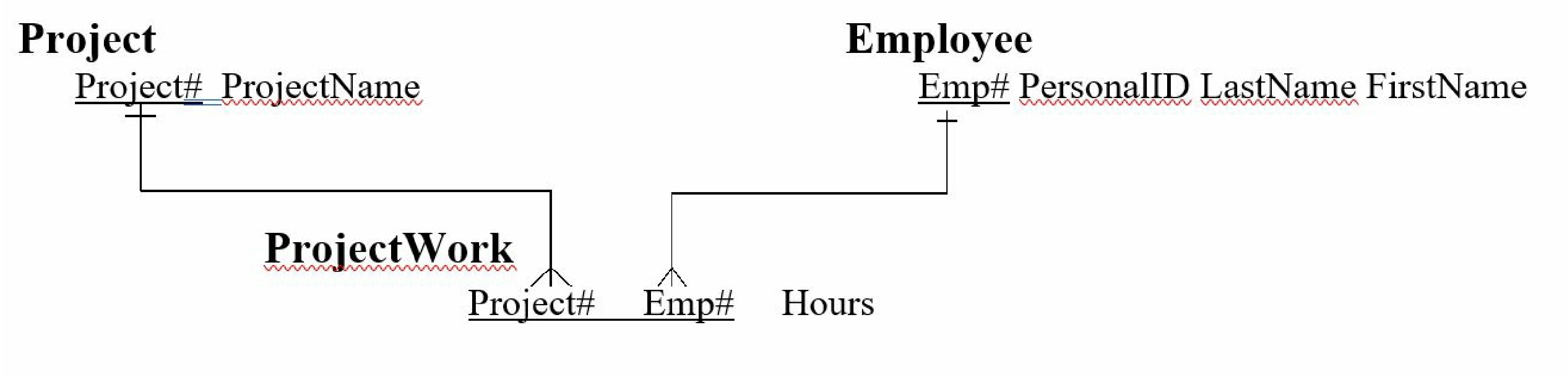
- Multiple Choice and True and False (50 points)

Results:
Test results will be available on Studentweb 20 December 2018.

1 Task 1

Task 1.   Time: 60 minutes.
The questions to be answered in relational algebra uses the three relations below (from an earlier project). Use latin letters (in Inspera) or write the answers on a separate sheet.

(Syntax found as .PDF file).



a. List **FirstName, LastName** for persons who have worked more than 200 hours in Project# = 7001.

b. List **Project#, ProjectName** where PersonlID = ‘010203 12312’ has participated.
Note: preferably using semijoin.

c. List Emp#, FirstName together with any projects (Project#, ProjectName) for person(s) with LastName = ‘Johnson’. The list should look something like:

```
Emp#  FirstName  Project#   ProjectName
  34         Ann           7001     New pavement .....
  34         Ann           7101     Renew .....
  35         Bart           null       null  (i.e. has no project)
  36        Cindy         7001     Renew .....
  ............................................................
```

d. List **Emp#, LastName, FirstName** for employee who have never participated in any projects.
Tip: First, find the Emp# for Employees which you cannot find in ProjectWork. Then, join with Project.

e. List **Emp#, LastName, FirstName** for employees who have participated in all the projects.

f. (not connected to the 3 relations above). How can relational algebra be used to explain distributed databases?

Fill in your answer here

2 Task 2

Task 2. Triggers. 30 minutes

a. What is the difference between stored procedures, stored functions and triggers?

b. When is it appropriate to use triggers, when is it not appropriate?

Fill in your answer here

3 Task 3

Task 3. Big data. 30 minutes

a. When would you recommend using noSQL systems, when would you reccomend using SQL systems?

b. Describe and compare the key-value, the document based and the column-based database types.

.

Fill in your answer here

i

Appendix

Use this attachment to help you answer some of the questions in section 2.

Appendix_A_BigData

4

Q1

Can data mining be applied in the following area: Image screening, e.g. detecting potential oil slicks in the sea from satellite data?

Select an alternative:

○ True

○ False

5

Q2

Can data mining be applied in the following area: Load forecasting, e.g. forecasting the demand for electricity for a particular hour, day, month, and year?

Select an alternative:

○ True

○ False

6

Q3

Data mining is about solving problems by analyzing data that is currently not available in the databases.

Select an alternative:

○ True

○ False

7 **Q4**

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage e.g. an economic advantage.

Select an alternative:

○ True

○ False

8 **Q5**

The four interfaces in Weka are: Explorer, Experimenter, Knowledge Flow and Workbench?

Select an alternative:

○ True

○ False

9 **Q6**

ZeroR is the baseline classifier?

Select an alternative:

○ True

○ False

10 **Q7**

The default percentage split is set to 10%.

Select an alternative:

○ True

○ False

11 **Q8**

Weka accepts numeric files only.

Select an alternative:

○ True

○ False

12 **Q9**

A confusion matrix is also known as an error matrix.

Select an alternative:

○ True

○ False

13 **Q10**

Supervised files are the ones that use a class value for their operation.

Select an alternative:

○ True

○ False

14 **Q11**

In the visualization tab, the classifier errors are displayed using the symbol ”square”.

Select an alternative:

○ True

○ False

15 **Q12**

Looking at the segment-challenge dataset. With a training set percentage of 99% it gives a figure of 100% accuracy on the test set. Does this mean that this generates a perfect classifier?

Select an alternative:

○ True

○ False

16 **Q13**

How many instances are there in the IRIS data set?
Select an alternative:

○ 100

○ 200

○ 250

○ 150

17

Q14

How many attributes are there?
Select an alternative:

○ 7

○ 10

○ 4

○ 5

18

Q15

How many possible values does the class attributes have?
Select an alternative:

○ 50

○ 1

○ 2

○ 3

19

Q16

Weka data files are___files.
Select an alternative:

○ ARFF

○ ARC

○ CSV

○ DOC

20 **Q17**

Examine the IRIS file (header) and say when the dataset was first used by looking at the NotePad handout?

Select an alternative:

○ 1988

○ 1980

○ 1936

○ 1973

21 **Q18**

Choose the J48 tree classifier, and run it (with default parameters). How many instances are misclassified?

Select an alternative:

○ 1

○ 6

○ 4

○ 2

22 **Q19**

Now switch the classifier to Simple Logistic, which you will find in the functions category, and run it (with default parameters). How many instances are misclassified now?

Select an alternative:

○ 12

○ 6

○ 9

○ 3

○ 15

23 **Q20**

If you did the above experiment with 10 different random seeds rather than 5, how would you expect this to affect the mean and standard deviation?

Select an alternative:

○ The mean would be a bit bigger but the standard deviation would be about the same.

○ The mean would be about the same and the standard deviation would be a little smaller.

○ Both the mean and standard deviation would be a bit smaller.

○ They would both stay about the same

24 **Q21**

What should be the first row of a Comma Separated Values (.csv) format file that contains the nominal Weather data?

Select an alternative:

○ Outlook, temperature, humidity, windy, play

○ Sunny, hot, high, TRUE, no

○ Rainy, mild, high, TRUE, no

○ Sunny, hot, high, FALSE, no

25 **Q22**

The problem of finding hidden structure in unlabeled data is called

Select an alternative:

○ None of the above

○ Supervised learning

○ Reinforcement learning

○ Unsupervised learning

26 **Q23**

You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is an example of

Select an alternative:

○ Unsupervised learning

○ Reduction learning

○ All of the above

○ Supervised learning

27

Q24

Algorithm is

Select an alternative:

○ Computational procedure that takes some value as input and produces some value as output.

○ It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations.

○ Science of making machines perform tasks that would require intelligence when performed by humans.

○ None of the above

28

Q25

Classification is

Select an alternative:

○ The task of assigning a classification to a set of examples

○ None of the above

○ A subdivision of a set of examples into a number of classes

○ A measure of the accuracy, of the classification of a concept that is given by a certain theory.

29

Q26

Self-organizing maps (type of artificial neural network) are an example of

Select an alternative:

○ Supervised learning

○ Reinforcement learning

○ Unsupervised learning

○ Missing data inputation

30

Q27

Bayesian classifiers is

Select an alternative:

○ An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.

○ A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.

○ None of the above.

○ Any mechanism employed by learning system to constrain the search space of a hypothesis

31 Q28

One of the most frequently used data mining algorithms is

Select an alternative:

○ Cluster analysis

○ Decision trees

○ Evolution analysis

○ Outlier analysis

32 Q29

Which of the following is not a data mining method?

Select an alternative:

○ A priori algorithms

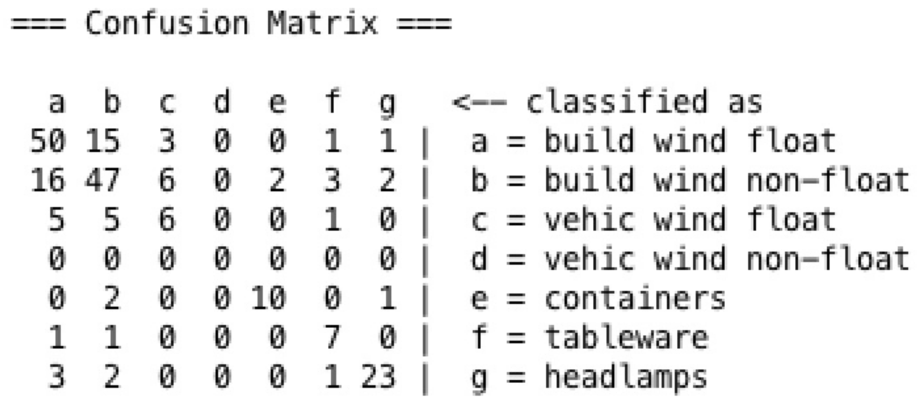
○ Data description

○ Classification and prediction

○ Dependency analysis

33 Q30

Look at the glass dataset, go to the Classify panel, choose the J48 tree classifier, and run it (with default parameters). Using the confusion matrix to determine how many headlamps instances were misclassified as build wind float?



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Select an alternative:

○ 3

○ 1

○ 6

○ 2

○ 7

34 **Q31**

Which of the following statements seems to be correct, based on your experiments?

Select an alternative:

○ The more test data, the greater the classifier's success rate.

○ Training set size has no influence on the classifier's success rate.

○ The more training data, the greater the classifier's success rate.

35 **Q32**

When the percentage split option is used for evaluation, how good is the performance if (a) almost none of the data is used for testing; (b) almost all of the data is used for testing?

Select an alternative:

○ has poor performance, (b) has good performance.

○ The performance for (a) and (b) are similar.

○ (a) has better performance than (b).

36 **Q33**

In supervised learning:

Select one alternative:

○ the desired output and the algorithms are known

○ both the input and the desired outputs are known

○ the network is controlled by the user

○ only input stimuli are fed into the network

37 **Q34**

In unsupervised learning:

Select one alternative:

○ the desired output and the algorithms are known

○ both the inputs and the desired outputs are known

○ both the inputs and the desired outputs are known

○ only input stimuli are fed into the network

38 Q35

Data mining requires a separate, dedicated database.

Select one alternative:

○ True

○ False

39 Q36

Data mining is a way for companies to develop business intelligence from their data to gain a better understanding of their customers and operations and to solve complex organizational problems.

Select one alternative:

○ True

○ False

40 Q37

The first step in the data mining process is to understand the relevant data from the available databases.

Select one alternative:

○ True

○ False

41 Q38

The term *data mining* was originally used to ______.

Select one alternative:

- ○ include most forms of data analysis in order to increase sales
- ○ describe the prices through which previously unknown patterns in data were discovered
- ○ describe the analysis of huge datasets stored in data warehouses
- ○ All of the above

42 Q39

Why has data mining gained the attention of the business world?

Select one alternative:

- ○ All of the above
- ○ More intense competition at the global scale driven by customers´ever-changing needs and wants in an increasingly saturated marketplace.
- ○ Consolidation and integration of database records, which enables a single view of customers and vendors.
- ○ Significant reduction in the cost of hardware and software for data storage and processing.

43 Q40

What is a major characteristic of data mining?

Select one alternative:

- ○ Because of the large amounts of data and massive search efforts, it is sometimes necessary to use serial processing for data mining
- ○ The miner needs sophisticated programming skills.
- ○ Data mining tools are readily combined with spreadsheets and other software development tools.
- ○ Dat are often buried within numerous small large databases, which sometimes contain data from several years.

44 Q41

Because the latter steps in the data mining process are built on the outcome of the former ones, one should:

Select one alternative:

○ pay extra attention to the earlier steps in order not to put the whole study on an incorrect path from the onset.

○ start with an understanding of the relevant data.

○ work quickly through the early steps and work in-dept on the latter steps.

○ start by cleaning the relevant data and storing it in a single data warehouse.

45 **Q42**

The simple split methodology splits the data into two mutually exclusive subsets called ____________ set and a ____________ set.

Select one alternative:

○ training; test

○ holdout; training

○ positive; negative

○ matrix; test

46 **Q43**

Data mining is a prime candidate for better management of companies that are data-rich, but knowledge-poor.

Select one alternative:

○ True

○ False

47 **Q44**

In class we ran an experiment with 10 different random seeds rather than 5, if we ran 5 how would you expect this to affect the mean and standard deviation.

Select one alternative:

○ The mean would be a bit bigger but the standard deviation would be about the same.

○ They would both stay about the same.

○ Both the mean and standard deviation would be a bit smaller.

○ The mean would be about the same and the standard deviation would be a little smaller.

48 **Q45**

The complexity of today's business environment creates many new challenges for organizations, such as global competition, but creates few new opportunities in return.

Select one alternative:

○ True

○ False

49
Q46

In ________, the problem is to group an unlabelled collection of objects, such as documents, customer comments, and Web pages into meaningful groups without any prior knowledge.

Select one alternative:

○ classification

○ grouping

○ search recall

○ clustering

50
Q47

Unsupervised learning is a process of inducing knowledge from a set of observations.

Select one alternative:

○ True

○ False

51
Q48

Supervised learning uses a set of inputs for which the desired outputs are known. For example, a dataset of loan applications with the success or failure of borrowers to repay their loans has a set of input parameters and known outputs.

Select one alternative:

○ True

○ False

52
Q49

Data mining can be very useful in detecting patterns such as credit card fraud, but is of little help in improving sales.

Select one alternative:

○ True

○ False

53

Q50

According to the Simplified Taxonomy of Machine Learning, all of the following are types of learning except:

Select one alternative:

○ Reinforcement learning

○ Unsupervised learning

○ Supervised learning

○ Unreinforced learning

Question 1

Attached

Relational algebra – useful operators.

Set operators	*Notation, variant-1*	*Notation without special characters*
Union	$R \cup S$	R union S
Intersect / Snitt	$R \cap S$	R intersect S
Set difference / Mengdedifferanse	$R - S$, $R \setminus S$	R difference S
Set product / Mengdeprodukt, cartesian product (“all joined with all”)	$R \times S$	R product S R times S
Relational operators		
Restriction[1]/Horizontal selection (sigma)	$\sigma_{<cond.>}(R)$	SIGMA $_{<cond.>}$ (R)
Projection/Vertical selection (pi)	$\pi_{<attribute\ list>}(R)$	PI$_{<attribute\ list>}$ (R)
Set division. (Given R[c,d] and S[d]. c is a member of the set R divided by S if c in R is found with all d-s in S.)	$R \div S$ R / S	R divideby S or R div S
Different types of product		
θ-join (product with some kind of condition on compatible attributes, e.g. >, <, =, !=, incl. combinations.)	$R \bowtie_{\theta} {}_{<cond.>} S$	R THETAJOIN$_{<condition>}$ S
Equi-join (the θ-operation is =)	$R \bowtie_{=} {}_{<cond.>} S$	R EQUIJOIN$_{<condition>}$ S
Natural join (equi-join where duplicate attributes are removed) ** most common join type **	$R \bowtie_{<cond.>} S$	R NATURALJOIN$_{<condition>}$ S, or just R JOIN$_{<condition>}$ S
Varieties for product[2]		
Outer join, normally left (all in R, plus all in S complying the join condition)	R ⟕ S	R LEFTOUTERJOIN$_{<cond.>}$ S
Right outer join (all in S, plus all in R complying the join condition)	R ⟖ S	R RIGHTOUTERJOIN$_{<cond.>}$ S
Full join (all in R, all in S, plus all complying with the join condition)	R ⟗ S	R FULLJOIN$_{<cond.>}$ S
Semijoin (those in R complying with R join$_{<condition>}$ S)	$R \triangleright_{<cond.>} S$[3]	R SEMIJOIN$_{<cond.>}$ S

Note: the operations are set-based, i.e. duplicates are deleted – corresponding to select distinct i SQL. Relational algebra is **set-based**, while SQL is **bag-based.**

[1] Originally, restriction was called selection (thereby sigma - σ). However, this makes it easy to be confused with SQL-seletion, whick in fact is a projection.

[2] If joining on primary/foreign key combination, we may drop <cond>.

[3] Some use this symbol for antijoin (not in), and use ⋉ for semijoin.