UPPSALA
UNIVERSITET

**Department of Information Technology**

| | |
|---|---|
| **FRONT SHEET FOR EXAMS** | **DATE:** Aug 15, 2022 |
| **Course name**<br>Data Mining I (1DL360) | |
| **Your exam code**<br>□□□□ – □□□□ – □□□ | |
| Semester when you were first registered to the course: | Programme: |
| Time for submitting the exam: | Table number: |

## INSTRUCTIONS

Please check that you have the correct exam!

This sheet should always be submitted, even if you haven't solved any of the exam problems.

Each solution should be written on a separate paper.

**Write your exam code on each paper.**

Please use only *one side* of the papers and do not use a pencil with red colour.

Sort the solutions in question order, with question 1 first, before you submit them.

| No. | Solved problems (mark with X) | Points earned | Comments from the teacher |
|---|---|---|---|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |
| 11 | | | |
| 12 | | | |
| 13 | | | |
| 14 | | | |
| 15 | | | |
| 16 | | | |
| 17 | | | |
| 18 | | | |
| 19 | | | |
| 20 | | | |
| 21 | | | |
| 22 | | | |
| 23 | | | |
| 24 | | | |
| 25 | | | |
| 26 | | | |
| 27 | | | |
| 28 | | | |
| 29 | | | |
| 30 | | | |

$\Sigma$

□ Exam with bonus points: Grade is not shown.[1]

5 ≥ □ 4 ≥ □ 3 ≥ □

**Exam grade:**

[1] The final result (points including bonus points and grade) can be found in Ladok after certification.

**Answer sheet:**

0 0 0 0
1 1 1 1
2 2 2 2
3 3 3 3
4 4 4 4
5 5 5 5
6 6 6 6
7 7 7 7
8 8 8 8
9 9 9 9

⟵ please write your exam code in the box below (full code), and also encode it on the left (only the number). For example, if your code is AB-0037-CDE you should fill in 0 in the first column, 0 in the second, 3 in the third and 7 in the fourth.

Full exam code:

......................................

## 1 Data

Question 1: A B C D E F G
Question 2: A B C D E F G
Question 3: A B C D E F G
Question 4: A B C D E F G
Question 5: A B C D E F
Question 6: A B C D E F
Question 7: A B C D E F
Question 8: A B C D E F
Question 9: A B C D E F G H
Question 10: A B C D E F G

## 2 Association rules

Question 11: A B C D E
Question 12: A B C D E F
Question 13: A B C D E F
Question 14: A B C D E F
Question 15: A B C D E F
Question 16: A B C D E F G
Question 17: A B C D E F G
Question 18: A B C D E
Question 19: A B C D E F G
Question 20: A B C D E F G

## 3 Classification

Question 21: A B C D E
Question 22: A B C D E
Question 23: A B C D E
Question 24: A B C D E F
Question 25: A B C D E F
Question 26: A B C D E
Question 27: A B C D E
Question 28: A B C D E F G H
Question 29: A B C D E F G H
Question 30: A B C D E F G H

## 4 Clustering

Question 31: A B C D E F G
Question 32: A B C D E
Question 33: A B C D E F G
Question 34: A B C D E F G
Question 35: A B C D E F G
Question 36: A B C D E F G
Question 37: A B C D E F G
Question 38: A B C D E F G
Question 39: A B C D E F
Question 40: A B C D E F G

Uppsala University
Department of Information Technology
Data Mining I (1DL360) – Aug 15, 2022

**Instructions:** Read through the complete exam and note any unclear directives before you start solving the questions. For each question there can be one or more correct answers, but you can choose only one. If you choose a correct answer, you gain 3 points. A wrong answer does not generate negative points – but the teacher reserves the right to penalize answers that are outrageously wrong. The questions are divided into four sections. To achieve a grade of 3, you must gain at least 60% of the points in each section. To achieve a grade of 4, you must gain at least 75% of the points in the whole exam. To achieve a grade of 5, you must collect at least 90% of the points in the whole exam. You are allowed to use dictionaries to and from English and a calculator, but no other material. Answers must be given exclusively on the answer sheet: answers given on the other sheets will be ignored. To mark an answer fill in the box *completely* (that is, not just crossing it). When the result of a question is numeric, it is possible that your result is slightly different from the proposed answer because of numeric approximation: for example, if the proposed answer is .36 and your computation returns .3563, then .36 must be selected as the correct answer. For distance-based methods, whenever it is not mentioned which distance function to use and the data is numerical please use Euclidean distance.

**Please, submit only the page with the answer sheet, thanks.**

## 1 Data

**Question 1** Consider the following two vectors:

- $v1 = \langle 0, 1, 1, 0, 0, 1 \rangle$
- $v2 = \langle 0, 1, 0, 1, 0, 1 \rangle$

What is the Jaccard coefficient for $v1$ and $v2$?

$\boxed{A}$ 2/6

$\boxed{B}$ 4/6

$\boxed{C}$ 3/6

$\boxed{D}$ 6/6

$\boxed{E}$ 5/6

$\boxed{F}$ 1/6

$\boxed{G}$ None of the previous answers

**Question 2** Consider the following two vectors:

- $v1 = \langle 0, 1, 0, 1, 0, 1\rangle$
- $v2 = \langle 0, 1, 1, 0, 0, 1\rangle$

What is the Simple Matching coefficient for $v1$ and $v2$?

A 3/6<br>
B 6/6<br>
C 1/6<br>
D 5/6<br>
E 4/6<br>
F 2/6<br>
G None of the previous answers

**Question 3** Consider the following two vectors:

- $v1 = \langle 1, 2, 3\rangle$
- $v2 = \langle 1, 4, 6\rangle$

What is the Manhattan distance between $v1$ and $v2$?

A $\sqrt{5}$<br>
B 2<br>
C $\sqrt{13}$<br>
D 4<br>
E 0<br>
F 3<br>
G None of the previous answers

**Question 4** Consider the following four documents (already coded as a list of terms):

1. dog dog elephant gharial pangolin
2. cat cat elephant
3. cat cat turtle
4. cat cat cat cat cat

What is the tf*idf weight for term dog in document 1? (use the definitions seen in the course, with number of occurrences and base-2 logarithm)

A .5<br>
B 0<br>
C .125<br>
D 1<br>
E 4<br>
F .25<br>
G None of the previous answers

**Question 5** Consider the following table:

| ID | a1 | a2 |
|---|---|---|
| o1 | 1 | -1 |
| o2 | 2 | 2 |
| o3 | 3 | 5 |
| o4 | 5 | 8 |

What is the value of attribute a2 for object o2 after min-max rescaling?

A .23

B 2

C 0

D 1

E .25

F None of the previous answers

**Question 6** Consider the following table:

| ID | a1 | a2 |
|---|---|---|
| o1 | 1 | -1 |
| o2 | 2 | 2 |
| o3 | 3 | 5 |
| o4 | 5 | 8 |

What is the value of attribute a1 for object o1 after standardisation? (use the sample standard deviation in the calculation)

A 0

B .25

C 1

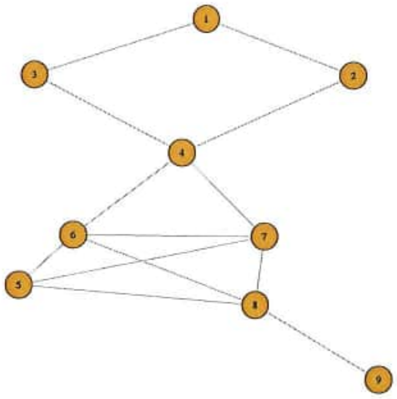
D -2.5

E -1.24

F None of the previous answers

---

Consider the following undirected graph:



**Question 7** What is the probability of 0 in the degree distribution of the graph?

- A 1
- B .67
- C 0
- D .5
- E .33
- F None of the previous answers

**Question 8** What is the betweenness of node 4 in the graph?

- A 0
- B 3
- C .5
- D 1.5
- E 1
- F None of the previous answers

**Question 9** What is the clustering coefficient of the graph, using the definition based on triangles?

- A 1.32
- B .35
- C .64
- D 1.14
- E .58
- F 0
- G .31
- H None of the previous answers

**Question 10** What is $p$ in an ER G(9,p) model whose expected vertex degree equals the average vertex degree of the graph?

- A -.02
- B .36
- C 1
- D .67
- E 1.2
- F .53
- G None of the previous answers

## 2 Association rules

**Question 11** Which of the following statements is true? (s indicates support)

□ A s(ABD) $\geq$ s(ABC)

□ B s(ABCD) $\geq$ s(ABC)

□ C s(ABC) $\geq$ s(ABD)

□ D s(ABCD) $\geq$ s(ABD)

□ E None of the previous answers

**Question 12** Which of the following statements is true?

□ A $\text{lift}(X \to Y) \geq \text{lift}(Y \to X)$

□ B $\text{lift}(X \to Y) = \text{confidence}(X \to Y)/\text{support}(X \to Y)$

□ C $\text{lift}(X \to Y) \leq \text{confidence}(X \to Y)$

□ D $\text{lift}(X \to Y) = \text{confidence}(X \to Y)/\text{support}(Y \to X)$

□ E $\text{lift}(X \to Y) = \text{confidence}(X \to Y)$

□ F None of the previous answers

---

Consider the following transactions:

1. i1, i2, i3
2. i4, i5, i2, i6
3. i5, i1
4. i6, i5, i1, i2
5. i6, i5, i1, i2
6. i6, i1
7. i6, i5
8. i2, i6

**Question 13** What is the support of {i5,i6} → {i2}?

□ A 1

□ B .375

□ C .5

□ D .625

□ E .75

□ F None of the previous answers

**Question 14** What is the confidence of {i1,i5,i6} → {i2}?

□ A .625

□ B .5

□ C 1

□ D .75

□ E .250

□ F None of the previous answers

**Question 15** What is the lift of {i2,i6} → {i5}?

A .75

B 1.2

C .625

D 1

E .5

F None of the previous answers

**Question 16** Which of the following 3-itemsets is in the list of *candidate* 3-itemsets generated by the APRIORI algorithm, if the list of frequent 2-itemsets is {i1, i2}, {i1, i3}, {i2, i4}, {i4, i3}, {i3, i2}?

A {i2, i3, i4}

B I cannot answer using only the information provided in the question

C {i1, i2, i3}, {i2, i3, i4}

D {i1, i2, i3}

E There are no candidate 3-itemsets

F All the subsets of {i1, i2, i3, i4} with three elements

G None of the previous answers

---

Consider the following transactions:

1. A,B,E,F,G
2. B,D,E,H
3. A,C,E,F,H,I
4. B,D,F,I
5. A,B,C,G,I
6. A,D,H
7. B,C,D,G,H,I
8. A,B,C,D,G,I

**Question 17** How many frequent 3-itemsets exist with minimum support .35?

A 1

B 11

C 4

D 0

E 7

F 2

G None of the previous answers

---

Consider the following frequent itemsets, with their support:

1. {B,G} 0.500
2. {C,I} 0.500

3. {B,D} 0.500
4. {B,I} 0.500
5. {B} 0.750
6. {I} 0.625
7. {A} 0.625
8. {D} 0.625
9. {C} 0.500
10. {G} 0.500
11. {H} 0.500

**Question 18** What is the maximum support of a closed frequent itemset?

A .625
B .5
C 0
D .750
E None of the previous answers

**Question 19** How many frequent itemsets are closed?

A 0
B 8
C 9
D 4
E 2
F 6
G None of the previous answers

**Question 20** How many rules with confidence $\geq .75$ exist, with only one item in the right-hand-side and at least one item in the left-hand-side of the rule?

A 5
B 0
C 11
D 2
E 7
F 1
G None of the previous answers

## 3 Classification

**Question 21** Consider the following confusion matrix:

| | True 0 | True 1 |
|---|---|---|
| Predicted 0 | 35 | 20 |
| Predicted 1 | 20 | 30 |

What is the accuracy of the classifier?

A .35

B .62

C .54

D .78

E None of the previous answers

**Question 22** Consider the following confusion matrix:

| | True 0 | True 1 |
|---|---|---|
| Predicted 0 | 35 | 25 |
| Predicted 1 | 10 | 30 |

where 0 is the positive class. What is the precision of the classifier?

A .78

B .68

C .25

D .35

E None of the previous answers

**Question 23** Consider the following confusion matrix:

| | True 0 | True 1 |
|---|---|---|
| Predicted 0 | 35 | 30 |
| Predicted 1 | 20 | 25 |

where 0 is the positive class. What is the recall of the classifier?

A .64

B .78

C .35

D .54

E None of the previous answers

**Question 24** Consider a classifier tested on five records and assigning the following probabilities of the records to belong to the positive class: .95, .92, .87. .43, .21. We indicate the positive and negative class, respectively, as + and -. Assume that the actual class of these records, in the same order, is -, -, -, +, +. What value of TPR has the Roc curve for this classifier when the FPR is $\frac{2}{3}$?

A $\frac{2}{3}$

B 0

C $\frac{1}{3}$

D 1

E .5

F None of the previous answers

**Question 25** Consider a decision tree where a node has been split into two leaves. The first leaf contains 4 records, 2 of class c0 and 2 of class c1. The second leaf contains 4 records, 4 of class c0 and 0 of class c1. What is the classification error of this split?

A .40

B .15

C .30

D .25

E .55

F None of the previous answers

**Question 26** In boosting, records that are wrongly classified in previous rounds

A do not change their probability of being included in the test set.

B always have their weights decreased.

C always have their weights increased.

D do not change their probability of being included in the training set.

E None of the previous answers

**Question 27** Assume to have an ordinal attribute (which is not the class label) in your dataset, and that you decide to transform it into a numerical attribute, preserving the order. How will this affect the construction of a decision tree using the C4.5 algorithm?

A The resulting tree will be the same, but it will take a significantly longer time to produce it because numerical attributes have a larger number of possible splitting values.

B This transformation has no effect on decision trees

C The resulting tree will be the same, but it will be much faster to produce it because with numerical attributes we do not have to check all combinations of values.

D Decision trees are only defined for nominal and numerical attributes, so the transformation is necessary to build the tree.

E None of the previous answers

---

Consider the following training set TRAIN:

| ID | a1 | a2 | a3 | Class |
|---|---|---|---|---|
| r1 | 78 | .5 | 2 | C1 |
| r2 | 70 | .7 | 1 | C1 |
| r3 | 60 | 1.3 | 2 | C1 |
| r4 | 60 | .8 | 2 | C2 |
| r5 | 49 | 1.5 | 2.5 | C2 |
| r6 | 49 | .9 | 2 | C2 |

**Question 28** Assume that you want to build a decision tree from the TRAIN dataset. What are the GINI impurities for all possible binary splits of attribute a2? (of course, do not consider splits generating empty nodes)

A 0

B 0 and .36

C .36 and .75

D 0, .36 and .75

E 0, .18, .25, .36, .4 and .75

F .18, .36 and .75

G .25, .4 and .65

H None of the previous answers

---

Now consider the following training set TEST:

| ID | a1 | a2 | a3 | Class |
|---|---|---|---|---|
| t1 | 70 | .5 | 2 | C1 |
| t2 | 60 | .5 | 2 | C1 |

**Question 29** What is the classification error of a 1-NN classifier trained on TRAIN and tested on TEST?

A 6/6

B 0

C 1/6

D 4/6

E 2/6

F 3/6

G 5/6

H None of the other answers

**Question 30** What is the classification error of a 2-NN classifier with distance-based weighting trained on TRAIN and tested on TEST?

A 5/6

B 0

C 4/6

D 3/6

E 6/6

F 2/6

G 1/6

H None of the other answers

## 4 Clustering

**Question 31** Let p = number of points, c = number of clusters, t = number of iterations, v = average number of distinct values in the attributes, d = number of attributes. What is the time complexity of k-means?

A O(p c v d)

B O(p c t v)

C O(p c)

D O(p c t v d)

E O(p c t)

F O(p t)

G None of the other answers

**Question 32** Among the following features of the data, which one is in general the most problematic for the db-scan algorithm?

A clusters with non-globular shapes

B clusters of different sizes

C clusters with different densities

D regions between clusters without data points

E None of the other answers

**Question 33** What is the minimum number of clusters that can be found by db-scan in a dataset with n records?

A the same as the maximum number of points found within eps from any record in the dataset.

B 1

C eps

D 0

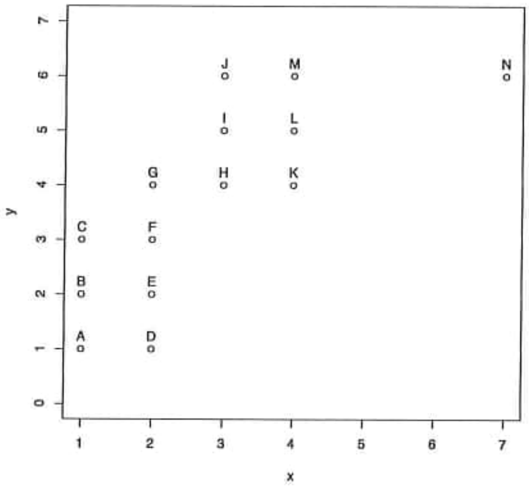
E minPts

F n

G None of the other answers

---

Consider the following data CL:



**Question 34** What clusters are generated by the k-means algorithm applied to the data CL with initial centroids (3,3) and (7,6)?

A Cluster 1: A, B, C, D, E, F, G, H, I, J. Cluster 2: K, L, M, N

B Cluster 1: A, B, C, D, E, F, G, H, I, J, K, L, M. Cluster 2: N

C Both clusters have the same final centroid in (7,6) and the input points can be assigned to any of the two clusters arbitrarily.

D Cluster 1: A, B, C, D, E, F. Cluster 2: G, H, I, J, K, L, M, N

E Cluster 1: A, B, C, D, E, F, G. Cluster 2: H, I, J, K, L, M. Cluster 3: N

F Both clusters have the same final centroid in (7,6) and contain all the points.

G None of the previous answers

**Question 35** What clusters are generated by the complete-link algorithm applied to the data CL, cutting the dendrogram to obtain two clusters?

A Cluster 1: A, B, C, D, E, F, G. Cluster 2: H, I, J, K, L, M, N

B Cluster 1: A, B, C, D, E, F. Cluster 2: G, H, I, J, K, L, M, N

C Cluster 1: A, B, C, D, E, F, G, H, I, J. Cluster 2: K, L, M, N

D More than one result is possible.

E Cluster 1: A, B, C, D, E, F, G, H, I, J, K, L, M. Cluster 2: N

F Cluster 1: A, B, C, D, E, F, G. Cluster 2: H, I, J, K, L, M. Cluster 3: N

G None of the previous answers

**Question 36** What clusters are generated by the single-link algorithm applied to the data CL, cutting the dendrogram to obtain three clusters?

A More than one result is possible.

B Cluster 1: A, B, C, D, E, F, G, H, I, J. Cluster 2: K, L, M, N

C Cluster 1: A, B, C, D, E, F. Cluster 2: G, H, I, J, K, L, M, N

D Cluster 1: A, B, C, D, E, F, G, H, I, J, K, L, M. Cluster 2: N

E Cluster 1: A, B, C, D, E, F, G. Cluster 2: H, I, J, K, L, M, N

F Cluster 1: A, B, C, D, E, F, G. Cluster 2: H, I, J, K, L, M. Cluster 3: N

G None of the previous answers

**Question 37** Apply the db-scan algorithm to the data CL with eps=1 (inclusive) and minPts=3 (not counting the point that is being classified). How many clusters will be identified?

A 10

B 2

C 3

D 8

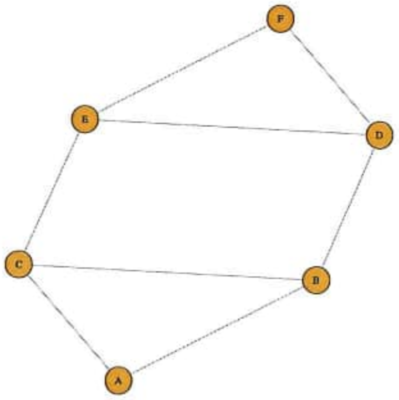
E 1

F 4

G None of the previous answers

---

Consider the following undirected graph:



**Question 38** How many communities would the clique percolation algorithm find in the graph with k=4?

A 3

B 2

C 1

D 4

E 6

F 0

G None of the previous answers

**Question 39** What part of the graph would be removed first by the Girvan-Newman algorithm?

A One of the edges B–D, C–E

B One of the edges B–C, D–E

C One of the edges A–B, A–C, D–F, E–F

D One of the nodes B, C, D, E

E One of the nodes A, F

F None of the previous answers

**Question 40** What is the modularity of the graph when the nodes are assigned to two communities, one with nodes A, B, C, one with nodes D, E, F?

A .43

B .25

C .08

D 1

E -.06

F .70

G None of the previous answers