Example Questions Data Mining, with Answers

Lecturer: dr Arno Knobbe

This example exam is provided for the students' benefit. The number of questions provided here is not an indication of the number of questions of the actual exam (which may be longer or shorter). Also, the nature and topic of the questions may be different in the actual exam. A topic not featuring in these questions is not an indication that it should not be studied for the exam.

General remarks

- The exam is multiple choice, and answers will need to be marked in a separate answer sheet. Be sure to provide your personal details, including your student number (just numbers, no 's').
- For each question, there is exactly one correct answer.
- The sequence of the answers per question has been produced by a random sequence generator, so will not contain any patterns.
- It might be wise to first note your answer in draft on the question form, before copying it to the answer form.
- A calculator is allowed. Mobile phones are not permitted. Please switch off your phone to avoid disturbing your fellow students.
- The grades will be registered within 15 working days in uSis. Due to new privacy regulations, the grades cannot be posted publicly.
- You are allowed to take home the exam questions, should you wish to do so.
- Cheating in any form will have serious consequences.

Question 1

Name a data mining tool that works with a canvas to design you data mining workflow.

(A) Cortana

(B) KNIME

(C) Python

(D) Weka

Question 2

Which of the following statements is correct?

(A) Leave-One-Out is a method aimed at missing values.

(B) Leave-One-Out is a better choice than 10-fold cross-validation when dealing with large data (large n).

(C) Leave-One-Out is simply cross-validation with $k = n$.

(D) When I'm not satisfied with the accuracy using 10-fold cross-validation, I can try Leave-One-Out to get a better score.

Question 3

What does the *Aprioriprinciple* refer to?

(A) The fact that the Apriorialgorithm is one of the most principal ones.

(B) The principle that there is no one algorithm that works best on all datasets.

(C) The fact that a superset of an itemset X has at most the support of X.

(D) The notion that the prior of the target is a good lower bound for the accuracy of a classifier.

Question 4

What is the formula for the entropy of a nominal attribute?

(A) $\sum_i p_i lg(p_i)$

(B) $-\sum_i p_i lg(p_i)$

(C) $-\sum_i lg(p_i)$

(D) $plg(p)$

Question 5

Which of the following statements is correct?

(A) The information gain of an attribute can be any positive number.

(B) The information gain of an attribute can be less than 0.

(C) The information gain of an attribute is at most 1.

(D) The information gain of an attribute is bounded (from above) by the entropy of the target.

Question 6

Which of the following are sets of quality measures for subgroups in Subgroup Discovery? Pick the largest correct set.

(A) WRAcc, z-score, R^2.

(B) WRAcc, z-score, Explained Variance, information gai1.

(C) WRAcc, z-score, joint entropy.

(D) WRAcc, z-score, Explained Variance.

Story A, Frequent Pattern Mining

The following 'story' asks you to complete an itemset lattice with the following labels: *I*= infrequent itemset, *F*=frequent itemset, *M*=maximal frequent itemset, *C*=closed frequent itemset. The upcoming questions test whether you computed the labels correctly. The lattice provided here can be used as a draft, and doesn't need to be handed i1. Only the answers to the multiple choice questions are relevant.

Given a transactional database with the following itemsets over $\{A, \ldots, E\}$, and a minimal support $minsup = 0.3$:

tid	Items
1	$\{A, C, B, E\}$
2	$\{D\}$
3	$\{A, B, E\}$
4	$\{A, C\}$
5	$\{A, D, C\}$
6	$\{C, B, E\}$
7	$\{D, C\}$
8	$\{A, C, B, E\}$
9	$\{B, E\}$
10	$\{D, C\}$

Question 7 (Story A)

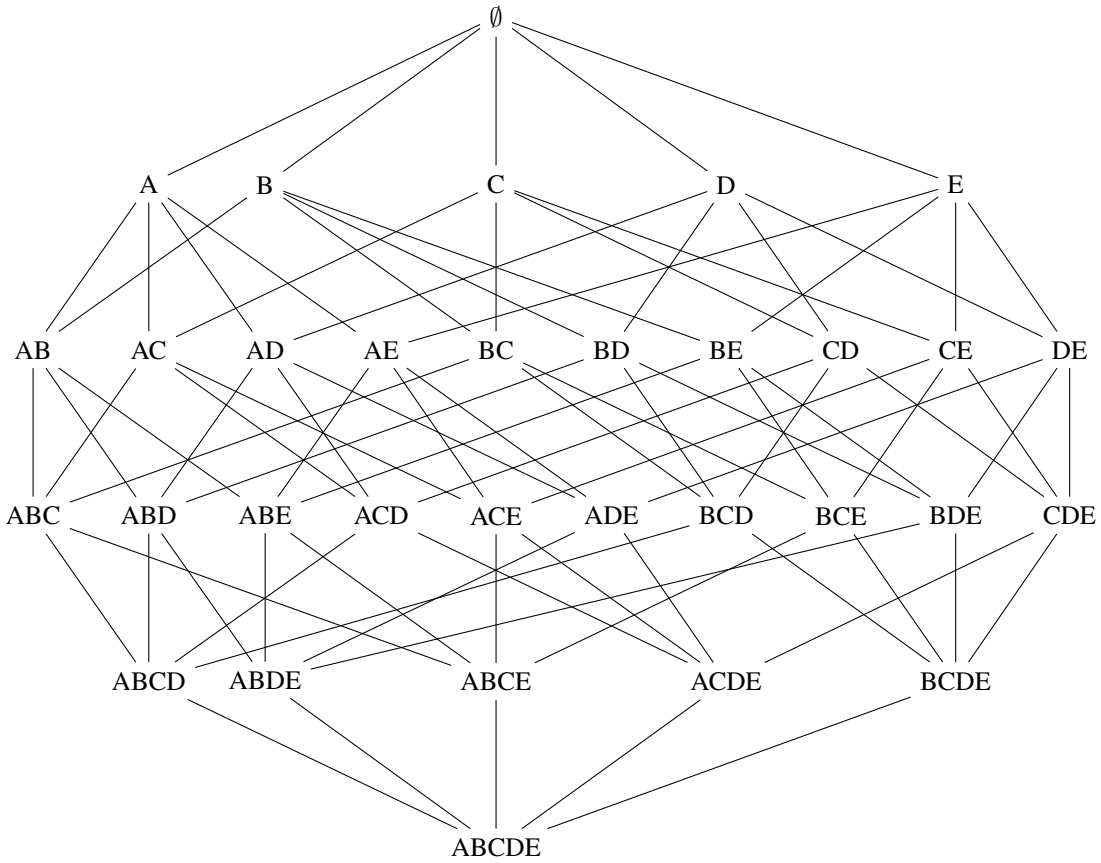
Which itemset in Story A is frequent?

(A) $\{D, E\}$

(B) $\{A, D\}$

(C) $\{A, B\}$

(D) $\{B, D\}$



Question 8 (Story A)

What are the maximal (frequent) itemsets in the dataset of Story A?

(A) $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$, and $\{E\}$

(B) $\{A, C\}$, $\{C, D\}$, and $\{A, B, E\}$.

(C) $\{A, B, E\}$, and $\{B, C, E\}$

(D) $\{A, C\}$, $\{C, D\}$, $\{A, B, E\}$, and $\{B, C, E\}$.

Question 9 (Story A)

Which of the following statements is **not** correct?

(A) $\{A, C\}$, $\{C, D\}$, $\{A, B, E\}$, and $\{B, C, E\}$ are closed itemset.

(B) $\{A, B\}$, $\{B\}$ and $\{E\}$ are closed itemset.

(C) $\{A\}$, $\{C\}$, $\{D\}$, and $\{B, E\}$ are closed itemset.

(D) $\{A, C, D\}$ is a closed itemset.

Question 10

What is the definition of a maximal (frequent) itemset?

(A) An itemset is *maximal frequent* if none of its immediate supersets has the same support.

(B) An itemset is *maximal frequent* if none of its immediate subsets is frequent.

(C) An itemset is *maximal frequent* if none of its immediate supersets is frequent.

(D) An itemset is *maximal frequent* if none of its immediate subsets has the same support.

Question 11

Rank the following in order of entropy, from low to high: 1) a person's gender, 2) whether they own a Ferrari, 3) a person's social security number, 4) a person's highest educatio1.

(A) a person's gender, whether they own a Ferrari, a person's social security number, a person's highest educatio1.

(B) whether they own a Ferrari, a person's gender, a person's highest education, a person's social security number.

(C) a person's gender, a person's highest education, whether they own a Ferrari, a person's social security number.

(D) a person's social security number, a person's highest education, a person's gender, whether they own a Ferrari.

Question 12

What is a disadvantage of using histograms to estimate the density of an attribute?

(A) Bin boundaries can be placed at unfortunate locations, causing empty bins, or too full bins.

(B) It performs worse than Kernel Density Estimatio1.

(C) It assumes a normal distributio1.

(D) It is an unsupervised method.

Question 13

Which of the methods below is an example of an unsupervised learning algorithm?

(A) k-N1.

(B) k-means Clustering.

(C) Subgroup Discovery.

(D) Linear Regressio1.

Question 14

Looking at the descriptions, as well as the feature values in the table below, which data types match which features?

- genre: Contains various names of movie styles.
- rating: Movies are rated on a 5 point scale from very bad to very good.
- gross: Money that the movie made.
- cinema: If the movie was shown in cinemas.

id	genre	rating	gross	cinema
1	horror	very bad	5000	0
2	drama	good	8000	1
3	comedy	very good	9000	1

(A) genre = nominal, rating = ordinal, gross = ordinal, cinema = nominal

(B) genre = nominal, rating = nominal, gross = numeric, cinema = binary

(C) genre = ordinal, rating = nominal, gross = numeric, cinema = binary

(D) genre = nominal, rating = ordinal, gross = numeric, cinema = binary

Question 15

Which of the following statements about clustering is correct?

(A) In k-means the initial assignment of an instance (before the algorithm converges) is dependent on its nearest neighbour.

(B) In k-means clustering, k is learned and reflects the number of clusters.

(C) In k-medoids, the number of observed data points is equal to the number of clusters.

(D) In k-medoids the cluster representative (central point) is always an observed data point whereas in k-means this is not the case.

Question 16

Which description does **not** apply to the parameter k in the k-NN algorithm?

(A) It is the number of classes for the classification problem.

(B) It influences the smoothness of the decision boundary.

(C) It determines how many neighbours are considered for classifying a new example.

(D) It controls how well the model fits the training data.

Question 17

When using k-means on data that has circular properties, what is a possible undesirable outcome?

(A) The algorithm doesn't converge.

(B) The algorithm gets stuck in a local optimum.

(C) Different results on each different run of the algorithm.

(D) Cluster centers move to the circular data.

Question 18

Say you need to distribute 100 balls over 5 boxes. Explain in what situation the entropy of the distribution is the highest.

(A) When all boxes contain the same number of balls.

(B) When a single box contains all balls.

(C) When all boxes contain different numbers of balls.

(D) When the number of balls in each box is $lg(100)$

Question 19

What two criteria are being balanced in a typical SD quality measure for binary classification?

(A) The false positive rate and the false negative rate.

(B) The number of positives within the subgroup, and the size of the subgroup.

(C) The unusualness of the distribution of the target, and the size of the subgroup.

(D) The Weighted Relative Accuracy and the information gai1.

Story B: Maximally Informative k-Itemsets

Consider the following dataset of binary attributes:

A	B	C	D
1	0	0	1
1	0	0	1
1	1	1	0
1	1	1	0
1	1	0	0
0	1	0	0
0	0	1	1
0	0	1	1

For answering the following questions, you may need to consult the following table of entropies.

p	$H(p)$
0	0
1/8	0.54
2/8	0.81
3/8	0.95
4/8	1
5/8	0.95
6/8	0.81
7/8	0.54
1	0

Question 20 (Story B)

What is the entropy of each of the four attributes over the entire dataset of Story B?

(A) $H(A) = 0.625$, $H(B) = 0.5$, $H(C) = 0.5$, $H(D) = 0.5$

(B) $H(A) = 0.95$, $H(B) = 1$, $H(C) = 1$, $H(D) = 1$.

(C) $H(A) = 0.95$, $H(B) = 0$, $H(C) = 0$, $H(D) = 0$.

(D) $H(A) = 0.287$, $H(B) = 1$, $H(C) = 1$, $H(D) = 1$

Question 21 (Story B)

Which itemset(s) is/are a miki with $k = 2$?

(A) $\{B, C\}$.

(B) $\{B, D\}$.

(C) $\{B, C\}$ and $\{C, D\}$.

(D) $\{B, C, D\}$

Question 22 (Story B)

Give the joint entropy of $\{A, B, C, D\}$?

(A) 2

(B) 3.95

(C) 2.25

(D) 3

Answers

1. B
2. C
3. C
4. B
5. A
6. B
7. C
8. D
9. B
10. C
11. B
12. A
13. B
14. D
15. D
16. A
17. C
18. A
19. C
20. B
21. C
22. C