☑ Instructions (2024)

To pass the exam you must obtain at least 60% of the points. For higher grades, you must obtain at least 75% (for a 4) or 90% (for a 5) of the available points. You are allowed to use a calculator, but no other material (notice that the exam does not require to remember formulas, with the exception of simple ones such as calculating the distance between two points). If there are questions requiring to specify a numerical result, 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, if it is not mentioned which distance function to use and the data is numerical please use Euclidean distance.

Please specify your study programme (optional, used to evaluate the course):

□

1 Attribute types (properties)

For which of the following attribute types is a monotonic function a permissible transformation in general?

Choose one answer:

○ Interval

○ Quantitative

○ None of the other answers

○ Ordinal

○ Ratio

○ Nominal

Maximum marks: 1

2 Boosting (theory)

In boosting, records that are wrongly classified in previous rounds:

Choose one answer:

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

○ None of the other answers

○ always have their weights decreased.

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

○ always have their weights increased.

Maximum marks: 1

3 PCA (theory)

What happens in general if PCA is applied to a table where all the attributes are standardized, then one attribute is multiplied by a very large number?

Select one alternative:

○ PCA will identify a component for each outlier

○ PCA will identify a single component

○ PCA will identify a very large number of components

○ The first component will be almost orthogonal to that attribute

○ A single component will explain most of the variability in the data

Maximum marks: 1

4 Classification (theory)

Which of the following methods partitions the dataset into a training and a test set (that is, each record is used only once and is included either in the training or in the test set)?

Choose one answer:

○ Leave one out

○ k-fold validation

○ Boosting

○ Bagging

○ Bootstrap

○ None of the other answers

Maximum marks: 1

5 k-NN (calculation)

Consider the following training set:

a1	a2	a3	class
13	1	19	C1
5	5	1	C1
10	3	9	C1
14	4	13	C2
9	2	20	C2

And the following test set:

a1	a2	a3	class
7	2	10	C1
11	4	6	C1

What is the classification error of a 3-NN classifier with distance-based weighting? (use Manhattan distance)

Select one alternative:

○ .5

○ .33

○ .17

○ 0

○ .66

○ None of the other answers

Maximum marks: 3

6 Complete link (calculation)

Consider the following data:

ID	x	y
A	1	1
B	1	2
C	1	3
D	2	4
E	3	5
F	4	6
G	3	0
H	4	1
I	5	2
J	6	3

Which clusters will be merged first by the complete-link algorithm, using Euclidean distance?
Select one alternative:

○ {A,B,C,D,E,F} and {G,H,I,J}

○ {A} and {J}

○ {A,B,C} and {D,E,F,G}

○ None of the other answers

○ {A} and {B}, or {B} and {C}

○ {G, H} and {I, J}

Maximum marks: 2

7 Apriori (candidates)

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}?

Select one alternative:

○ {i1, i2, i3}

○ There are no candidate 3-itemsets

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

○ {i2, i3, i4}

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

○ None of the other answers

Maximum marks: 2

8 Apriori (calculation)

Consider the following frequent itemsets, with their support:

- {B,G} 0.500
- {C,I} 0.500
- {B,D} 0.500
- {B,I} 0.500
- {B} 0.750
- {I} 0.625
- {A} 0.625
- {D} 0.625
- {C} 0.500
- {G} 0.500
- {H} 0.500

How many rules with confidence ≥ .75 exist, with only one item in the right-hand-side and at least one item in the left-hand-side of the rule?

Select one alternative:

○ 3

○ 0

○ 11

○ 8

○ 1

○ 4

○ None of the other answers

Maximum marks: 3

9 Classification (theory)

Which of the following methods partitions the dataset into a training and a test set (that is, each record is used only once and is included either in the training or in the test set)?

Välj ett alternativ:

○ Holdout

○ Bootstrap

○ Leave one out

○ Cross validation

○ None of the other answers

Maximum marks: 1

10 k-NN (theory)

Which of the following does not have an impact on the complexity of the k-NN algorithm during a classification process?

Choose one answer:

○ The parameter k

○ The number of records in the training data

○ The number of attributes

○ None of the other answers

Maximum marks: 1

11 DB-Scan (theory)

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

Select one alternative:

○ clusters of different sizes

○ clusters with non-globular shapes

○ regions between clusters without data points

○ None of the other answers

○ clusters with different densities

Maximum marks: 1

12 Apriori (hashing)

In the APRIORI algorithm, consider a hashing function with two branches: items 1, 2, 3 are associated to the left branch, and items 4, 5, 6 are associated to the right branch. A hash tree with maximum node capacity 2 is generated to store the following candidate 2-itemsets: (1,2), (1,4), (2,3), (2,5), (2,6). How many itemsets are stored in the left-most non-empty leaf of the hash tree?

Select one alternative:

○ 0

○ 4

○ 1

○ 5

○ 3

○ 2

Maximum marks: 3

13 Attribute types

What is the type of a social class attribute whose values can be “lower”, “middle”, “upper”?

Select one alternative:

○ Nominal

○ Interval

○ Ratio

○ Ordinal

Maximum marks: 1

14 Stemming (theory)

Stemming:

Select one alternative:

○ None of the other answers

○ Tends to increase precision

○ Tends to increase recall

○ Tends to increase dimensionality

○ Tends to increase the number of distinct tokens

Maximum marks: 1

15 DB-scan (theory)

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

Select one alternative:

○ 1

○ 0

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

○ minPts

○ None of the other answers

○ eps

○ n

Maximum marks: 1

16 Single-link (theory)

Single-link:

Välj ett alternativ:

○ Tends to produce clusters of similar sizes.

○ Is equivalent to k-means when we consider k clusters.

○ Can produce overlapping clusters.

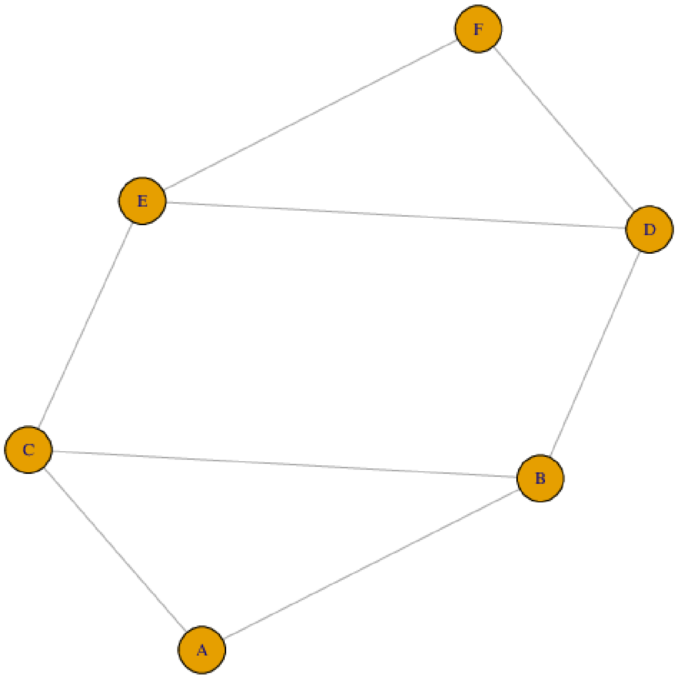
○ None of the other answers.

○ Is a divisive hierarchical algorithm.

Maximum marks: 1

17 Degree centrality (calculation)

Consider the following graph:



What is the highest probability in the degree distribution of the graph?

Choose one answer:

○ .33

○ None of the other answers

○ 0

○ 1

○ .67

○ .5

Maximum marks: 1

18 K-means (theory)

K-means:

Select one alternative:

○ None of the other answers

○ produces k clusters with the maximum possible SSE

○ produces k clusters with the minimum possible SSE

○ has only one possible set of initial centroids leading to the minimum possible SSE

Maximum marks: 1

19 Standardisation (theory)

For which of the following cases is standardisation typically useful to improve the result of the data mining process?

Choose one answer:

○ None of the other answers

○ The presence of correlated attributes

○ The presence of nominal attributes

○ The presence of too few attributes

○ The presence of attributes with different scales

Maximum marks: 1

20 Decision trees (theory)

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?

Select one alternative:

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

○ 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.

○ None of the other answers

○ This transformation has no effect on decision trees

○ 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.

Maximum marks: 1