Politecnico di Torino

1859

Data Science & Machine Learning for engineering applications

AA 2022-2023

DataBase and Data Mining Group

Prof. Tania Cerquitelli

Exam structure & Grading criteria

- The exam includes
  - homework (3 points)
  - a group project (20 points)
  - a written part (10 points)
- The final grade is given by the sum of all three parts.
- The professor may request an integrative test to confirm the obtained evaluation.
- Constraints
  - Grade of the group project is greater than or equal to 12,
  - Grade of the written part is greater than or equal to 6,
  - Group project + written part must be greater than or equal to 18.
  - Homework points will be considered only if the (Group project + written part) >= 18
- If the final score is strictly greater than 31 the registered score will be 30 with honor.

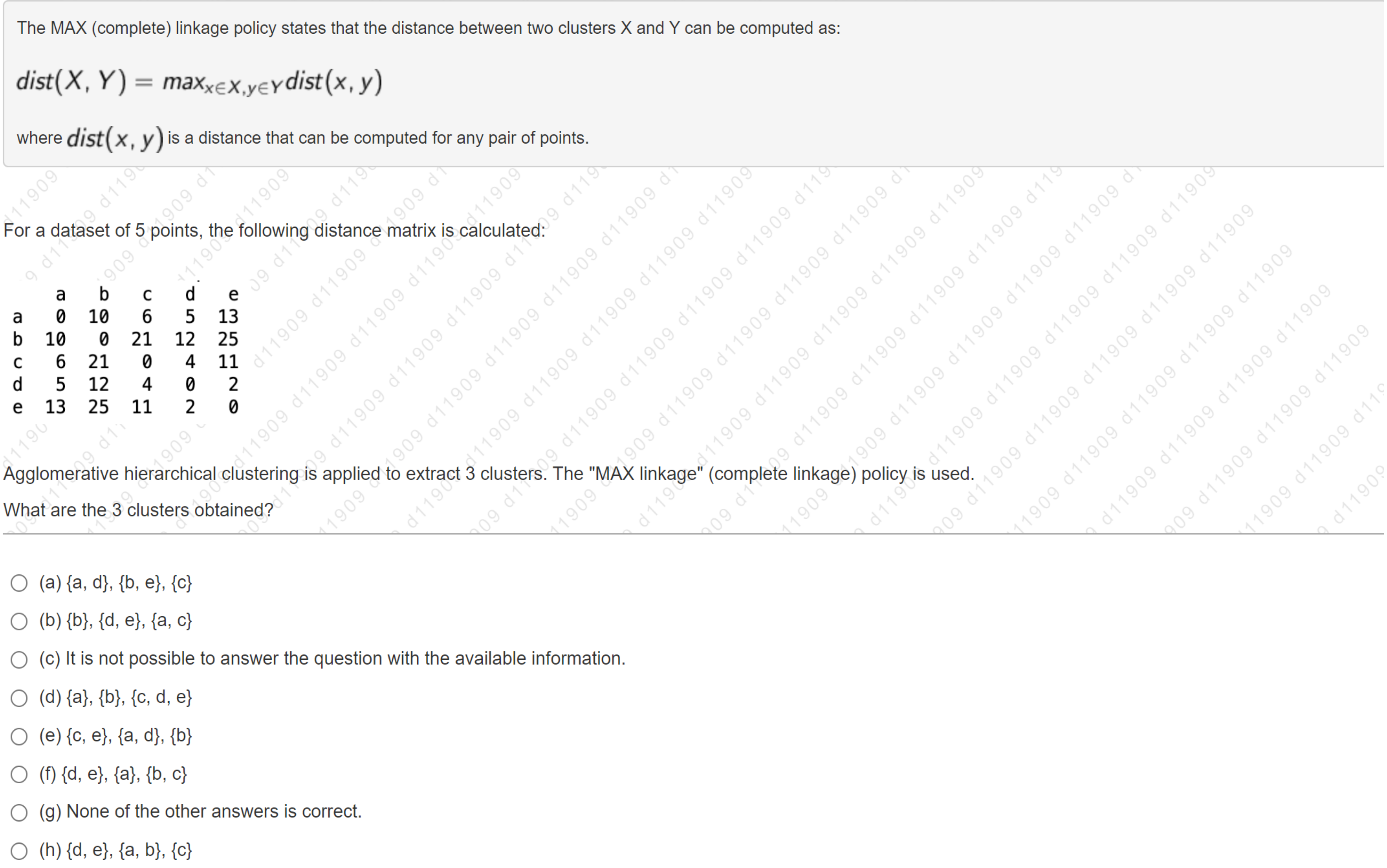
Homework

- During the course, we will assign homework
  - Participants can practice data science and machine learning algorithms in Python and the major data mining and machine learning libraries and become proficient.
- Each homework (hands-on activity) handed in by the deadline will give 0,5/30
  - Overall, 3 points on the final score
- The points for the homework are valid until the exam session in January 2024 (included)

Written part

- It covers the theoretical part of the course.
  - It includes multiple choice and box-to-fill questions related to the theoretical part of the course.
  - For multiple choice questions, wrong answers are penalized.
  - The written exam lasts 60 minutes.
  - Textbooks, notes, electronic devices of any kind are not allowed.

Question 1



The MAX (complete) linkage policy states that the distance between two clusters X and Y can be computed as:

$$dist(X, Y) = max_{x \in X, y \in Y} dist(x, y)$$

where $dist(x, y)$ is a distance that can be computed for any pair of points.

For a dataset of 5 points, the following distance matrix is calculated:

```
    a   b   c   d   e
a   0  10   6   5  13
b  10   0  21  12  25
c   6  21   0   4  11
d   5  12   4   0   2
e  13  25  11   2   0
```

Agglomerative hierarchical clustering is applied to extract 3 clusters. The "MAX linkage" (complete linkage) policy is used.

What are the 3 clusters obtained?

○ (a) {a, d}, {b, e}, {c}

○ (b) {b}, {d, e}, {a, c}

○ (c) It is not possible to answer the question with the available information.

○ (d) {a}, {b}, {c, d, e}

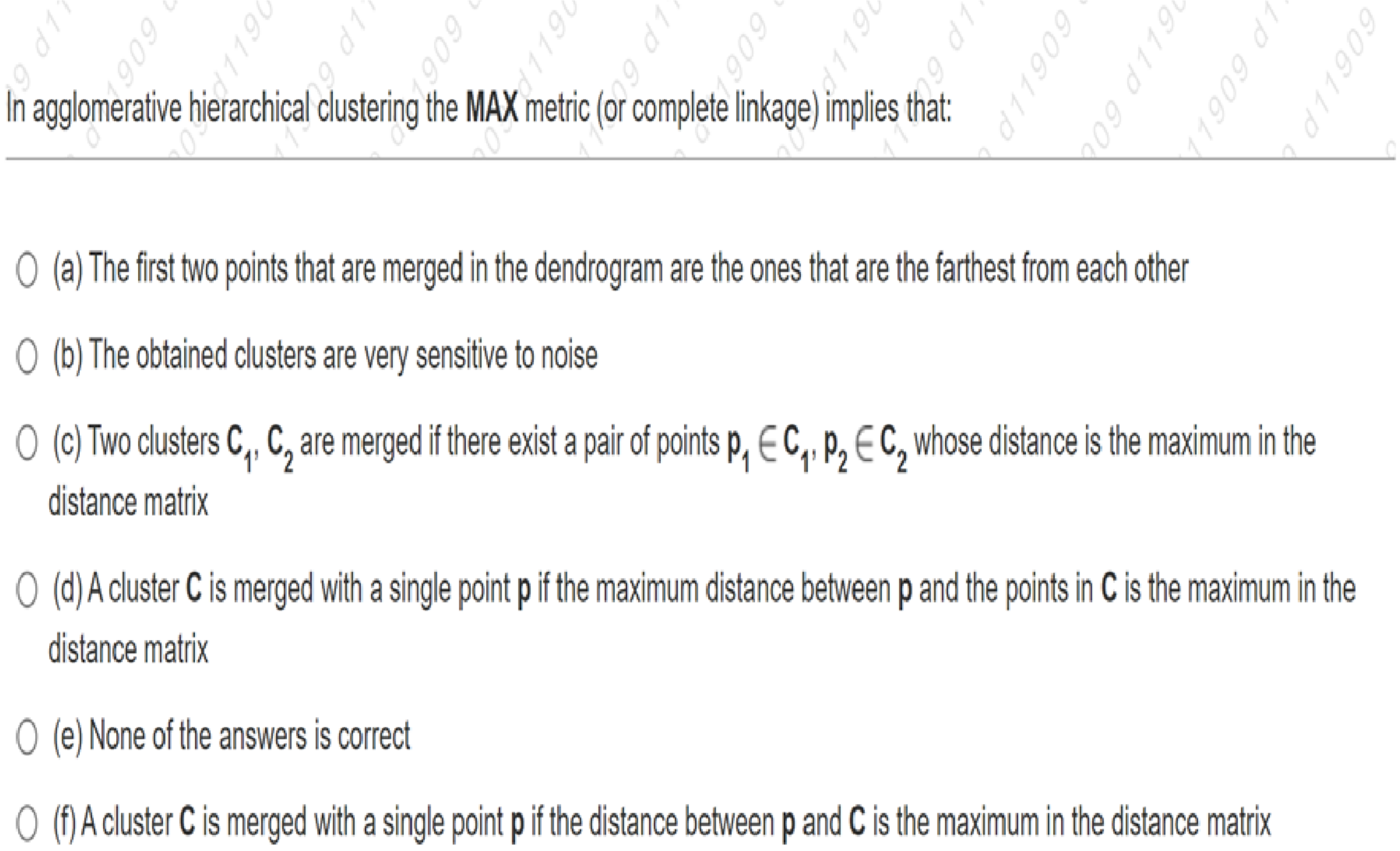
○ (e) {c, e}, {a, d}, {b}

○ (f) {d, e}, {a}, {b, c}

○ (g) None of the other answers is correct.

○ (h) {d, e}, {a, b}, {c}

Question 2



In agglomerative hierarchical clustering the **MAX** metric (or complete linkage) implies that:

○ (a) The first two points that are merged in the dendrogram are the ones that are the farthest from each other

○ (b) The obtained clusters are very sensitive to noise

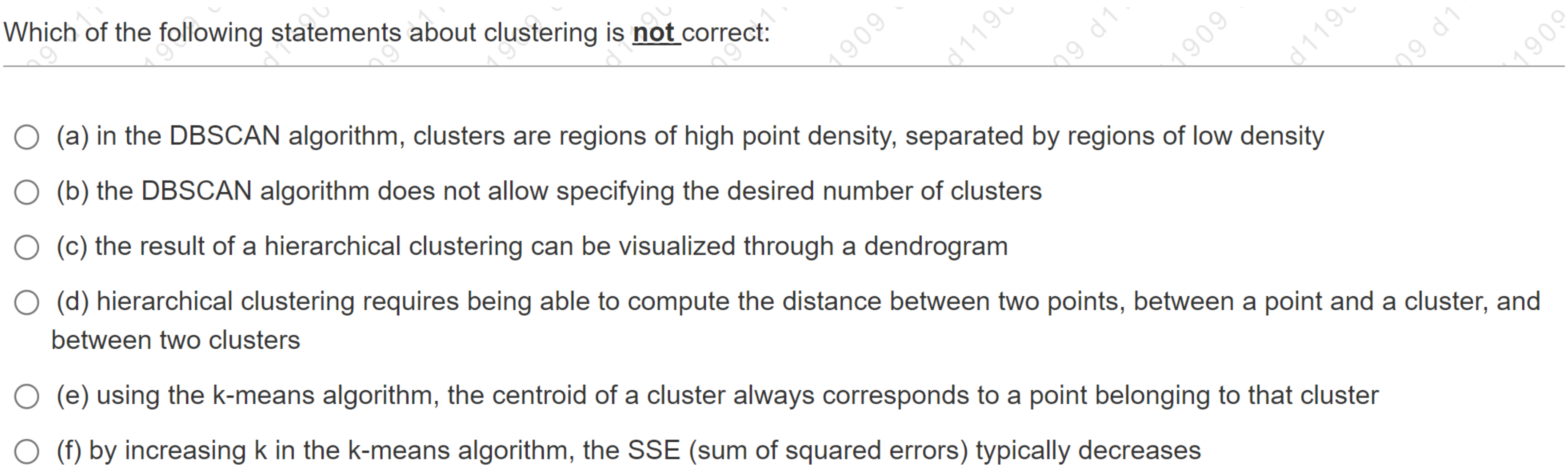
○ (c) Two clusters $\mathbf{C_1}$, $\mathbf{C_2}$ are merged if there exist a pair of points $\mathbf{p_1} \in \mathbf{C_1}$, $\mathbf{p_2} \in \mathbf{C_2}$ whose distance is the maximum in the distance matrix

○ (d) A cluster **C** is merged with a single point **p** if the maximum distance between **p** and the points in **C** is the maximum in the distance matrix

○ (e) None of the answers is correct

○ (f) A cluster **C** is merged with a single point **p** if the distance between **p** and **C** is the maximum in the distance matrix

Question 3



Which of the following statements about clustering is **not** correct:

○ (a) in the DBSCAN algorithm, clusters are regions of high point density, separated by regions of low density

○ (b) the DBSCAN algorithm does not allow specifying the desired number of clusters

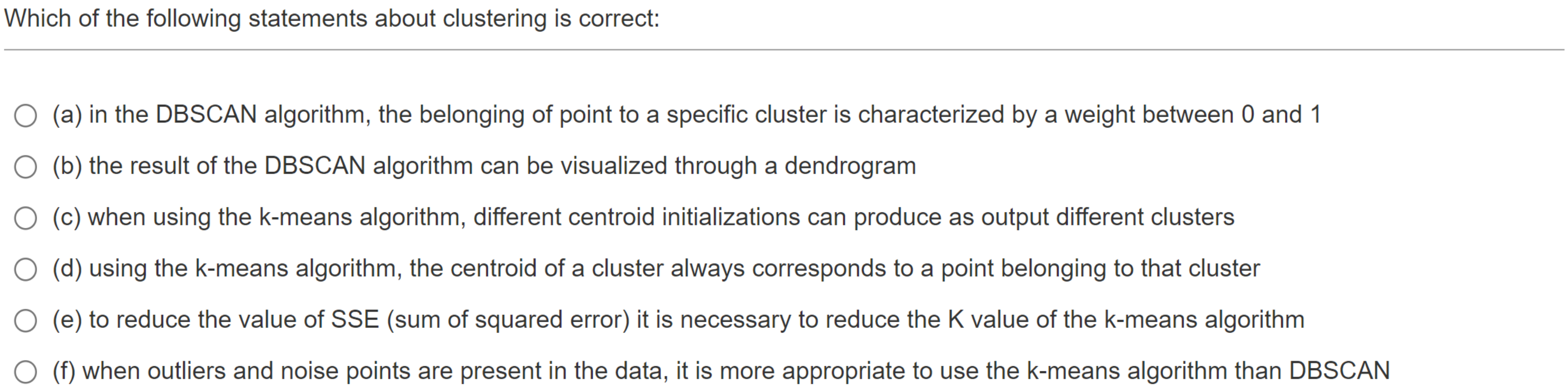
○ (c) the result of a hierarchical clustering can be visualized through a dendrogram

○ (d) hierarchical clustering requires being able to compute the distance between two points, between a point and a cluster, and between two clusters

○ (e) using the k-means algorithm, the centroid of a cluster always corresponds to a point belonging to that cluster

○ (f) by increasing k in the k-means algorithm, the SSE (sum of squared errors) typically decreases

Question 4



Which of the following statements about clustering is correct:

○ (a) in the DBSCAN algorithm, the belonging of point to a specific cluster is characterized by a weight between 0 and 1

○ (b) the result of the DBSCAN algorithm can be visualized through a dendrogram

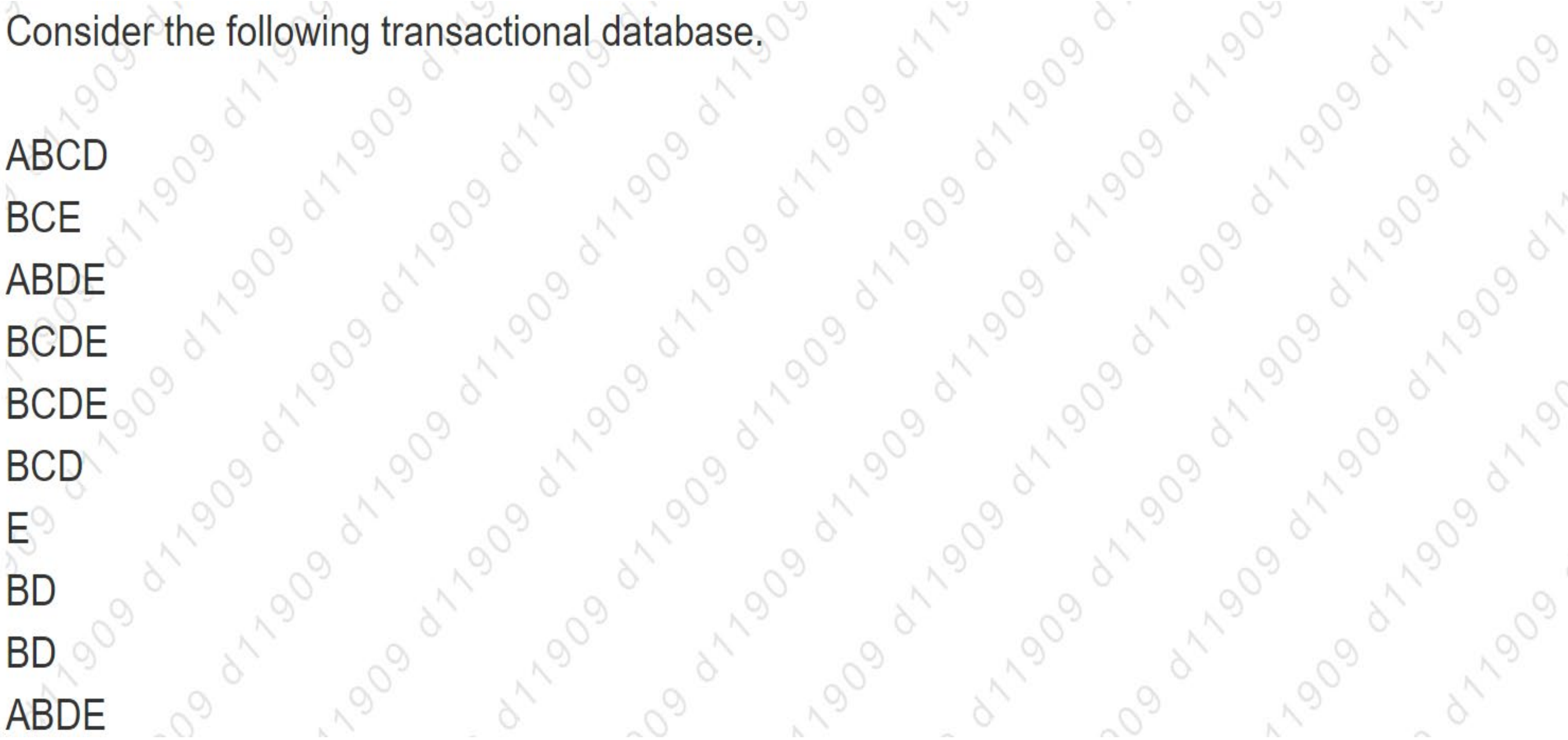
○ (c) when using the k-means algorithm, different centroid initializations can produce as output different clusters

○ (d) using the k-means algorithm, the centroid of a cluster always corresponds to a point belonging to that cluster

○ (e) to reduce the value of SSE (sum of squared error) it is necessary to reduce the K value of the k-means algorithm

○ (f) when outliers and noise points are present in the data, it is more appropriate to use the k-means algorithm than DBSCAN

Question 5

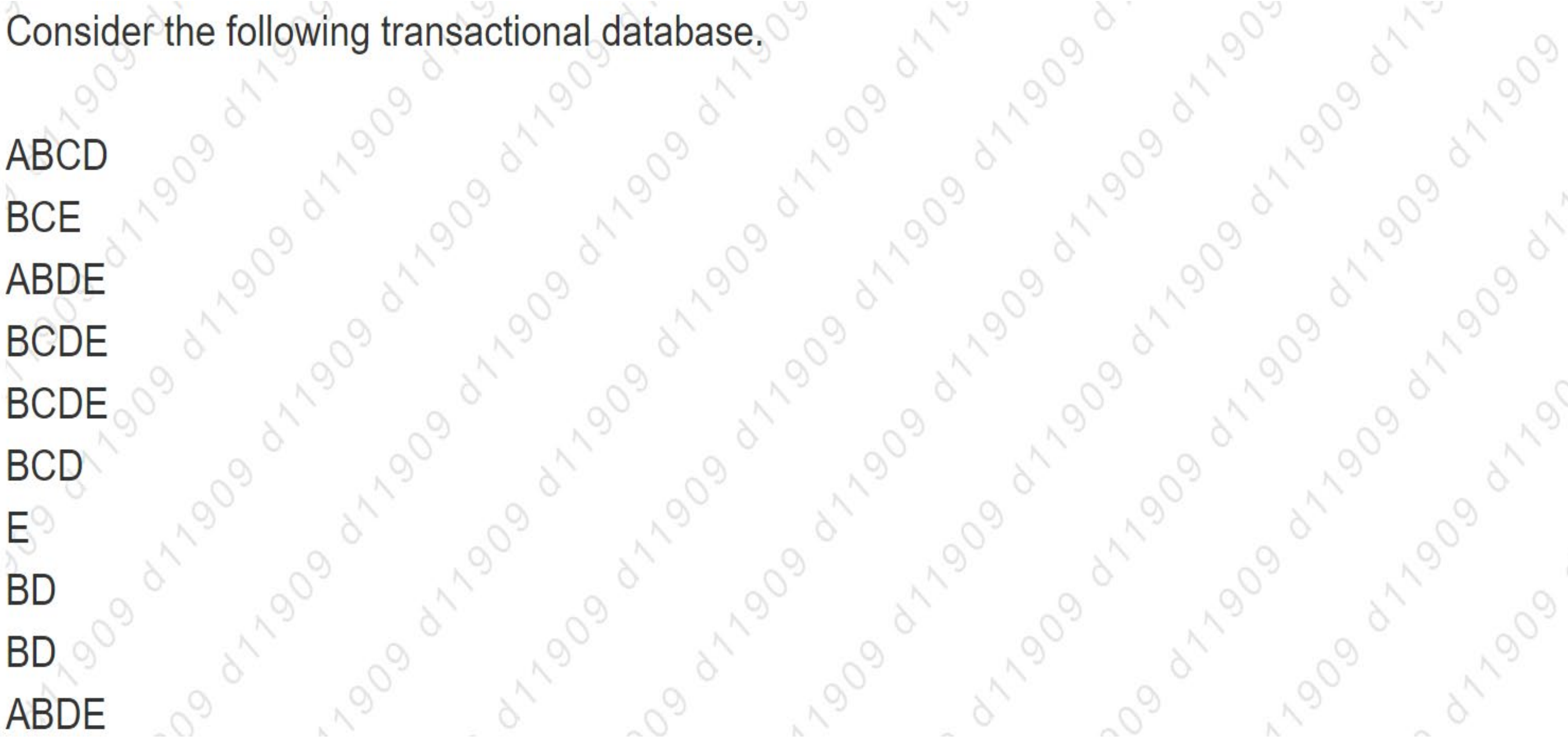


Consider the following transactional database.

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Write the header table of the corresponding FP-Tree with MinSup>2.

Question 6

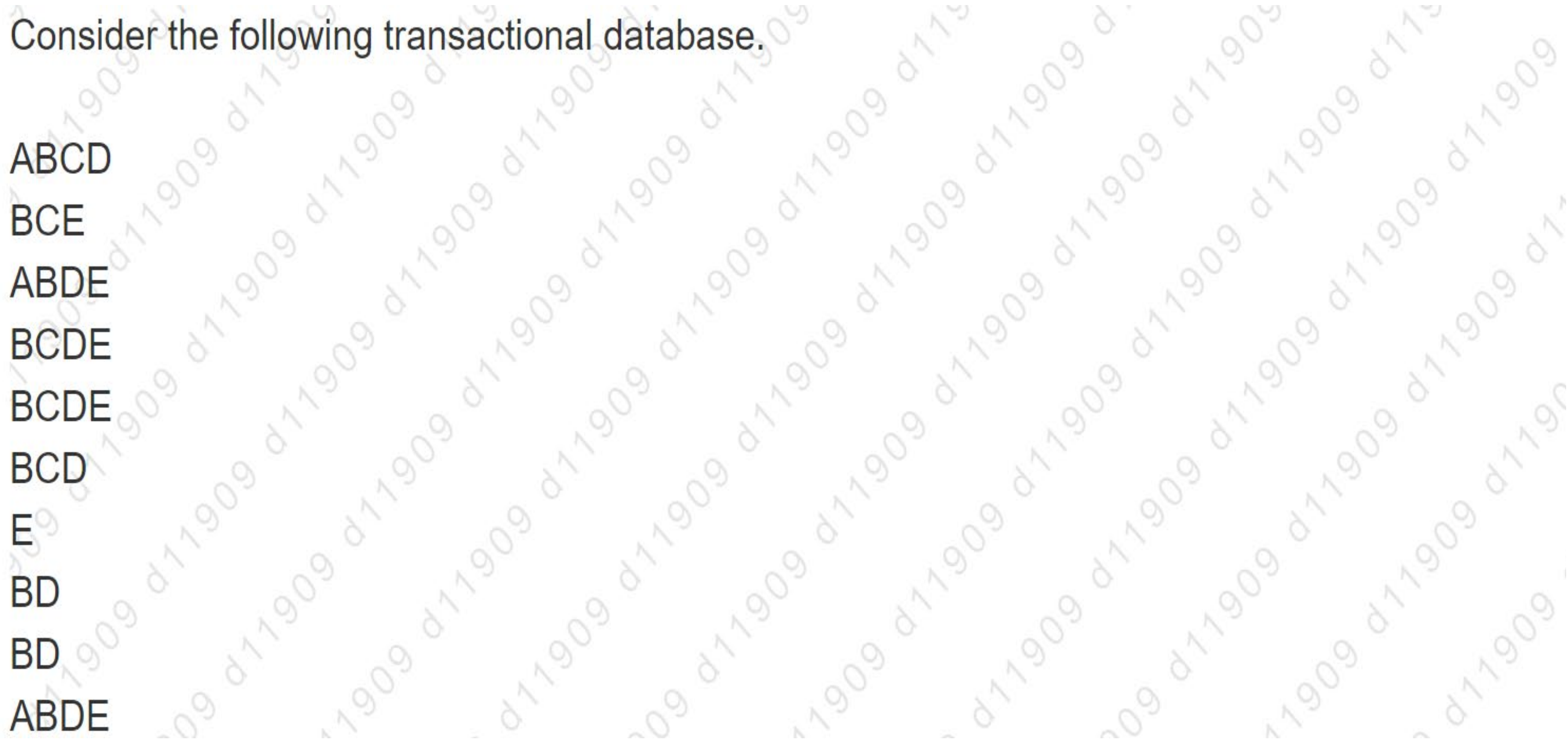


Consider the following transactional database.

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Write the FP-Tree with MinSup>2. Specifically, report the list of paths characterizing the FP-Tree. For each path specify the sequence of nodes in the form (item, local support).

Question 7

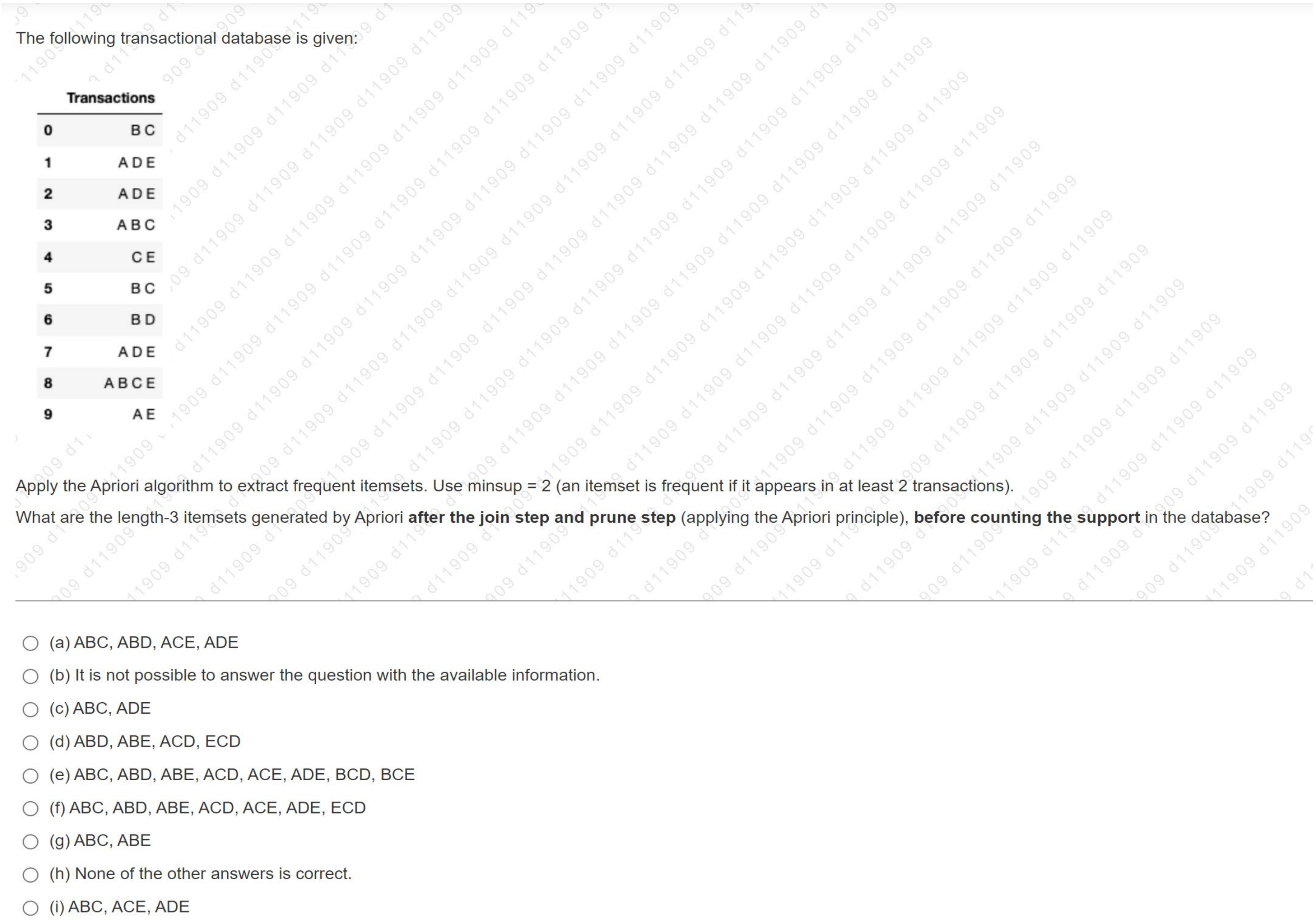


Consider the following transactional database.

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Write the node link chain for item E

Question 8



The following transactional database is given:

	Transactions
0	B C
1	A D E
2	A D E
3	A B C
4	C E
5	B C
6	B D
7	A D E
8	A B C E
9	A E

Apply the Apriori algorithm to extract frequent itemsets. Use minsup = 2 (an itemset is frequent if it appears in at least 2 transactions).

What are the length-3 itemsets generated by Apriori **after the join step and prune step** (applying the Apriori principle), **before counting the support** in the database?

○ (a) ABC, ABD, ACE, ADE

○ (b) It is not possible to answer the question with the available information.

○ (c) ABC, ADE

○ (d) ABD, ABE, ACD, ECD

○ (e) ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE

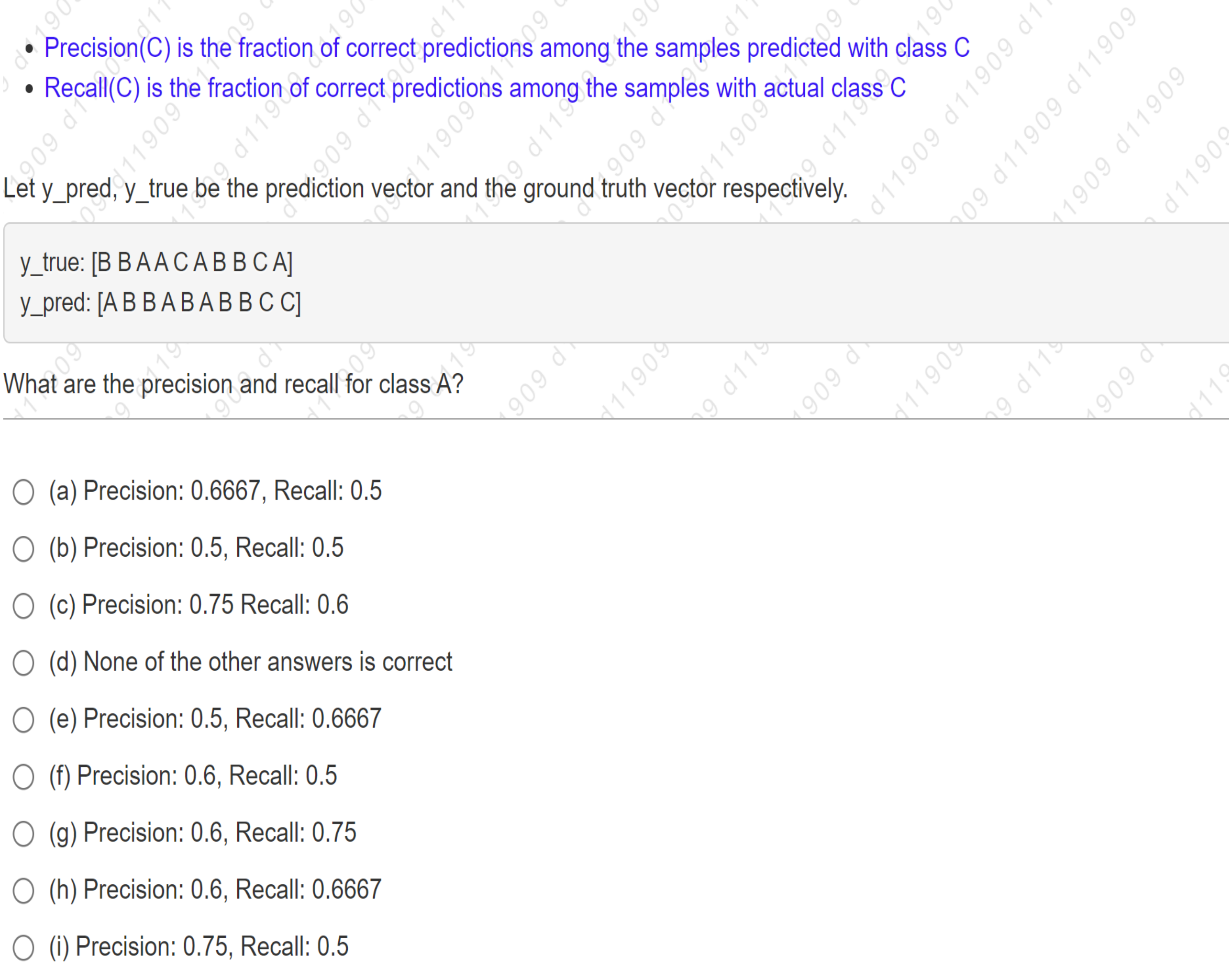
○ (f) ABC, ABD, ABE, ACD, ACE, ADE, ECD

○ (g) ABC, ABE

○ (h) None of the other answers is correct.

○ (i) ABC, ACE, ADE

Question 9



- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C

Let y_pred, y_true be the prediction vector and the ground truth vector respectively.

```
y_true: [B B A A C A B B C A]
y_pred: [A B B A B A B B C C]
```

What are the precision and recall for class A?

○ (a) Precision: 0.6667, Recall: 0.5

○ (b) Precision: 0.5, Recall: 0.5

○ (c) Precision: 0.75 Recall: 0.6

○ (d) None of the other answers is correct

○ (e) Precision: 0.5, Recall: 0.6667

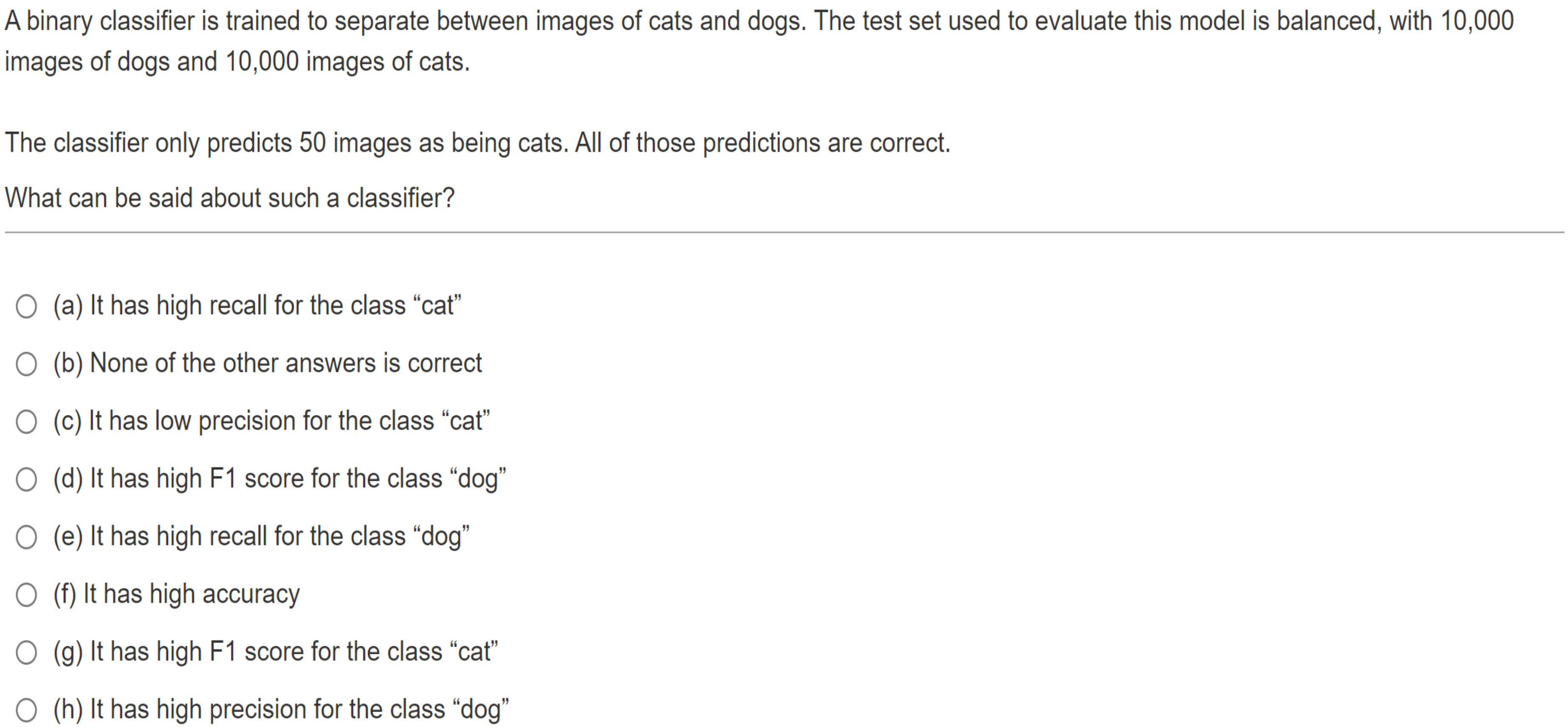
○ (f) Precision: 0.6, Recall: 0.5

○ (g) Precision: 0.6, Recall: 0.75

○ (h) Precision: 0.6, Recall: 0.6667

○ (i) Precision: 0.75, Recall: 0.5

Question 10



A binary classifier is trained to separate between images of cats and dogs. The test set used to evaluate this model is balanced, with 10,000 images of dogs and 10,000 images of cats.

The classifier only predicts 50 images as being cats. All of those predictions are correct.

What can be said about such a classifier?

○ (a) It has high recall for the class “cat”

○ (b) None of the other answers is correct

○ (c) It has low precision for the class “cat”

○ (d) It has high F1 score for the class “dog”

○ (e) It has high recall for the class “dog”

○ (f) It has high accuracy

○ (g) It has high F1 score for the class “cat”

○ (h) It has high precision for the class “dog”