**1. Question 1**

Given the items and the set of transactions :

|  |  |
| --- | --- |
| TransID | Items |
| 1 | A B C E G H I |
| 2 | A B D E F H I |
| 3 | A B D E H |
| 4 | A B E F H |
| 5 | A B E H |
| 6 | A D F G I |
| 7 | A F I |
| 8 | B C D E G I |
| 9 | C G I |
| 10 | D E F G H I |
| 11 | D G I |
| 12 | F |

For the minimum support of 3 , we already determined the frequent 3 itemsets with the APRIORI algorithm:

Which of the following 4-itemsets are preliminary candidates in the next step of APRIORI (i.e., after the merging step but before pruning)?

1. BEHI
2. CDGI
3. CEGI
4. DEFI
5. DEGI
6. DEHI 9.

**2. Solution:**

Generate candidate set using . Condition of joining two itemsets is that they should have elements in common, which is 2 in this case.

Candidate set would be

Thus itemsets are preliminary candidates in step 4 .

**3. Question 2**

For some transaction database we found that the rule has a confidence below the confidence threshold.

Which of the following rules will therefore have a confidence below the confidence threshold as well?

**4. Solution:**

Recall when we have an association rule .

Support is the amount of times and appear together.

Confidence is the support, divided by the times that the first part of the expression appears alone in the table, i.e. .

Support would be the same.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, the in confidence would be the same, but the amount that it would be divided by would be equal or more. So in conclusion it would be below the confidence threshold. 2.

Support would be more than or equal, since the total amount of items is smaller.

The amount of times is in the table is the same.

Thus, in order to calculate the confidence, would be more than or equal, and the bottom would be the same. So it would not be guaranteed to be below the confidence threshold.

Support would be more than or equal, since the total amount of items is smaller.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, in order to calculate the confidence, would be more than or equal, and the bottom also be more. So it would not be guaranteed to be below the confidence threshold.

Support would be the same.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, the in confidence would be the same, but the amount that it would be divided by would be equal or more. So in conclusion it would be below the confidence threshold.

It would not be guaranteed to be below the confidence threshold.

It would not be guaranteed to be below the confidence threshold.

Support would be the same.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, the in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold. 8.

It would not be guaranteed to be below the confidence threshold.

Support would be the same.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, the in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold.

Support would be the same.

The amount of times would appear in the table would be more than, or equal to, the amount of times appear in the table.

Thus, the in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold.

**5. Question 3**

We have the following one-dimensional dataset:

|  |  |
| --- | --- |
| ID | Value |
|  | 1 |
|  | 3 |
|  | 5 |
|  | 7 |
|  | 10 |
|  | 11 |
|  | 12 |

In three attempts, k-means delivered the following three clustering solutions:

We want to compare the solutions using . Which of the following statements are correct?

1. is better than in terms of
2. is better than in terms of .
3. and are equally good in terms of .
4. is better than in terms of .

**6. Solution:**

is a measure of compactness for a cluster. Recall how is calculated for a cluster

We want to calculate the centroids for the three clusterings, and then their respective values.

The correct statements are statement 2 and 3 .

**7. Question 4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | forecast | humidity | wind | play tennis? |
| 1 | sunny | high | weak | no |
| 2 | sunny | high | strong | no |
| 3 | sunny | high | weak | yes |
| 4 | sunny | normal | weak | yes |
| 5 | sunny | normal | strong | no |
| 6 | rainy | high | weak | no |
| 7 | rainy | normal | weak | yes |
| 8 | rainy | normal | weak | yes |
| 9 | rainy | normal | strong | yes |
| 10 | rainy | high | strong | no |

A decision tree is being trained on the above data set. As root of the tree, the attribute "forecast" was already selected.

Which attributes are selected as test nodes at the next level based on the Gini index?

1. For the branch of forecast sunny, we test wind.
2. For the branch of forecast=sunny, we test humidity.
3. For the branch of forecast=rainy, we test wind.
4. For the branch of forecast rainy, we test humidity.

**8. Solution:**

We choose the attribute and the split that minimizes the Gini index.

For forecast sunny:

no" and yes"

* humidity - high: persons
* normal: persons 4,5

Thus, we can calculate the Gini index

* wind
* weak: persons

strong: persons 2,5

Thus, we can calculate the Gini index

* Since (wind) (humidity), we choose to split on wind for forecast sunny.

For forecast =rainy:

no" and yes".

* (humidity)
* high: persons 6,10
* normal: persons

Thus, we can calculate the Gini index

wind

* weak: persons

strong: persons 9,10

Thus, we can calculate the Gini index

* Since (humidity) (wind), we choose to split on humidity for forecast = rainy.

Thus we can conclude that the correct statements are 1 and 4 .

**9. Question 5**

In a dataset with ten points and are labeled outliers.

Four outlier detection methods, , deliver the following rankings (from left-to-right: top-rank to bottom-rank):

|  |  |
| --- | --- |
| method | ranking |
|  |  |
|  |  |
|  |  |
|  |  |

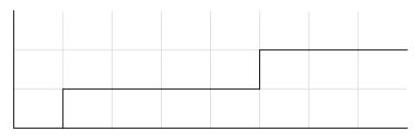
Based on ROC AUC as evaluation measure, which of the following statements is correct?

1. and perform equally well.
2. is better than .
3. is better than
4. is better than

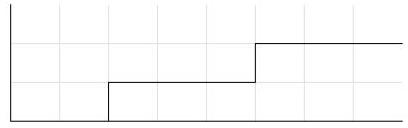
**10. Solution:**

We can create the ROC AUC for each detection method.

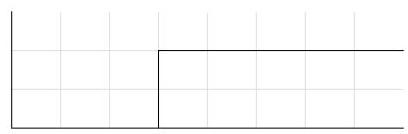
For For



For



For



Thus we can see that the correct statements are statements 1 and 2 .