Mansoura University
Faculty of Computer & Information
Information System Department

# Data Mining

## Workbook

**4th year**

**IS, IT, SWE, Bio**

Dr. Amira Rezk

2019

## Part1: Pre-Processing

- Suppose that the data for analysis includes the attribute *age*. The *age* values are

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) Find the *mean*, *median*, and *mode* of the data.

b) Give the *five-number summary* of the data. And Show a *boxplot* of the data.

c) Partition the data into three bins by each of equal-frequency and equal-width partitioning

d) Use smoothing by bin boundaries to smooth these data

e) Use min-max normalization to transform the value 30 for *age* onto the range [0:0; 1:0].

f) Use z-score normalization to transform the value 30 for *age*, where the *SD* of *age* is 12.94 years.

g) Plot an equal-width histogram of width 10.

- Why do you need to perform the pre-processing operations before perform the Data mining techniques?

- Discuss the different pre-processing operations and declare how and why you use each one.

## Part2: Mining Frequent Patterns, Associations, & Correlations

- Suppose we have market basket data, consisting of 100 transactions and 20 items. If the support for item A is 25%, the support for item B is 90% and the support for itemset {A, B} is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule {A} → {B}. Is the rule interesting according to the confidence measure?

(b) Compute the interest measure for the association pattern {A, B}. Describe the nature of the relationship between item A and item B in terms of the interest measure.

(c) What conclusions can you draw from the results of parts (a) and (b)?

- Consider the data set shown in Table 1, Let min sup = 40% and min conf. = 80%.

a. Find all frequent itemsets using Apriori Algorithm

b. List all the strong association rules.

c. Find the correlation the strong association rules using ***lift***, what is the meaning of the computed value?

Table 1: Market basket transactions

| Transaction ID | Items bought |
|---|---|
| 1001 | {i1, i4, i5} |
| 1024 | {i1, i2, i3, i5} |
| 1012 | {i1, i2, i4, i5} |
| 1031 | {i1, i3, i4, i5} |
| 1015 | {i2, i3, i5} |
| 1022 | {i2, i4, i5} |
| 1029 | {i3, i4} |
| 1040 | {i1, i2, i3} |
| 1033 | {i1, i4, i5} |
| 1038 | {i1, i2, i5} |

*The transaction data shown in the Table 2 from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9). There are 5 meal items that are involved in the transactions. For simplicity the meal items short names (M1 – M5).* ***The minimum support is 2/9 (.222) and the minimum confidence is 7/9 (.777).***

| Table2 | |
|---|---|
| Meal Item | List of Item IDs |
| Order:1 | M1, M2, M5 |
| Order:2 | M2,M4 |
| Order:3 | M2,M3 |
| Order:4 | M1,M2.M4 |
| Order:5 | M1,M3 |
| Order:6 | M2,M3 |
| Order:7 | M1,M3 |
| Order:8 | M1,M2,M3,M5 |
| Order:9 | M1,M2,M3 |

**a. Apply the Apriori algorithm to the dataset of transactions and identify all frequent k-itemsets.** *Show all of your work. You must show candidates but can cross them off to show the ones that pass the minimum support threshold. Note: if a candidate itemset is pruned because it violates the Apriori property, you must indicate that it fails for this reason and not just because it does not achieve the necessary support count (i.e., in these cases there is no need to actually compute the support count). So, explicitly tag the itemsets that are pruned due to violation of the Apriori property. (If you do not know what the Apriori property is, do not panic. You will ultimately get the exact same answer but will just lose a few points).*

**b. Find all strong association rules of the form: X $\wedge$ Y → Z and note their confidence values.**

- Consider the data set shown in Table3

| Table 3:Transactions . | |
|---|---|
| **TID** | **Items bought** |
| 1001 | {i1, i4, i5} |
| 1024 | {i1, i2, i3, i5} |
| 1012 | {i1, i2, i4, i5} |
| 1031 | {i1, i3, i4, i5} |
| 1015 | {i2, i3, i5} |
| 1022 | {i2, i4, i5} |
| 1029 | {i1, i3, i4} |
| 1040 | {i1, i2, i3} |
| 1033 | {i1, i4, i5} |
| 1038 | {i1, i2, i5} |

Let *min sup* = 30% and *min conf.* = 75%.

a. Construct the FP-tree for these transaction

b. Compute the support for item-sets : {i1}, {i4}, {i5}, {i1;i4},{i1; i5},{i4; i5} and {i1; i4; i5}

c. Compute the confidence for the association rules:
{i1; i4} → {i5}; {i1, i5}→ {i4} and {i4; i5} → {i1}.
Which one is a strong rule?

d. Compute the interest measure for the strong association rules in (c). What is the meaning of the computed value

- Consider the data set shown in Table 4

a. Compute the support for item-sets

{i5}, {i2; i4}, and {i2; i4; i5} by treating each transaction ID as a market basket.

b. Use the results in (a) to compute the confidence for the association rules

{i2; i4} → {i5} and {i5} → {i2; i4}.

c. Is confidence a symmetric measure?

d. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

e. Use the results in part (d) to compute the confidence for the association rules {i2; i4} → {i5} and {i5} → {i2; i4}.

f. Discuss whether there are any relationships between support and confidence of parts {a, b} and {d, e}.

g. Compute the lift for the association rules {i2; i4} → {i5} and {i5} → {i2; i4} in parts {b, e}. what is the meaning of the computed value.

Table4: Market basket transactions.

| **Customer ID** | **Transaction ID** | **Items bought** |
|---|---|---|
| 1 | 1001 | {i1, i4, i5} |
| 1 | 1024 | {i1, i2, i3, i5} |
| 2 | 1012 | {i1, i2, i4, i5} |
| 2 | 1031 | {i1, i3, i4, i5} |
| 3 | 1015 | {i2, i3, i5} |
| 3 | 1022 | {i2, i4, i5} |
| 4 | 1029 | {i3, i4} |
| 4 | 1040 | {i1, i2, i3} |
| 5 | 1033 | {i1, i4, i5} |
| 5 | 1038 | {i1, i2, i5} |

- **The following table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, $\overline{\text{hot dogs}}$ refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and $\overline{\text{hamburgers}}$ refers to the transactions that do not contain hamburgers.**

| | *hot dogs* | $\overline{\textit{hot dogs}}$ | $\Sigma_{row}$ |
|---|---|---|---|
| *hamburgers* | 2000 | 500 | 2500 |
| $\overline{\textit{hamburgers}}$ | 1000 | 1500 | 2500 |
| $\Sigma_{col}$ | 3000 | 2000 | 5000 |

a. Suppose that the association rule "*hot dogs* → *hamburgers*" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

*b.* Based on the given data, is the purchase of *hot dogs* independent of the purchase of *hamburgers*? If not, what kind of *correlation* relationship exists between the two? *Note: the* $\chi^2$ *value needed to reject the hypothesis is 10.828*

- A database has four transactions. Let min sup = 60% and min conf = 80%.

| *cust_ID* | *TID* | *items_bought* (in the form of *brand-item_category*) |
|---|---|---|
| 01 | T100 | {King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread} |
| 02 | T200 | {Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread} |
| 01 | T300 | {Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie} |
| 03 | T400 | {Wonder-Bread, Sunset-Milk, Dairyland-Cheese} |

(a) At the granularity of *item_category* (e.g., $item_i$ could be "*Milk*"), for the rule template,

$$\forall X \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c],$$

list the frequent $k$-itemset for the largest $k$, and *all* the *strong* association rules (with their support $s$ and confidence $c$) containing the frequent $k$-itemset for the largest $k$.

(b) At the granularity of *brand-item_category* (e.g., $item_i$ could be "*Sunset-Milk*"), for the rule template,

$$\forall X \in customer,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3),$$

list the frequent $k$-itemset for the largest $k$ (but do not print any rules).

## Part 3: Classification

- Consider the sample data shown in Table1,for a binary classification problem.

(a) What is the entropy of this collection of data with respect to the positive class?
(b) What are the information gains of a1 and a2 relative to this data?
(c) What is the best split (among a1 and a2) according to the information gain?

Table1

| Instance | A1 | A2 | A3 | class |
|---|---|---|---|---|
| 1 | T | T | 1.0 | + |
| 2 | T | T | 6.0 | + |
| 3 | T | F | 5.0 | - |
| 4 | F | F | 4.0 | + |
| 5 | F | T | 7.0 | - |
| 6 | F | T | 3.0 | - |
| 7 | F | F | 8.0 | - |
| 8 | T | F | 7.0 | + |
| 9 | F | T | 5.0 | - |

- Consider the following decision tree

a. Extract IF-THEN rules from this decision tree

b. Evaluate this classification model using a test data set in table2;

**Calculate the following measurements**: accuracy, error rate, sensitivity, specificity, and precision

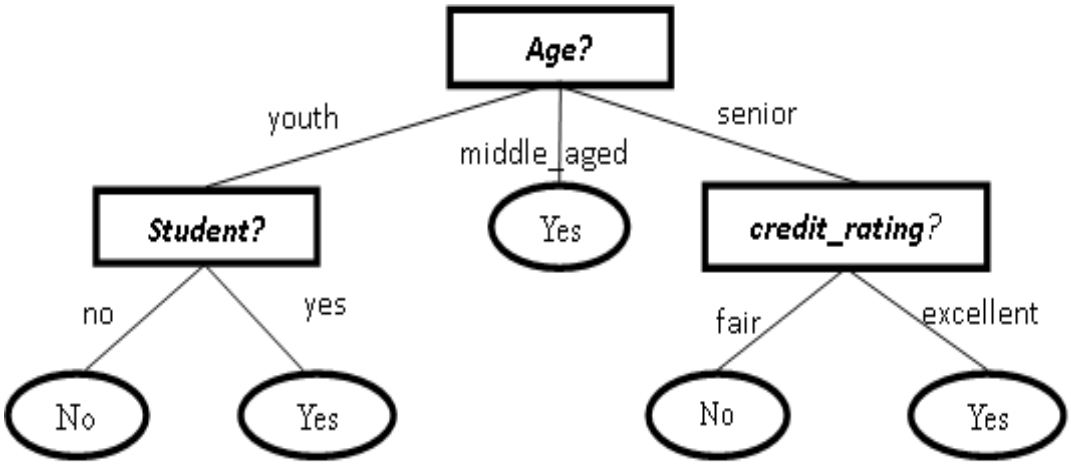


**Table2: test data set**

| RID | age | income | student | Credit_rating | Class: buys_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | yes |
| 2 | youth | high | no | excellent | no |
| 3 | youth | medium | no | fair | no |
| 4 | middle aged | high | no | fair | yes |
| 5 | middle aged | low | yes | excellent | yes |
| 6 | middle aged | high | yes | fair | no |
| 7 | senior | low | yes | excellent | no |
| 8 | senior | low | yes | fair | yes |
| 9 | senior | medium | no | fair | no |
| 10 | senior | medium | yes | fair | yes |

- Table3 presents a training set D of class-labeled tuples randomly selected from customer database. The two decision classes are buy from shop (YES) or does not buy from shop (NO).

a. Generate a decision tree from the training tuples of data partition, D

b. Extract IF_THEN rules from a decision tree. How to assess the goodness of a rule?

c. Use Naïve Bayesian Classification method to classify an individual who has the following attributes:

***{Car ownership="N0"; Marital status= "married", taxable income= medium}***

**Table3: data set**

| Taxable income | Car ownership | Marital status | Buy from shop |
|---|---|---|---|
| high | Yes | Single | No |
| medium | No | Married | No |
| low | No | Single | Yes |
| medium | Yes | Married | No |
| medium | No | Divorced | Yes |
| low | No | Married | No |
| high | Yes | Divorced | No |
| low | No | Single | Yes |
| low | No | Married | No |
| low | No | Single | Yes |

- Consider the sample data shown in Table4

Table4

| A1 | A2 | A3 | class |
|---|---|---|---|
| T | T | P | + |
| T | T | P | + |
| T | F | N | - |
| F | F | N | + |
| F | T | N | - |
| F | T | N | - |
| F | F | P | - |
| T | F | P | + |
| F | T | N | - |
| T | F | N | - |

a- What is the entropy of this collection of data

b- What are the information gains of A1, A2 and A3 relative to this data?

c- What is the best split according to the information gain in part (b)?

d- Use Naïve Bayesian Classification method to classify an object which has the following attributes: {A1:T, A2:F, A3:N}

- For a given data in table5, count represents the number of data tuples having the values for department, status, age, and salary. Let status be the class label attribute.

a- What is the entropy of this collection of data .

b- What are the information gains of department, age, and salary

c- What is the best split according to the information gain in (b)?

d- Use Naïve Bayesian Classification method to classify an individual who has the following attributes:

***{ department ="marketing"; age= "youth", salary= low}***

| Table5 | | | | |
|---|---|---|---|---|
| department | age | salary | status | count |
| sales | Middle aged | medium | senior | 30 |
| sales | youth | low | junior | 30 |
| sales | Middle aged | low | junior | 40 |
| systems | youth | medium | junior | 20 |
| systems | Middle aged | high | senior | 20 |
| systems | senior | high | senior | 10 |
| marketing | senior | medium | senior | 10 |
| marketing | Middle aged | medium | junior | 20 |
| secretary | senior | medium | senior | 10 |
| secretary | youth | low | junior | 10 |

## Part4 : Clustering

- Suppose that the data mining task is to cluster the following nine points (with (x, y) representing location) into three clusters, the distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster.

  **A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9), C3(2,8):**

  - Use the *k-means* algorithm to show the three cluster centers after the first-round execution

- For the following points

  **(2, 5), (3, 10), (8, 4), (5, 8), (8, 5), (6, 4), (1, 2), (4, 9):**

Assume that k = 2 and initially assign (2, 5) and (3, 10) as the center of each cluster.

- Apply the k-means algorithm until the clusters do not change, using the Manhattan distance.

*(Hint: The Manhattan distance is: $d(i, j) = |x_{i1}-x_{j1}|+ |x_{i2}-x_{j2}|+ ....+ | x_{in}-x_{jn}|$.)*

- Suppose that the data mining task is to cluster the following points (with (x, y) representing location), The distance function is the Manhattan distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster.

**A1(1, 2), A2(7, 5), A3(6, 4), A4 (3, 4),**
**B1 (5, 8), B2(2, 5), B3(8, 4),**
**C1(2, 10), C2(4, 9), C3(2,8)**

- Use the k-means algorithm to show the three cluster centers after the first round execution**. (Hint: The Manhattan distance is: $d(i, j) = |x_{i1}-x_{j1}|+ |x_{i2}-x_{j2}|+ ....+ | x_{in}-x_{jn}|$.)**

- For the following nine points (1,2), (2,2), (2,3), (3,4), (4,3), (5,4), (6,6), (7,5), (8,4). Assume that k = 2 and initially assign (2,2) and (5,4) as the center of each cluster.
  - Apply the k-means algorithm using the Manhattan distance and show the new cluster centers after the first round execution (Hint: The Manhattan distance is: d(i, j) = |xi1-xj1|+ |xi2-xj2|+ ….+ | xin-xjn|.)
  - Compute the silhouette coefficient for object (3, 4). What is the meaning of the computed value?

- For the following nine points (2,2), (2,7), (3,1), (3,5), (4,3), (4,8), (5,2), (6,2), (6,5). Assume that k = 2 and initially assign (2,2) and (2,7) as the center of each cluster.
  - Apply the k-means algorithm using the Manhattan distance and show the new cluster centers after the first-round execution
  - Suppose that your result in (a) is the final cluster, how can you use it to detect the outlier? what is the object which likely be an outlier?

- Describe the principles and ideas regarding Agglomorative Hierarchical Clustering. - Show the different steps of the algorithm using the dissimilarity matrix below and complete link clustering. Give partial results after each step.

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **1** | 0 | | | | |
| **2** | 2 | 0 | | | |
| **3** | 4 | 3 | 0 | | |
| **4** | 10 | 7 | 9 | 0 | |
| **5** | 8 | 5 | 6 | 1 | 0 |

## Choose the Correct Answer.

1. For the following association rule: Computer → Webcam (60%, 100%): Which of the following is true?

I. 100% of costumers bought both a computer and a webcam
II. 60% of costumers bought both a computer and a webcam
III. 100% of costumers who bought a computer bought also a webcam
IV. 60% of costumers who bought a computer bought also a webcam

a. II only b III Only c. I and IV d. II and III

2. We have Market Basket data for 1,000 rental transactions at a Video Store. There are four videos for rent -- Video A, Video B, Video C and Video D. The probability that both Video C and Video D are rented at the same time is known as _____ .

a. Correlation b. support c. lift d. confidence

***Consider the following transaction database: Suppose that minsup is set to 40% and minconf. to 70%.***

| TransID | Items |
|---|---|
| T100 | A, B, C, D |
| T200 | A, B, C, E |
| T300 | A, B, E, F, H |
| T400 | A, C, H |

3. The support of the item set A, B, E is……

a. 50% b.40% c. 70% d. 66%

4. Based on the given minimum support the item set A,B,E is……..

a. frequent b. not frequent c. strong d. not strong

5. The confidence of the rule A, B → E is

a. 50% b.40% c. 100% d. 66%

6. Based on the given minimum confidence the rule A, B → E is……..

a. frequent b. not frequent c. strong d. not strong

7. The lift of the rule A, B → E is…….

a. 1.33 b.1 c. 0.89 d. 0.66

8. The value of the lift in the previous question means that items are…….

a. positive correlated b. negative correlated
c. independent d. strong

9. For the given data {33, 25, 42, 25, 31, 37, 46, 29, 38} the five numbers summery will be …

a. 25, 27, 32, 35, 46 b. 25, 27, 33, 35, 46 c. 14, 27, 33, 35, d. 19, 29, 32, 38, 43

10. If you use min-max normalization to transform the value 33 onto the range [1.0, 2.0] the new value is

a. 0.38 b. 1.38 c. 0.038 d. 1.038

11. Identify the outlier for the given data? 23, 34, 27, 7, 30, 26, 28, 31, 34

a. 7 b. 23 c. 31 d. 34

***From the given Confusion Matrix***

12. Accuracy is…….

a.0.99 b.0.95 c.0.86 d.0.05

13. Error rate is…….

a.0.99 b.0.95 c.0.86 d.0.05

14. Sensitivity is…….

a.0.99 b.0.95 c.0.86 d.0.05

15. Specificity is …….

a.0.99 b.0.95 c.0.86 d.0.05

| ***Confusion Matrix*** | | | | |
|---|---|---|---|---|
| | | Predicted | | Total |
| | | Yes | No | |
| Actual | Yes | 6954 | 46 | 7000 |
| | No | 412 | 2588 | 3000 |
| Total | | 7366 | 2634 | 10000 |

***Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:*** **C1: {(2, 2), (4, 4), (6, 6)}C2: {(0, 4), (4, 0)} C3: {(5, 5), (9, 9)}**

16. What will be the cluster centroids if you want to proceed for second iteration?

a. C1: (4, 4), C2: (2, 2), C3: (7, 7) b. C1: (6, 6), C2: (4, 4), C3: (9, 9)

c. C1: (2, 2), C2: (0, 0), C3: (5, 5) d. None of these

17. What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in second iteration?

a. 10 b. $5\sqrt{2}$ c. $13\sqrt{2}$ d. None of these

18. Consider the given data: {3, 4, 5, 10, 21, 32, 43, 44, 46, 52, 59, 67}, Using equal-width partitioning and four bins, how many values are there in the first bin?

a. 3 b. 4 c. 5 d. 6

19. If smooth by median is applied to the previous bins, what is the new value of the data in the first bin?

a. 4 b. 4.5 c. 5 d. 7.5

20. Which of the following lists all parts of the five-number summary?

a. Mean, Median, Mode, Range, and Total

b. Minimum, Quartile1, Median, Quartile3, and Maximum

c. Smallest, Q1, Q2, Q3, and Q4

d. Minimum, Maximum, Range, Mean, and Median

## Answer the following Questions

1. Define: A centroid in k-means.
2. Define: A core point in DBSCAN
3. Define: association and correlation analysis. Give an example
4. Define: cluster analysis. Give an example
5. Define: Data Cleaning, Data integration, Data reduction, Data transformation, Discretization
6. Define: outlier analysis. Give an example
7. Define: regression. Give an example
8. Give an example for nonparametric data reduction strategies.
9. Give an example for parametric data reduction strategies.
10. How does K-means differ from DBSCAN
11. How to assess the goodness of a rule?
12. How you can solve missing values problems
13. If a person's height is measured in inches then what kind of attribute you will use?
14. If the correlation coefficient of the items bred and rice is equal to 1.5. This means what?
15. If the covariance of the items bred and rice is equal to 1. This means what?
16. If the information gain of age and income attributes are 0.24 and 0.024 respectively which one you will chose as the splitting attribute
17. If the lift measure of the items bred and rice is equal to 0.5. This means what?
18. If the lift measure of the items bred and rice is equal to 1. This means what?
19. If the lift measure of the items bred and rice is equal to 1.5. This means what?
20. If the mean is equal to the median then this might be an indication that the data is what?

21. If the mean is larger than the median then this might be an indication that the data is what?
22. If the mean is smaller than the median then this might be an indication that the data is what?
23. If you have 100 values in my data and I add 5.0 to all of the values, then how will this change the median?
24. If you have 100 values in my data and I add 5.0 to all of the values, then how will this change the median?
25. List the Cluster Analysis Methods
26. List the Major Preprocessing Tasks That Improve Quality of Data
27. List the steps of knowledge discovery
28. List the transformation strategies
29. List the types of outliers. Give an example for each one.
30. The confidence for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean?
31. The support for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean?
32. What are rules conflicts? How can you solve it?
33. What are the data smoothing techniques?
34. What are the different strategies of data reduction?
35. What are the main advantages and disadvantages of Decision Tree classification algorithms?
36. What are the terminating conditions in decision tree induction?
37. What classifiers are normally considered to be easy to interpret?
38. What clustering algorithms can find clusters of arbitrary shape?
39. What data mining task should be used to detect fraudulent usage of credit cards?
40. What is over fitting? Briefly describe one method to prevent over-fitting in classification trees.
41. What is the Apriori property?
42. What is the bootstrap sampling?

43. What is the different between noise and outlier?
44. What is the different between symmetric and asymmetric binary attribute?
45. What is the five numbers summery of the data? How is it represented graphically?
46. What is the majority voting? When you use it?
47. What is the means of  association rule computer → webcam (60%, 100%)
48. What is the mode of the data? What is the mean of (bimodal, trimodel)
49. What is the problem that related to calculate the mean? How you can fix it?
50. What is the problem that related to use global constant to fill in the missing values.
51. What is the redundant attribute? How can you detect it?
52. What Kinds of Data Can Be Mined?
53. What Kinds of Patterns Can Be Mined?
54. When are objects *q & m* density-connected.
55. When is object *p* density-reachable from another object *q?*

## True or False

56. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms density-based algorithms
57. All continuous variables are ratio
58. Association rules provide information in the form of "if-then" statements.
59. Attributes are sometimes called variables and objects are sometimes called observations
60. Binary variables are sometimes continuous
61. Cluster is the process of finding a model that describes and distinguishes data classes or concepts.
62. Computing the total sales of a company. Is a data mining task?
63. Correlation analysis divides data into groups that are meaningful, useful, or both.

64. Database mining refers to the process of deriving high-quality information from text.
65. Dissimilarity matrix stores n data objects that have p attributes as an n-by-p matrix
66. Dividing the customers of a company according to their profitability. is a data mining task?
67. For an association rule, if we move one item from the right-hand-side to the left-hand-side of the rule, then the confidence will never change.
68. If all the proper subsets of an itemset are frequent, then the itemset itself must also be frequent.
69. In decision tree algorithms, attribute selection measures are used to rank attributes
70. In decision tree algorithms, attribute selection measures are used to reduce the dimensionality
71. In lazy learner we interest in the largest distance.
72. Intrinsic methods measure how well the clusters are separated
73. Multimedia Mining is the application of data mining techniques to discover patterns from the Web.
74. Regression is a method of integration
75. Strategies for data transformation include chi-square test
76. The Pruning make the decision tree more complex
77. An object is an outlier if its density is equal to the density of its neighbors.
78. A common weakness of association rule mining is that it is not produce enough interesting rules
79. Accuracy is interestingness measures for association rules
80. Binning is a method of reduction
81. Core object is an object whose $\epsilon$-neighborhood contains objects less than MinPts
82. Correlation analysis is used to eliminate misleading rules.
83. Correlation is a method of cleaning

84. Data matrix stores a collection of proximities for all pairs of n objects as an n-by-n matrix
85. Extracting the frequencies of a sound wave. Is a data mining task?
86. Incomplete data problem can be solved by binning
87. K-Nearest Neighbor Classifiers do classification when new test data is available
88. Median is a value that occurs most frequently in the attribute values
89. Mode is a middle value in set of ordered values
90. One strength of a Bayesian Classifier is that it can be easily trained
91. Outlier analysis is a method of transformation
92. Predicting the outcomes of tossing a (fair) pair of dice. Is a data mining task?
93. Recall is interestingness measures for association rules
94. Redundancy is an important issue in data cleaning
95. Sampling methods smooth noisy data
96. Sorting a student database based on student identification numbers. Is a data mining task?
97. The bottleneck of the Apriori algorithm is caused by the number of association rules
98. The goal of clustering analysis is to maximize the number of clusters
99. the object is local outlier if it is deviate significantly from the rest of the dataset
100. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms

**My best wishes;**