

Câu hỏi 1

Chưa được trả lời

Chấm điểm của
0,33T¹ Cờ câu hỏi

In logistic regression the cost function ($J(\theta) = \text{cost}(h_\theta(x), y)$) is calculated as...

- a. $\text{cost}(h_\theta(x), y) = y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))$
- b. $\text{cost}(h_\theta(x), y) = y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$
- c. $\text{cost}(h_\theta(x), y) = \frac{1}{2}(h_\theta(x)-y)^2$
- d. $\text{cost}(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$
- e. $\text{cost}(h_\theta(x), y) = \frac{1}{2}(h_\theta(x)-y)^{1/2}$

Câu hỏi 2

Chưa được trả lời

Chấm điểm của
0,33T¹ Cờ câu hỏi

(LO.4.4) Chose the correct statement about applying correlation analysis to detect and remove the redundancy between two categorical attributes A and B

- a. $r_{A,B} \sim 0$ then we can remove both A and B
- b. $r_{A,B} \sim 0$ then we can remove A or B
- c. Another solution
- d. $r_{A,B} \sim 1$ then we can remove A or B
- e. $r_{A,B} \sim 1$ then we can remove both A and B

Câu hỏi 3

Chưa được trả lời

Chấm điểm của
0,33T¹ Cờ câu hỏi

(LO.4.4) In chi-square (χ^2) analysis the _____ is a parameter describing the strength of the hypothesis.....

- a. degree of freedom, A and B are correlated
- b. no answer is correct
- c. degree of freedom, A and B are independent
- d. significance level, A and B are independent
- e. significance level, A and B are correlated

Câu hỏi 4

Chưa được trả lời

Chấm điểm của
0,33

T^o Cờ câu hỏi

(LO.4.3) In data cube aggregation.....

- a. the number of dimension is 3
- b. data are categorical
- c. data is in the highest level of details to improve the data mining effectiveness
- d. the data details are reserved
- e. data are numerical

Câu hỏi 5

Chưa được trả lời

Chấm điểm của
0,33

T^o Cờ câu hỏi

(LO.3.4) In the Apriori algorithm:

- a. $|C_k| \geq |L_k|$
- b. $|C_k| \geq |C_{k+1}|$
- c. the data set D is scan m times, where m is the length of the longest frequent itemset
- d. A and C are correct
- e. A, B and C are correct

Clear my choice

Câu hỏi 6

Chưa được trả lời

Chấm điểm của
0,33

T^o Cờ câu hỏi

(LO.3.3) Chose the correct statement:

- a. Compared to the K-means, the K-medoids algorithm is better in dealing with noise data
- b. Both the partition-based and the hierarchical clustering methods require a number of clusters as input
- c. Partition-based clustering methods commonly work well with sphere-shape clusters
- d. An advantage of a partition-based clustering method compared to a hierarchical clustering method is that it can return the previous step/loop in iterations
- e. A and C are correct

Câu hỏi 7

Chưa được trả lời

Chấm điểm của
0,33T^ǐ Cờ câu hỏi

(L.O.3.2) Which statement is correct about ANN?

- a. is a computing model that simulates the mechanism of actions in human brain
- b. the number of output nodes can be one or many depending the number of states of the data that need to be investigated by the system
- c. commonly used for classification
- d. A, B and C are correct
- e. all of the above are wrong

Câu hỏi 8

Chưa được trả lời

Chấm điểm của
0,33T^ǐ Cờ câu hỏi

(L.O.5.2) Which function below is not supported in Weka?

- a. build/train the model and then store it for later execution with new dataset
- b. attribute selection based on the correlation between the independent attributes and response attributes
- c. read data from ARFF file
- d. read data from CSV file
- e. all of the above are incorrect

Câu hỏi 9

Chưa được trả lời

Chấm điểm của
0,33T^ǐ Cờ câu hỏi

(L.O.1.1) Choose the wrong answer for the following statement:

Knowledge discovered by data mining could be....

Thời gian còn lại 0:40:33

- a. nontrivial
- b. understandable
- c. implicit
- d. useful for decision making
- e. explicit in the database

[Clear my choice](#)

1	2	3
10	11	12
19	20	21
28		

Hoàn thành bài I

Câu hỏi 10

Chưa được trả lời

Chấm điểm của
0,33T^ǐ Cờ câu hỏi

(L.O.1.2) Choose the correct statement:

- a. A particular data mining algorithm can work with data from only one data source
- b. Pattern is a chart presented to users
- c. Data mining algorithm works with task-relevant data
- d. Data warehouse is used to store the detailed data used for mining later on
- e. Data mining algorithm works directly with data sources

Câu hỏi 11

Chưa được trả lời

Chấm điểm của
0,33

T¹ Câu hỏi

(LO.1.1) The various aspects of data mining methodologies is/are.....

- i) Mining various and new kinds of knowledge
- ii) Mining knowledge in multidimensional space
- iii) Pattern evaluation and pattern or constraint-guided mining.
- iv) Handling uncertainty, noise, or incompleteness of data

Which of the following answer is correct?

- a. i, ii and iv only
- b. All i, ii, iii and iv
- c. ii, iii and iv only
- d. All i, ii, iii and iv are incorrect
- e. i, ii and iii only

Câu hỏi 12

Chưa được trả lời

Chấm điểm của
0,33

T¹ Câu hỏi

In decision tree method, which of the following measurement helps to avoid generating arbitrary small partitions at each splitting step?

- a. Information Gain
- b. GainRatio
- c. GiniIndex
- d. B and C are correct
- e. another solution

Câu hỏi 13

Chưa được trả lời

Chấm điểm của 0,33

T' Cờ câu hỏi

(LO.2.6)provides ways to learn the rules in the data which support for data mining methods

- a. Data pre-processing
- b. Machine learning
- c. Visualization
- d. Information science
- e. Database technology

Câu hỏi 14

Chưa được trả lời

Chấm điểm của 0,33

T' Cờ câu hỏi

(LO.2.3) What of the following is a model?

- a. $Y = aX + b$
- b. $Y = 3X + 2$
- c. $p(Y>y_1|X>x_1) = p_1$
- d. A, B, and C are correct
- e. None of the above

Câu hỏi 15

Chưa được trả lời

Chấm điểm của 0,33

T' Cờ câu hỏi

(LO.3.3) Which clustering method below is the most suitable one for identifying clusters with pipe shape?

- a. DBSCAN
- b. K-Means
- c. K-Medoids
- d. BIRCH
- e. B and C are correct

Câu hỏi 16

Chưa được trả lời

Chấm điểm của 0,33

T' Cờ câu hỏi

(LO.1.1) Strategic value of data mining is.....

- a. cost-sensitive
- b. time-sensitive
- c. work-sensitive
- d. technical-sensitive
- e. All above are incorrect

Câu hỏi 17

Chưa được trả lời

Chấm điểm của 0,33

[? Cứu hỏi](#)

(LO.2.6) Knowledge base is used for.....

- a. Knowledge visualization
- b. User interface
- c. Data cleansing and integration
- d. Data mining engine and pattern evaluation
- e. Data warehousing

Câu hỏi 18

Chưa được trả lời

Chấm điểm của 0,33

[? Cứu hỏi](#)

(LO.4.2) What are the tasks of data cleansing?

- a. transform data to a normalized format, create new attributes
- b. integrate data from various sources, remove data redundancy
- c. identify outliers, remove noise
- d. find out the data tendency, remove redundant attributes
- e. extracting useful knowledge

Câu hỏi 19

Chưa được trả lời

Chấm điểm của 0,33

[? Cứu hỏi](#)

(LO.3.4) Which statement is correct in association rule mining?

- a. support($A \Rightarrow B$) is always greater than confidence($A \Rightarrow B$)
- b. confidence($A \Rightarrow B$) is always smaller than support($A \Rightarrow B$)
- c. support is more important than confidence
- d. support_count($A \Rightarrow B$) is the number of transactions in D that contain both A and B
- e. confidence($A \Rightarrow B$) is always greater than support($A \Rightarrow B$)

Câu hỏi 20

Chưa được trả lời

Chấm điểm của 0,33

[? Cứu hỏi](#)

(LO.4.4) The main tasks of data summarization are:

- a. Identify the outlier and attribute redundancy
- b. Describe tendency and dispersion of the dataset
- c. Removing data redundancy
- d. Integrate detailed data from various sources to summarized data in a data warehouse
- e. Improve the pattern evaluation

Thời gian còn lại 0:25:30

Câu hỏi 21

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.4.2) Which statement related to Inter-quartile range below is wrong?

- a. Q1 is always greater than the minimum value in the dataset
- b. Q2 and median are similar
- c. Removing data redundancy
- d. Can be applied with categorical data
- e. Can be used for detecting outliers

[Clear my choice](#)

Câu hỏi 22

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.3.3) In density-based clustering method, which of the following statement is correct?

- a. each object in a cluster has at least MinPts neighbors (in the predefined radius ϵ)
- b. This method cannot identify the sphere-shape clusters
- c. the distance from an object A to a core object P is smaller than ϵ then A belongs to cluster C containing P
- d. This method is always more efficient than the K-Mean method
- e. each cluster has only one core object, that is the cluster centroid

Câu hỏi 23

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.4.1) Which statement below is wrong?

- a. In positive skew data, Mean is greater than Median
- b. This statement is always true: Minimum < Q1 < Median < Q3
- c. There is no case when Median = Mode
- d. "Extreme" means "Extreme outlier"

Câu hỏi 24

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.4.4) Which statement is wrong about data redundancy?

- a. Chi-square (χ^2) analysis is a correlation analysis method which is specific to categorical data
- b. It does not affect data mining methods, it is not necessary to resolve
- c. Data from an attribute can be inferred from others
- d. It can be detected by correlation analysis

Câu hỏi 25

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.4.1) Attribute subset selection will change the.....

- a. number of objects in the dataset
- b. number of attributes in the dataset
- c. probability distribution of different object classes in the dataset
- d. A and B are correct
- e. all of the above statements

Câu hỏi 26

Chưa được trả lời

Chấm điểm của
0,33

T' Cờ câu hỏi

(LO.3.1) Given a regression equation: $Y = f(X, \theta)$, chose the wrong statement:

- a. θ is a set of parameters
- b. X is an input vector
- c. Y changes when X changes
- d. θ describes the effects of X on Y
- e. X is a single variable

Câu hỏi 27

Chưa được trả lời

Chấm điểm của
0,75

T' Cờ câu hỏi

(LO.3.2, 0.5 points) Given a dataset in the table below

RID	Tuoi	Thu_nhap	Sinh_vien	Tin_dung	Mua_may_tinh
1	tre-	cao-	no-	kha-	khong_mua-
2	tre-	cao-	no-	tol-	khong_mua-
3	trung-	cao-	no-	kha-	mua-
4	cao-	trung_binh	no-	kha-	mua-
5	cao-	thap-	yes-	kha-	mua-
6	cao-	thap-	yes-	tol-	khong_mua-
7	trung-	thap-	yes-	tol-	mua-
8	tre-	trung_binh	no-	kha-	khong_mua-
9	tre-	thap-	yes-	kha-	mua-
10	cao-	trung_binh	yes-	kha-	mua-
11	tre-	trung_binh	yes-	tol-	mua-
12	trung-	trung_binh	no-	tol-	mua-
13	trung-	cao-	yes-	kha-	mua-
14	cao-	trung_binh	no-	tol-	khong_mua-

If we use the information gain measure to identify the splitting attribute, then the value of $Gain_{Tuoi}$ is:

- a. 0.151
- b. 0.246
- c. 0.94
- d. 0.029
- e. 0.694

Thời gian còn lại 0:14:49

(LO.3.2, 0.5 points) Given a dataset as in the following table, the last column (Camping) is the label of the corresponding input data represented by the four features in the left.

Outlook	Temp	Humidity	Windy	Camping
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

If we use the Naive Bayesian method to classify a data point X ($\text{Outlook} = \text{Sunny}$, $\text{Temp} = \text{Cool}$, $\text{Humidity} = \text{High}$, $\text{Windy} = \text{True}$) then $P(X|\text{Camping} = \text{Yes}) * P(\text{Camping} = \text{Yes})$ is:

If we use the Naive Bayesian method to classify a data point X ($\text{Outlook} = \text{Sunny}$, $\text{Temp} = \text{Cool}$, $\text{Humidity} = \text{High}$, $\text{Windy} = \text{True}$) then $P(X|\text{Camping} = \text{Yes}) * P(\text{Camping} = \text{Yes})$ is:

- a. 0.0503
- b. 0.0026
- c. 0.0246
- d. 0.0053
- e. none of the above