PRACTICE EXAM

DLMDSML01 1887
EXAMID: 1273337

MASTERSOLUTION

QUESTION 1 OF 18

Marked out of 3.00

Which of the following is required by K-means clustering?
Answer option 1: defined distance metric
Answer option 2: number of clusters

Select one:

- □ Both Answer option 1 and Answer option 2 are correct.
- □ Neither Answer option 1 nor Answer option 2 is correct.
- □ Only answer option 1 is correct.
- □ Only answer option 2 is correct.

The correct answer is: Both Answer option 1 and Answer option 2 are correct.

QUESTION 2 OF 18

Marked out of 3.00

Generally, which of the following method(s) is used for predicting continuous dependent variables?
Answer option 1: Linear Regression
Answer option 2: Logistic Regression

Select one:

- □ Neither Answer option 1 nor Answer option 2 is correct.
- □ Both Answer option 1 and Answer option 2 are correct.
- □ Answer option 1 only
- □ Answer option 2 only

The correct answer is: Answer option 1 only

QUESTION 3 OF 18

Marked out of 3.00

Which of the following indicates a fairly strong relationship between X and Y?

Select one:

- □ Correlation coefficient = 0.3
- □ Correlation coefficient = 0.9
- □ Correlation coefficient = 0.1
- □ The training is finalized after few iterations.

The correct answer is: Correlation coefficient = 0.9

QUESTION 4 OF 18

Marked out of 3.00

Which of the following is correct for the Support Vector Machine learning model?
Answer option 1: Only data vectors defining the margins are called the support vectors.
Answer option 2: SVM can only classify linearly separable data.

Select one:

- □ Only answer option 1 is correct.
- □ Answer option 1 and answer option 2 are both correct.
- □ Neither answer option 1 nor answer option 2 is correct.
- □ Only answer option 2 is correct.

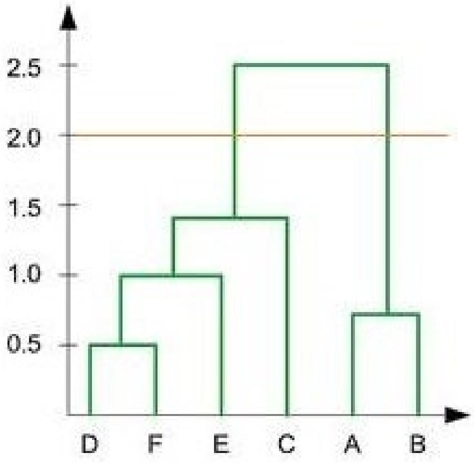
The correct answer is: Only answer option 1 is correct.

MASTERSOLUTION

QUESTION 5 OF 18

Marked out of 3.00

In the figure below, if you draw a horizontal line on y-axis for y=2.
What will be the number of clusters formed?



Select one:

□ 1
□ 3
□ 2
□ 4

The correct answer is: 2

QUESTION 6 OF 18

Marked out of 3.00

Classification problems fall under which category?

Select one:

- □ Supervised learning
- □ Reinforcement learning
- □ Cluster analysis
- □ Unsupervised learning

The correct answer is: Supervised learning

QUESTION 7 OF 18

Marked out of 3.00

… is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

Select one:

- □ Agglomerative clustering
- □ Non-hierarchical clustering
- □ Divisive clustering
- □ K-Means clustering

The correct answer is: Divisive clustering

MASTERSOLUTION

QUESTION 8 OF 18

Marked out of 3.00

Which of the following is/are **not** true about DBSCAN clustering algorithm:
Answer option 1: For data points to be in a cluster, they must be in a distance threshold to a core point
Answer option 2: It has strong assumptions for the distribution of data points in dataspace

Select one:

- □ Answer option 1 and Answer option 2 are correct.
- □ Neither Answer option 1 nor Answer option 2 is correct.
- □ Answer option 2 only
- □ Answer option 1 only

The correct answer is: Answer option 2 only

QUESTION 9 OF 18

Marked out of 3.00

Which of the following is correct in decision tree learning?
Answer option 1: The Information gain is calculated for discrete values only, while the gain ratio rank can handle both discrete and continuous values.
Answer option 2: The measurement of the information gain of a specific feature is poor if this feature is completely homogenous.

Select one:

- □ Only answer option 2 is correct.
- □ Answer option 1 and answer option 2 are both correct.
- □ Only answer option 1 is correct.
- □ Neither answer option 1 nor answer option 2 is correct.

The correct answer is: Only answer option 1 is correct.

QUESTION 10 OF 18

Marked out of 3.00

What is correct about evolutionary computation?
Answer option 1: Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.
Answer option 2: Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.

Select one:

- □ Answer option 1 and answer option 2 are both correct.
- □ Only answer option 2 is correct.
- □ Only answer option 1 is correct.
- □ Neither answer option 1 nor answer option 2 is correct.

The correct answer is: Only answer option 1 is correct.

QUESTION 11 OF 18

Marked out of 3.00

Genetic algorithms belong to the family of methods in which area?

Select one:

- □ Machine Learning and Optimization
- □ Complete enumeration methods
- □ Expert Systems
- □ Logical reasoning algorithms

The correct answer is: Machine Learning and Optimization

QUESTION 12 OF 18

Marked out of 3.00

A … regression can be employed to predict which grade a student will achieve in an exam based on hours of study, previous grades and other related factors.

Select one:

□ Linear
□ Multilinear
□ Quantile
□ Logistic

The correct answer is: Multilinear

QUESTION 13 OF 18

Marked out of 3.00

Which approach of the following is applied to improve the prediction performance of decision trees?
Answer option 1: Bagging
Answer option 2: Boosting

Select one:

□ Answer option 1 and answer option 2 are both correct.
□ Neither answer option 1 nor answer option 2 is correct.
□ Only answer option 1 is correct.
□ Only answer option 2 is correct.

The correct answer is: Answer option 1 and answer option 2 are both correct.

QUESTION 14 OF 18

Marked out of 3.00

Which of the following is correct for the simple Genetic Algorithm?
Answer option 1: Children compete with parents in survival selection.
Answer option 2: Both crossover and mutation are applied in each generation.

Select one:

□ Only answer option 2 is correct.
□ Neither answer option 1 nor answer option 2 is correct.
□ Only answer option 1 is correct.
□ Answer option 1 and answer option 2 are both correct.

The correct answer is: Only answer option 2 is correct.

QUESTION 15 OF 18

Marked out of 6.00

Mention three advantages of the decision tree technique.

1. Easy to Understand: Its graphical representation is very intuitive and users can easily relate their hypothesis.
2. Useful in Data exploration: Decision tree is one of the fastest ways to identify most significant variables and relations between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable.
3. Less data cleaning required: It is not influenced by outliers and missing values to a fair degree.
4. It can handle both numerical and categorical variables.
5. Non Parametric Method: This means that decision trees have no assumptions about the space distribution and the classifier structure.

[2 pts for each, maximum 6 pts]

QUESTION 16 OF 18

Marked out of 6.00

Suppose you are using RBF kernel in SVM with a certain Gamma value.

a. What does a low Gamma value signify?
b. What does a high Gamma value signify?

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.

a. For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape **(3 pts)**. It behaves like a linear separator between the density centers of the two classes
b. For a higher gamma, the model will capture the shape of the dataset well - yet incurs the prospect of overfitting **(3 pts)**.

QUESTION 17 OF 18

Marked out of 18.00

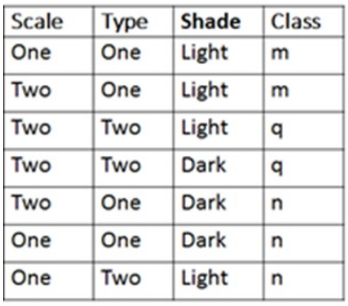
What trade-offs are incurred for choosing small or large values of the regularization parameter in regression models?
How can one choose an appropriate value of the regularisation parameter (λ)?

Selecting the regularisation parameter is a tricky business. If the value of λ is too high, it will lead to extremely small values of the regression coefficient β, which will lead to the model underfitting **(2 pts)** (high bias – low variance) **(2 pts)**. On the other hand, if the value of λ is 0 (very small), the model will tend to overfit the training data **(2 pts)** (low bias – high variance)**(2 pts)**.
There is no proper way to select the value of λ. What you can do is have a sub-sample of data**(2 pts)** and run the algorithm multiple times on different sets **(2 pts)**. Here, the person has to decide how much variance can be tolerated **(2 pts)**. Once the user is satisfied with the variance, that value of λ can be chosen for the full dataset **(2 pts)**.
One thing to be noted is that the value of λ selected here was optimal for that subset, not for the entire training data **(2 pts)**.

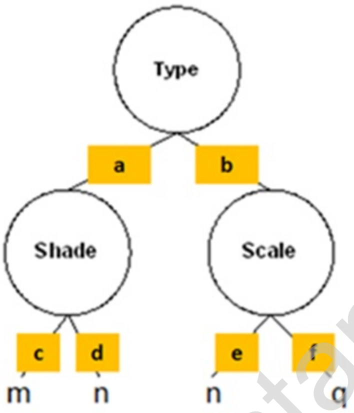
QUESTION 18 OF 18

Marked out of 18.00

What are the values of a, b, c, d, e, and f in the following decision tree?



Scale	Type	Shade	Class
One	One	Light	m
Two	One	Light	m
Two	Two	Light	q
Two	Two	Dark	q
Two	One	Dark	n
One	One	Dark	n
One	Two	Light	n



a= One
b= Two
c= Light
d= Dark
e= One
f= Two

[3 pts for each]