NPTEL Online Certification Courses Indian Institute of Technology Kharagpur

Introduction to Machine Learning

Assignment- Week 1

TYPE OF QUESTION: MCQ

Number of questions: **10** **Total mark: 10 X 2 = 20**

MCQ Question

QUESTION 1:

Which of the following is not a type of supervised learning?

A. Classification
B. Regression
C. Clustering
D. None of the above

Correct Answer: C. Clustering

Detailed Solution : Classification and Regression are both supervised learning as they need class labels or target values for training, but Clustering doesn't need target values.

QUESTION 2:

As the amount of training data increases

A. Training error usually decreases and generalization error usually increases
B. Training error usually decreases and generalization error usually decreases
C. Training error usually increases and generalization error usually decreases
D. Training error usually increases and generalization error usually increases

Correct Answer: A. Training error usually decreases and generalization error usually increases

Detailed Solution: When the training data increases, the decision boundary becomes very complex to fit the data. So, the generalization capability usually reduces with the increase in training data.

QUESTION 3:

Suppose I have 10,000 emails in my mailbox out of which 300 are spams. The spam detection system detects 150 mails as spams, out of which 50 are actually spams. What is the precision and recall of my spam detection system ?

A. **Precision = 33.33%, Recall = 16.66%**
B. Precision = 25%, Recall = 33.33%
C. Precision = 33.33%, Recall = 75%
D. Precision = 75%, Recall = 33.33%

Correct Answer: A. Precision = 33.33%, Recall = 16.66%

Detailed Solution:

Precision = $\frac{T_p}{T_p+F_p} = \frac{50}{50+100} = \frac{50}{150} = 33.33\%$

Recall = $\frac{T_p}{T_p+F_n} = \frac{50}{50+250} = 16.66\%$

QUESTION 4:

Which of the following are not classification tasks ?

A. Find the gender of a person by analyzing his writing style
B. **Predict the price of a house based on floor area, number of rooms etc**.
C. Predict whether there will be abnormally heavy rainfall next year
D. Detect Pneumonia from Chest X-ray images

Correct Answer: B. Predict the price of a house based on floor area, number of rooms etc.

Detailed Solution : House Price is a continuous real valued variable, so we have to use regression methods for this task.

QUESTION 5:

Occam's razor is an example of:

A. Inductive bias
B. Preference bias

Correct Answer: B. Preference bias

Detailed Solution : Prefer simplest hypothesis over complex one

---

QUESTION 6:

A feature F1 can take certain value: A, B, C, D, E, F and represents grade of students from a college. Which of the following statements is true in the following case?

A. Feature F1 is an example of a nominal variable.
B. Feature F1 is an example of ordinal variables.
C. It doesn't belong to any of the above categories.
D. Both of these

Correct Answer: B. Feature F1 is an example of ordinal variables.

Detailed Solution : Ordinal variables are the variables which have some order in their categories. For example, grade A should be considered as higher grade than grade B.

QUESTION 7:

Which of the following is a categorical feature?

A. Height of a person
B. Price of petroleum
C. Mother tongue of a person
D. Amount of rainfall in a day

Correct Answer: C. Mother tongue of a person

Detailed Solution : Categorical variables represent types of data which may be divided into groups. All other features are continuous)

QUESTION 8:

Which of the following tasks is NOT a suitable machine learning task?

A. **Finding the shortest path between a pair of nodes in a graph**
B. Predicting if a stock price will rise or fall
C. Predicting the price of petroleum
D. Grouping mails as spams or non-spams

Correct Answer : A. Finding the shortest path between a pair of nodes in a graph

Detailed Solution : Finding the shortest path is a graph theory based task, whereas other options are completely suitable for machine learning.

QUESTION 9:

Which of the following is correct for reinforcement learning?

A. The algorithm plans a sequence of actions from the current state.
B. The algorithm plans one action at each time step.
C. The training instances contain examples of states and best actions of the states.
D. The algorithm groups unseen data based on similarity.

Correct Answer : B.The algorithm plans one action at each time step.

Detailed Solution : In reinforcement learning, the agent tries to learn the policy function i.e. the best action to take at a given state.

QUESTION 10:

What is the use of Validation dataset in Machine Learning?

A. To train the machine learning model.
B. To evaluate the performance of the machine learning model
C. To tune the hyperparameters of the machine learning model
D. None of the above.

Correct Answer : C. To tune the hyperparameters of the machine learning model

Detailed Solution : The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyperparameters.

*********END*********

NPTEL Online Certification Courses Indian Institute of Technology Kharagpur

Introduction to
Machine Learning
Assignment- Week 2
TYPE OF QUESTION: MCQ

Number of questions: **8** **Total mark: 8 X 2 = 16**

MCQ Question

QUESTION 1:

Identify whether the following statement is true or false?

“Overfitting is more likely when the set of training data is small”

A. **True**
B. False

Correct Answer : **A.True**

Detailed Solution : With a small training dataset, it’s easier to find a hypothesis to fit the training data exactly,i.e., overfit.

QUESTION 2:

Which of the following criteria is typically used for optimizing in linear regression.

A. Maximize the number of points it touches.
B. Minimize the number of points it touches.
C. **Minimize the squared distance from the points.**
D. Minimize the maximum distance of a point from a line.

Correct Answer : **C. Minimize the squared distance from the points.**

Detailed Solution : Loss function of linear regression is squared distance from the points.

QUESTION 3:

Which of the following is false?

A. Bias is the true error of the best classifier in the concept class
B. Bias is high if the concept class cannot model the true data distribution well
C. **High bias leads to overfitting**
D. For high bias both train and test error will be high

Correct Answer : C. High bias leads to overfitting

Detailed Solution : High bias leads to underfitting.

QUESTION 4:

The following dataset will be used to learn a decision tree for predicting whether a person is happy (H) or sad (S), based on the color of shoes, whether they wear a wig and the number of ears they have.

Color	Wig	Num. Ears	Emotion (Output)
G	Y	2	S
G	N	2	S
G	N	2	S
B	N	2	S
B	N	2	H
R	N	2	H
R	N	2	H
R	N	2	H
R	Y	3	H

Which attribute should you choose as the root of the decision tree?

A. Color

B. Wig

C. Number of ears

D. Any one of the previous three attributes

Correct Answer : A. Color

Detailed Solution : We have to compute Information Gain w.r.t. each of these 4 attributes and the attribute with highest information gain will be chosen as the root of the decision tree.

QUESTION 5:

Consider applying linear regression with the hypothesis as $h_\theta(x) = \theta_0 + \theta_1 x$.The training data is given in the table.

X	Y
6	7
5	4
10	9
3	4

The cost function is $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$

What is the value of $J(\theta)$when **θ** = (2, 1) ?

A. 0

B. 1

C. 2

D. 2.5

Correct Answer: D. 2.5

Detailed Solution : Substitute θ_0 by 2 and θ_1 by 1 and compute $J(\theta)$.

QUESTION 6:

In a binary classification problem, out of 64 data points 29 belong to class I and 35 belong to class II. What is the entropy of the data set?

A. 0.97
B. 0
C. 1
D. 0.99

Correct Answer : D. 0.99

Detailed Solution : We can compute Entropy as

$ENTROPY(p_+, p_-) = -p_+ log_2 p_+ - p_- log_2 p_-$, here

$p_+ = 29/64$ *and* $p_- = 35/64$

QUESTION 7:

Decision trees can be used for the following type of datasets:

I. The attributes are categorical
II. The attributes are numeric valued and continuous
III. The attributes are discrete valued numbers

A. In case I only
B. In case II only
C. In cases II and III only
D. In cases I, II and III

Correct Answer : D. In cases I, II and III

Detailed Solution : Decision trees can be applied in all 3 cases.

QUESTION 8:

What is true for Stochastic Gradient Descent?

A. In every iteration, model parameters are updated for multiple training samples
B. In every iteration, model parameters are updated for one training sample
C. In every iteration, model parameters are updated for all training samples
D. None of the above

Correct Answer : B. In every iteration model parameters are updated for one training sample.

Detailed Solution : In batch gradient descent, multiple training samples are used and in stochastic gradient descent, one training sample is used to update parameters.

*******END*******

NPTEL Online Certification Courses Indian Institute of Technology Kharagpur

Introduction to Machine Learning

Assignment- Week 3

TYPE OF QUESTION: MCQ

Number of questions: 8 **Total mark: 8 X 2 = 16**

QUESTION 1:

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

X	Y	Class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Suppose, you want to predict the class of new data point x=1 and y=1 using euclidean distance in 7-NN. To which class the data point belongs to?

A. + Class
B. – Class
C. Can't say
D. None of these

Correct Answer: **B. – Class**

Detailed Solution : We have to compute the euclidean distance from the given point (1,1) to all the data points given in the dataset and based on that we have to check the dominating class for the 7 nearest points.

QUESTION 2:

Imagine you are dealing with 15 class classification problem. What is the maximum number of discriminant vectors that can be produced by LDA?

A. 20
B. 14
C. 21
D. 10

Correct Answer: B. 14

Detailed Solution : LDA produces at most c − 1 discriminant vectors, c = no of classes

QUESTION 3:

'People who bought this, also bought…' recommendations seen on amazon is a result of which algorithm?

A. User based Collaborative filtering
B. Content based filtering
C. Item based Collaborative filtering
D. None of the above

Correct Answer: C. Item based Collaborative filtering

Detailed Solution : Though both User based and Item based CF methods are used in recommendation systems, Amazon specifically uses Item based filtering.

QUESTION 4:

Which of the following is/are true about PCA?

1. PCA is a supervised method
2. It identifies the directions that data have the largest variance
3. Maximum number of principal components <= number of features
4. All principal components are orthogonal to each other

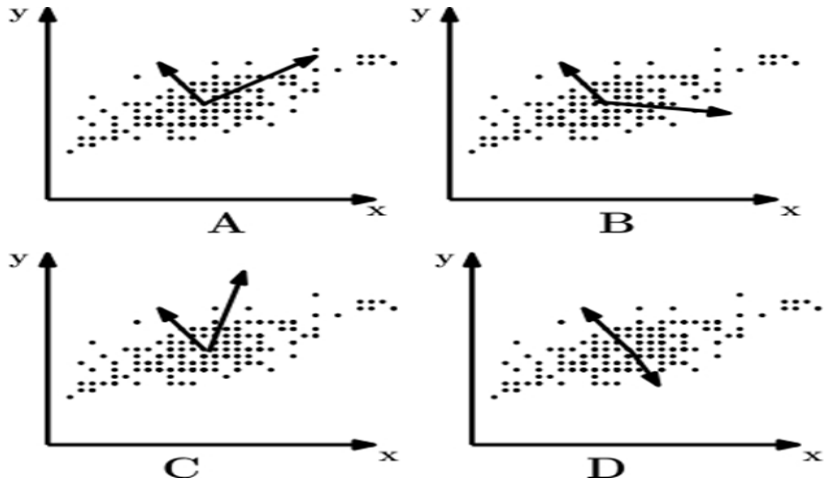
A. Only 2
B. 1, 3 and 4
C. 1, 2 and 3
D. 2, 3 and 4

Correct Answer: D

Detailed Solution : PCA is an unsupervised learning algorithm, so 1 is wrong. Other options are true about PCA.

QUESTION 5:

Consider the figures below. Which figure shows the most probable PCA component directions for the data points?



A. A

B. B

C. C

D. D

Correct Answer: **A. A**

Detailed Solution : PCA tries to choose the direction in such a way that maximizes the variance in the data.

QUESTION 6:

When there is noise in data, which of the following options would improve the performance of the KNN algorithm?

A. Increase the value of k

B. Decrease the value of k

C. Changing value of k will not change the effect of the noise

D. None of these

Correct Answer: A. Increase the value of k

Detailed Solution : Increasing the value of k reduces the effect of the noise and improves the performance of the algorithm.

QUESTION 7:

Which of the following statements is True about the KNN algorithm?

A. **KNN algorithm does more computation on test time rather than train time.**
B. KNN algorithm does lesser computation on test time rather than train time.
C. KNN algorithm does an equal amount of computation on test time and train time.
D. None of these.

Correct Answer: A. KNN algorithm does more computation on test time rather than train time.

Detailed Solution : The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.
In the testing phase, a test point is classified by assigning the label which are most frequent among the k training samples nearest to that query point – hence higher computation.

QUESTION 8:

Find the value of the Pearson's correlation coefficient of X and Y from the data in the following table.

AGE (X)	GLUCOSE (Y)
43	99
21	65
25	79
42	75

A. 0.47
B. 0.68
C. 1
D. 0.33

Correct Answer : B. 0.68

Detailed Solution : Pearson Coefficient $r = \frac{\sum_i (X_i - \overline{X})((Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2}\sqrt{\sum_i (Y_i - \overline{Y})^2}}$

********END*******

Introduction to Machine Learning

Assignment- Week 4

TYPE OF QUESTION: MCQ

Number of questions: 7 **Total mark:** 7 X 2 = 14

QUESTION 1:

A spam filtering system has a probability of 0.95 to correctly classify a mail as spam and 0.10 probability of giving false positives. It is estimated that 1% of the mails are actual spam mails.

Suppose that the system is now given a new mail to be classified as spam/ not-spam, what is the probability that the mail will be classified as spam?

A. 0.89575
B. 0.10425
C. 0.1085
D. 0.0995

Correct Answer: C. 0.1085

Detailed Solution:

Let S = 'Mails correctly marked spam by the system', T= 'Mails misclassified by the system'(Marked as spam when not spam or Marked as not spam when it is a spam) , M = 'Spam mails'.

P(S|M)= 0.95 , P(S|M') = 0.10 , P(M) = 0.01

We have to find the probability of mail being classified as spam which can either be if a spam mail is correctly classified as spam or if a mail is misclassified as spam.

$$P(S) = P(S|M) * P(M) + P(S|M') * P(M') = 0.95 * 0.01 + 0.10 * 0.99 = 0.1085$$

QUESTION 2:

Bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags and it is found to be black. Find the probability that it was drawn from Bag I.

A. 1/2
B. 2/3

C. 7/12

D. 9/23

Correct Answer : C. 7/12

Detailed Solution :

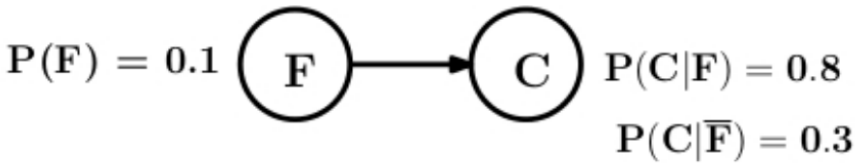
B1: "Ball is drawn from bag I", B2: "Ball is drawn from bag II", W: "Drawn ball is white", B: "Drawn ball is black"

We have to find $P(B1|B)$

$$P(B1|B) = \frac{P(B|B1)*P(B1)}{P(B|B1)*P(B1)+P(B|B2)*P(B2)} = \frac{(6/10)*(1/2)}{(6/10)*(1/2)+(3/7)*(1/2)} = \frac{3/10}{3/10+3/14} = \frac{7}{12}$$

QUESTION 3:

4. Consider the following Bayesian network, where F = having the flu and C = coughing:



Find P(C) and P(F|C).

A. 0.35, 0.23

B. 0.35,0.77

C. 0.24, 0.024

D. 0.5, 0.23

Correct Answer: A. 0.35, 0.23

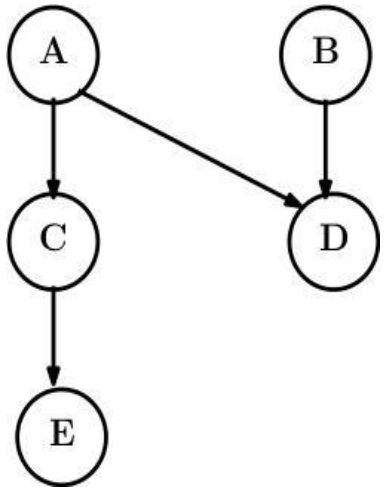
Detailed Solution :

$$P(C) = P(C|F) * P(F) + P(C|\bar{F}) * P(\bar{F})$$

$$P(F|C) = \frac{P(C|F)*P(F)}{P(C|F)*P(F)+P(C|\bar{F})*P(\bar{F})}$$

QUESTION 4:

Consider the following Bayesian network.



Thus, the independence expressed in this Bayesian net are that
A and B are (absolutely) independent.
C is independent of B given A.
D is independent of C given A and B.
E is independent of A, B, and D given C.

Suppose that the net further records the following probabilities:
$P(A) = 0.3$
$P(B) = 0.6$
$P(C|A) = 0.8$
$P(C|\overline{A}) = 0.4$
$P(D|A,B) = 0.7$
$P(D|A,\overline{B}) = 0.8$
$P(D|\overline{A},B) = 0.1$
$P(D|\overline{A},\overline{B}) = 0.2$
$P(E|C) = 0.7$
$P(E|C) = 0.7$

Find $P(D)$.

A. 0.32
B. 0.50
C. 0.40
D. 0.78

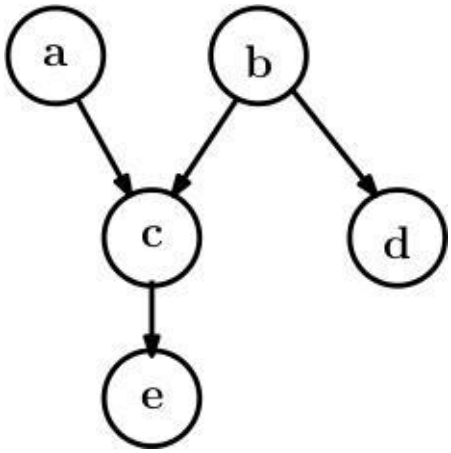
Correct Answer: A. 0.32

Detailed Solution :

$P(D) = P(D|AB) * P(AB) + P(D|\overline{A}B) * P(\overline{A}B) + P(D|A\overline{B}) * P(A\overline{B}) + P(D|\overline{A}\,\overline{B}) * P(\overline{A}\,\overline{B}) = 0.32$

QUESTION 5:

Consider the following graphical model, mark which of the following pair of random variables are independent given no evidence?



A. a,b
B. c,d
C. e,d
D. c,e

Correct Answer : A. a,b

Detailed Solution : Nodes a and b don't have any predecessor nodes. As they don't have any common parent node, a and b are independent.

QUESTION 6:

In a Bayesian network a node with only outgoing edge(s) represents

A. a variable conditionally independent of the other variables.
B. a variable dependent on its siblings.
C. a variable whose dependency is uncertain.
D. None of the above.

Correct Answer: A. a variable conditionally independent of the other variables.

Detailed Solution : As there is no incoming edge for the node, the node is not conditionally dependent on any other node.

QUESTION 7:

It is given that $P(A|B) = 2/3$ and $P(A|\overline{B}) = 1/3$. Compute the value of $P(B|A)$.

A. ½
B. ⅔
C. ¾
D. Not enough information.

Correct Solution : D. Not enough information.

Detailed Solution : There are 3 unknown probabilities $P(A), P(B), P(AB)$which can not be computed from the 2 given probabilities. So, we don't have enough information to compute $P(B|A)$.

**********END*******

NPTEL Online Certification Courses

Indian Institute of Technology Kharagpur

Course -Introduction to Machine Learning

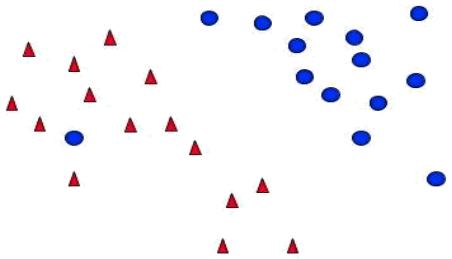
Assignment- Week 5 (Logistic Regression, SVM, Kernel Function, Kernel SVM)

TYPE OF QUESTION: MCQ/MSQ

Number of Question: 10 **Total Marks: 10x2 = 20**

1. What would be the ideal complexity of the curve which can be used for separating the two classes shown in the image below?

   A) Linear
   B) Quadratic
   C) Cubic
   D) insufficient data to draw conclusion



Answer: A
(The blue point in the red region is an outlier (most likely noise). The rest of the data is linearly separable.)

2. I. Logistic Regression is used for regression purposes.
   II. Logistic Regression is used for classification purposes.

   A) Only I is Correct
   B) Only II is Correct
   C) Both I and II are Correct
   D) Both I and II are Incorrect

Answer: C
Logistic Regression is used for both the calssfication and regression task.

3. Which of the following methods do we use to best fit the data in Logistic Regression?

   A) Least Square Error
   B) Maximum Likelihood
   C) Jaccard distance
   D) Both A and B

Answer: B
In logistic regression, both least square error and maximum likelihood are used as estimation methods for fitting the data.

4. Consider a following model for logistic regression: P(y=1|x,w)=g(w0+w1x) where g(z) is the logistic function.

In the above equation the P(y =1|x; w), viewed as a function of x, that we can get by changing the parameters w.

What would be the range of P in such a case?

A) (-inf,0)
B) (0,1)
C) (-inf, inf)
D) (0,inf)

Answer: B
For values of x in the range (-inf ,+inf), logistic function always give a output in the range (0,1).

5. **State whether True or False.**
After training an SVM, we can discard all examples which are not support vectors and can still classify new examples.

A) TRUE
B) FALSE

Answer: A
This is true because the support vectors only affect the boundary.

6. Suppose you are dealing with 3 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method.

How many times we need to train our SVM model in such case?

A) 1
B) 2
C) **3**
D) 4

Answer: C

In a N-class classification problem, we have to train the SVM at least N times in a one vs all method.

7. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

A) 1
B) 2
C) 1 and 2
D) None of these.

Answer: C
Kernels are used in SVMs to map low dimensional data into high dimensional feature space to classify non-linearly separable data. It is a similarity function between low-dimensional data points and its high dimensional feature space to find out what data points can be mapped into what sort of feature space.

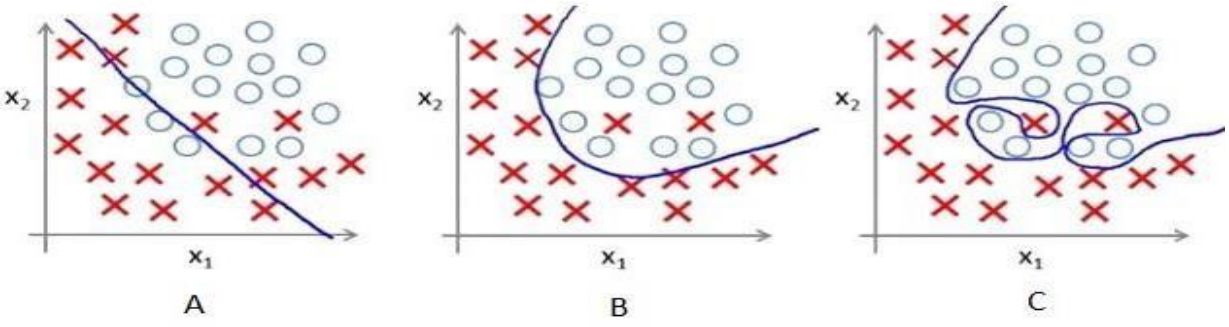
8. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

A) The model would consider even far away points from hyperplane for modelling.
B) **The model would consider only the points close to the hyperplane for modelling**.
C) The model would not be affected by distance of points from hyperplane for modelling.
D) None of the above

Answer: B
The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.
For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.
For a higher gamma, the model will capture the shape of the dataset well.

9. Below are the labelled instances of 2 classes and hand drawn decision boundaries for logistic regression. Which of the following figure demonstrates overfitting of the training data?

A) A
B) B
C) **C**
D) None of these



Answer: C
In figure 3, the decision boundary is very complex and unlikely to generalize the data.

10. What do you conclude after seeing the visualization in previous question?

C1. The training error in first plot is higher as compared to the second and third plot.
C2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
C3. Out of the 3 models, the second model is expected to perform best on unseen data.
C4. All will perform similarly because we have not seen the test data.

A) C1 and C2
B) C1 and C3
C) C2 and C3
D) C4

Answer: B
From the visualization, it is clear that the misclassified samples are more in the plot A when compared to B. So, C1 is correct. In figure 3, the training error is less due to complex boundary. So, it is unlikely to generalize the data well. Therefore, option C2 is wrong.
The first model is very simple and underfits the training data. The third model is very complex and overfits the training data. The second model compared to these models has less training error and likely to perform well on unseen data. So, C3 is correct.
We can estimate the performance of the model on unseen data by observing the nature of the decision boundary. Therefore, C4 is incorrect

End

NPTEL Online Certification Courses

Indian Institute of Technology Kharagpur

Course Name – Introduction To Machine Learning

Assignment – Week 6 (Neural Networks)

TYPE OF QUESTION: MCQ/MSQ

Number of Question: 8 **Total Marks: 8x2 = 16**

1. In training a neural network, we notice that the loss does not increase in the first few starting epochs: What is the reason for this?

   I) The learning Rate is low.
   II) Regularization Parameter is High.
   III) Stuck at the Local Minima.
   IV) **All of these could be the reason.**

   Answer: D
   The problem can occur due to any one of the reasons above.

2. What is the sequence of the following tasks in a perceptron?

   I) Initialize the weights of the perceptron randomly.
   II) Go to the next batch of data set.
   III) If the prediction does not match the output, change the weights.
   IV) For a sample input, compute an output.

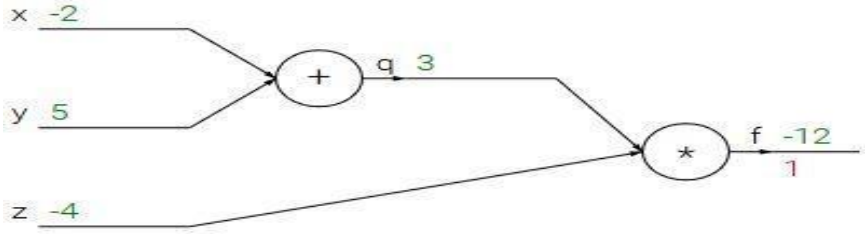
   A) I, II, III, IV
   B) IV, III, II, I
   C) III, I, II, IV
   D) I, IV, III, II

   Answer: D
   D is the correct sequence.

3. Suppose you have inputs as x, y, and z with values -2, 5, and -4 respectively. You have a neuron ‘q’ and neuron ‘f’ with functions:

   q = x + y
   f = q * z

   Graphical representation of the functions is as follows:

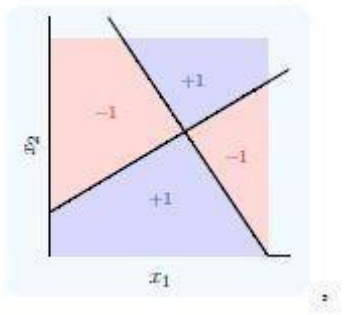


What is the gradient of F with respect to x, y, and z?

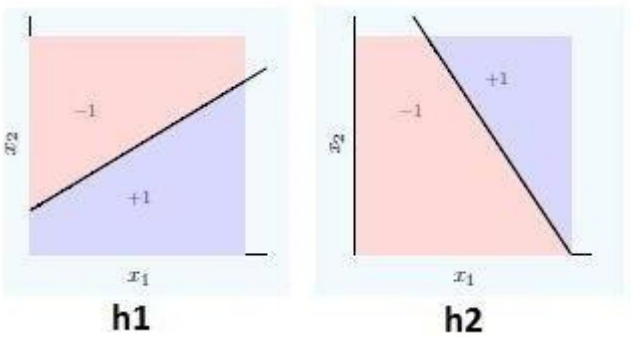
A) (-3, 4, 4)
B) (4, 4, 3)
C) (-4, -4, 3)
D) (3, -4, -4)

Answer: C
To calculate gradient, we should find out (df/dx), (df/dy) and (df/dz).

4. A neural network can be considered as multiple simple equations stacked together. Suppose we want to replicate the function for the below mentioned decision boundary.



Using two simple inputs h1 and h2,



What will be the final equation?

I) (h1 AND NOT h2) OR (NOT h1 AND h2**)**
II) (h1 OR NOT h2) AND (NOT h1 OR h2)
III) (h1 AND h2) OR (h1 OR h2)
IV) None of these

Answer: C
As you can see, combining h1 and h2 in an intelligent way can get you a complex equation.

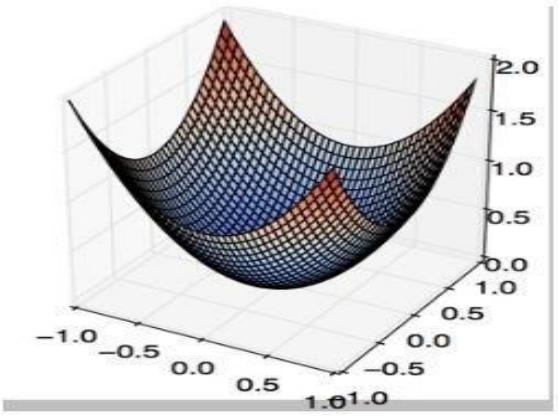
5. Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions)?

   I) **As number of hidden layers increase, model capacity increases**
   II) As dropout ratio increases, model capacity increases
   III) As learning rate increases, model capacity increases
   IV) None of these.

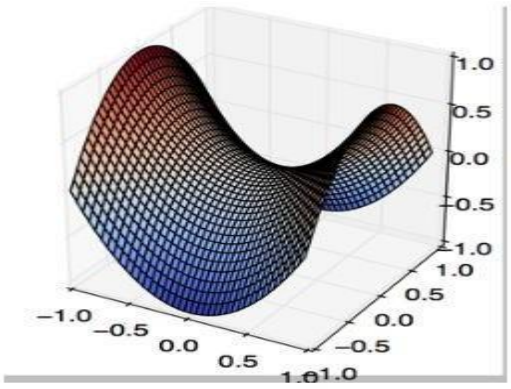
Answer: A
Option A is correct.

6. First Order Gradient descent would not work correctly (i.e. may get stuck) in which of the following graphs?

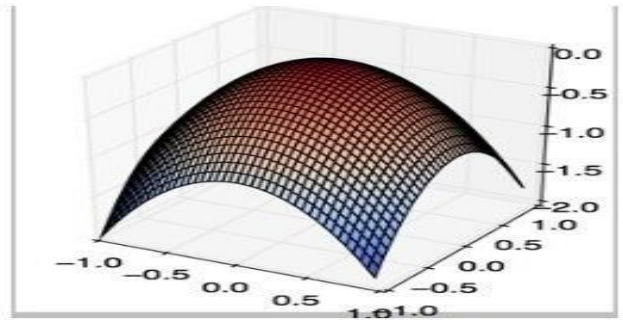
A)



B)



C)



D) None of These.

Answer: B
This is a classic example of saddle point problem of gradient descent.

7. Which of the following is true?
Single layer associative neural networks do not have the ability to

I) Perform pattern recognition
II) Find the parity of a picture
III) Determine whether two or more shapes in a picture are connected or not

A) II and III are true
B) II is true
C) All of the above
D) None of the above

Answer: A
Pattern recognition is what single layer neural networks are best at but they do not have the ability to find the parity of a picture or to determine whether two shapes are connected or not.

8. The network that involves backward links from outputs to the inputs and hidden layers is called as

A) Self-organizing Maps
B) Perceptron
C) Recurrent Neural Networks
D) Multi-Layered Perceptron

Answer: C

End

NPTEL Online Certification Courses
Indian Institute of Technology Kharagpur

Course Name: Introduction to Machine Learning
Assignment – Week 7 (Computational Learning theory, PAC Learning, Sample Complexity, VC Dimension, Ensemble Learning)
TYPE OF QUESTION: MCQ/MSQ

Number of Question: 8 **Total Marks: 8X2 = 16**

1. Which of the following option is / are correct regarding the benefits of ensemble model?

   1. Better performance
   2. More generalized model
   3. Better interpretability

   A) 1 and 3
   B) 2 and 3
   C) 1 and 2
   D) 1, 2 and 3

   Answer: C (1 and 2 are the benefits of ensemble modelling. Option 3 is incorrect because when we ensemble multiple models, we lose interpretability of the models).

2. In AdaBoost, we give more weights to points having been misclassified in previous iterations. Now, if we introduced a limit or cap on the weight that any point can take (for example, say we introduce a restriction that prevents any point's weight from exceeding a value of 10). Which among the following would be an effect of such a modification?

   A) We may observe the performance of the classifier reduce as the number of stages increase.
   B) It makes the final classifier robust to outliers.
   C) It may result in lower overall performance.
   D) None of these.

   Answer: B, C (Outliers tend to get misclassified. As the number of iterations increase, the weight corresponding to outlier points can become very large resulting in subsequent classifier models trying to classify the outlier points correctly. This generally has an adverse effect on the overall classifier. Restricting the weights is one way of mitigating this problem. However, this can also lower the performance of the classifier).

3. Which among the following are some of the differences between bagging and boosting?

   A) In bagging we use the same classification algorithm for training on each sample of the data, whereas in boosting, we use different classification algorithms on the different training data samples.

B) Bagging is easy to parallelize whereas boosting is inherently a sequential process.
C) In bagging we typically use sampling with replacement whereas in boosting, we typically use weighted sampling techniques.
D) In comparison with the performance of a base classifier on a particular data set, bagging will generally not increase the error whereas as boosting may lead to an increase in the error.

Answer: Options (B), (C) and (D) are correct.

4. What is the VC-dimension of the class of circle in a 4-dimensional plane?

A) 3
B) 4
C) **5**
D) 6

Answer: C is the correct option.

5. Considering the AdaBoost algorithm, which among the following statements is true?

A) In each stage, we try to train a classifier which makes accurate predictions on any subset of the data points where the subset size is at least half the size of the data set.
B) **In each stage, we try to train a classifier which makes accurate predictions on a subset of the data points where the subset contains more of the data points which were misclassified in earlier stages**.
C) The weight assigned to an individual classifier depends upon the number of data points correctly classified by the classifier.
D) The weight assigned to an individual classifier depends upon the weighted sum error of misclassified points for that classifier.

Answer: B, D (The classifier chosen at each stage is the one that minimizes the weighted error at that stage. The weight of a point is high if it has been misclassified more number of times in the previous iterations. Thus, maximum error minimization is performed by trying to correctly predict the points which were misclassified in earlier iterations. Also, weights are assigned to the classifiers depending upon their accuracy which again depends upon the weighted error (for that classifier).

6. Suppose the VC dimension of a hypothesis space is 6. Which of the following are true?

A) At least one set of 6 points can be shattered by the hypothesis space.
B) No sets of 6 points can be shattered by the hypothesis space.
C) All sets of 6 points can be shattered by the hypothesis space.
D) No set of 6 points can be shattered by the hypothesis space.

Answer: A, D (From the definition of VC dimension)

- If there exists at least one subset of X of size d that can be shattered then VC(H) $\geq d$.
- If no subset of size d can be shattered, then VC(H) $< d$.
- From the above facts, options A and D are correct.

7. Ensembles will yield bad results when there is a significant diversity among the models. Write True or False.
   A) True
   B) False

Answer: B
Ensemble is a collection of diverse set of learners to improve the stability and the performance of the algorithm. So, more diverse the models are, the better will be the performance of ensemble.

8. Which of the following algorithms are not an ensemble learning algorithm?
   A) Random Forest
   B) Adaboost
   C) Gradient Boosting
   D) Decision Tress

Answer: D.
Decision trees do not aggregate the results of multiple trees, so it is not an ensemble algorithm.

NPTEL Online Certification Courses

Indian Institute of Technology Kharagpur

Course Name: Introduction to Machine Learning

Assignment – Week 8 (Clustering)

TYPE OF QUESTION: MCQ/MSQ

Number of Question: 7 **Total Marks: 7x2 = 14**

1. For two runs of K-Mean clustering is it expected to get same clustering results?

   A) Yes
   B) No

Answer: (B)

K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.
However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

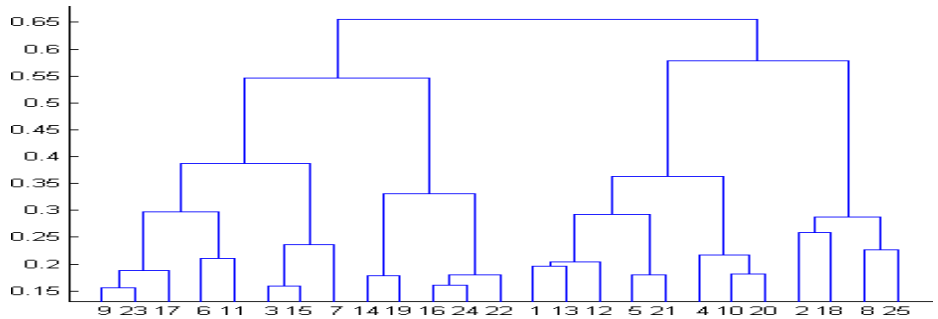
2. Which of the following can act as possible termination conditions in K-Means?

   I. For a fixed number of iterations.
   II. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
   III. Centroids do not change between successive iterations.
   IV. Terminate when RSS falls below a threshold

   A) I, III and IV
   B) I, II and III
   C) I, II and IV
   D) **All of the above**

Answer: D

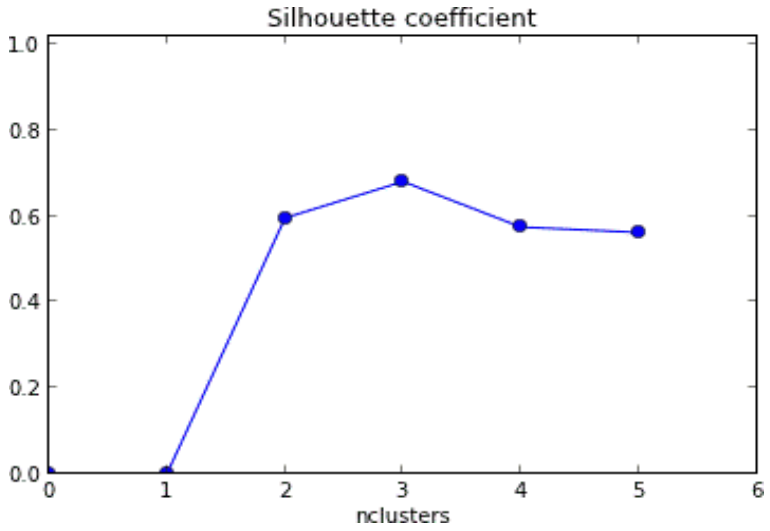
3. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?



A) There were 28 data points in clustering analysis.
B) The best no. of clusters for the analysed data points is 4.
C) The proximity function used is Average-link clustering.
D) The above dendrogram interpretation is not possible for K-Means clustering analysis.

Answer: A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

4. What should be the best choice of no. of clusters based on the following results:



A) 1
B) 2
C) **3**
D) 4

Answer: C
The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters.

5. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

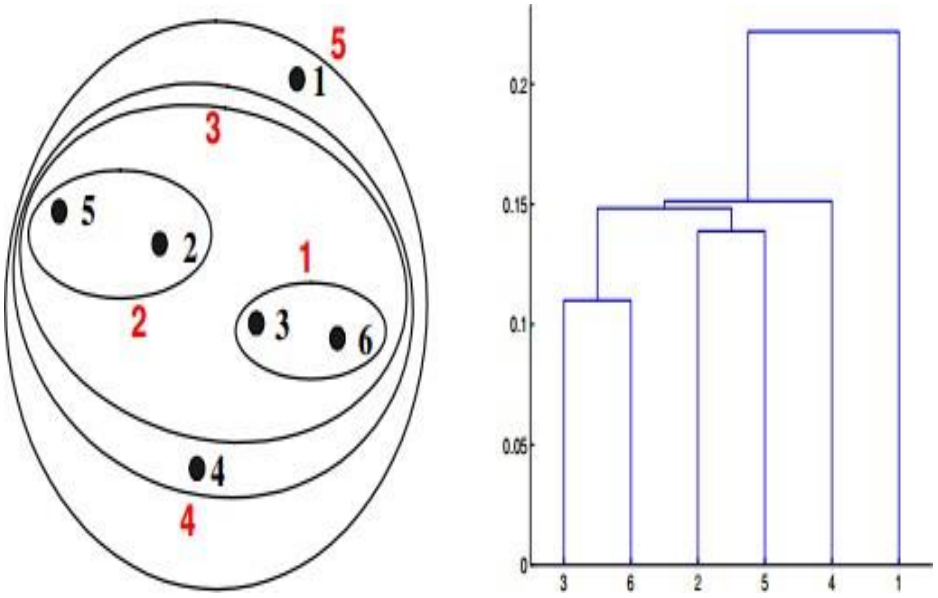
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

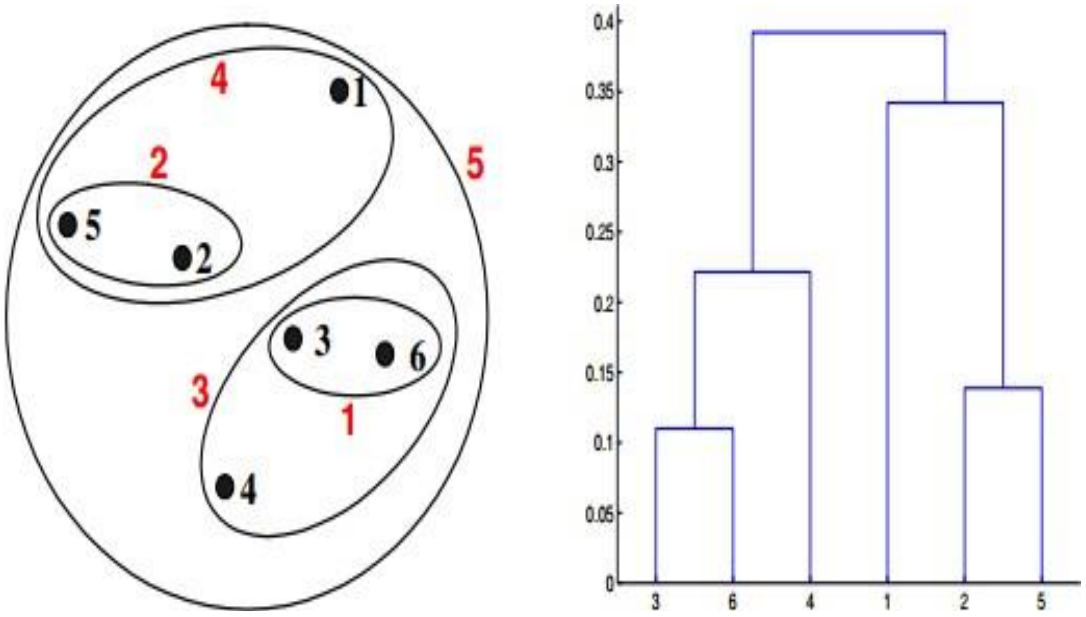
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

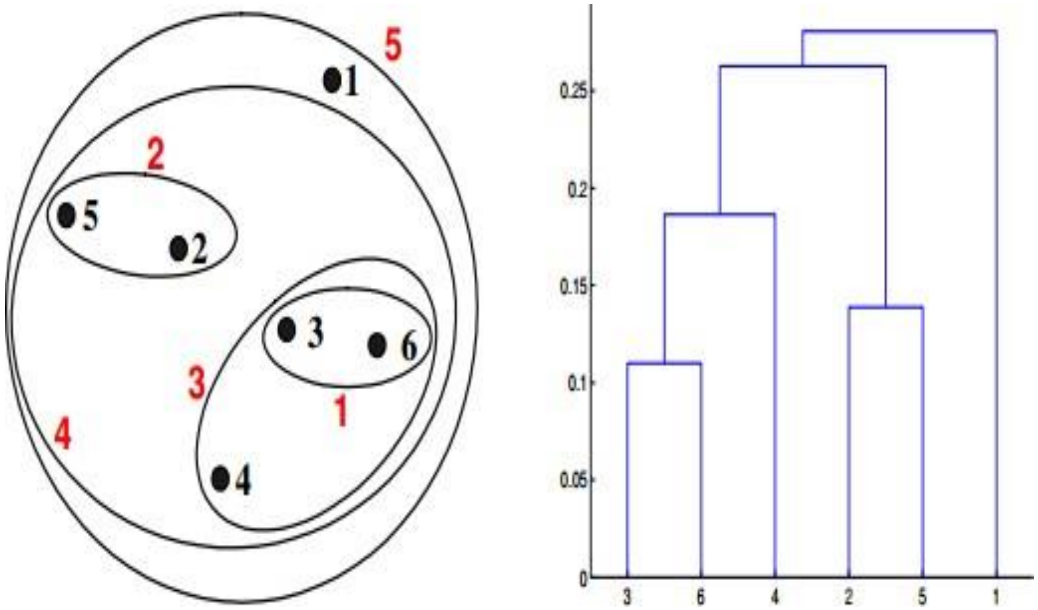
A)



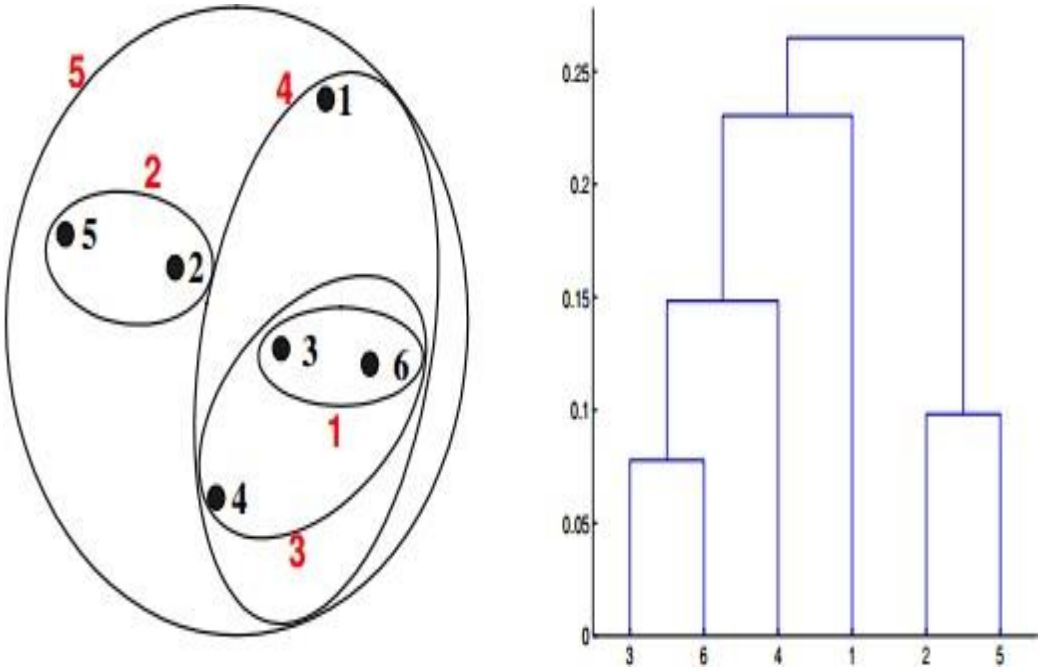
B)


4
1
2
5
5
2
3
6
3
1
4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

C)


5
1
2
5
2
3
6
3
1
4
4
0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

D)



Solution: A)

Answer: For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist ({3, 6}, {2, 5}) = min (dis (3, 2), dist (6, 2), dist (3, 5), dist (6, 5)) = min (0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

6. Which of the following algorithms are most sensitive to outliersWhatWh?

A) K-means clustering
B) K-medians clustering
C) K-modes clustering
D) K -medoids clustering

Answer: A
K-means is the most sensitive because it uses the mean of the cluster data points to find the cluster center.

7. What is the possible reason(s) for producing two different dendograms using agglomerative clustering for the same data set?

A) Proximity function
B) No. of data points
C) Variables used
D) All of these

Answer: E
Change in either of the proximity function, no of variables used and data points will change the dendograms.