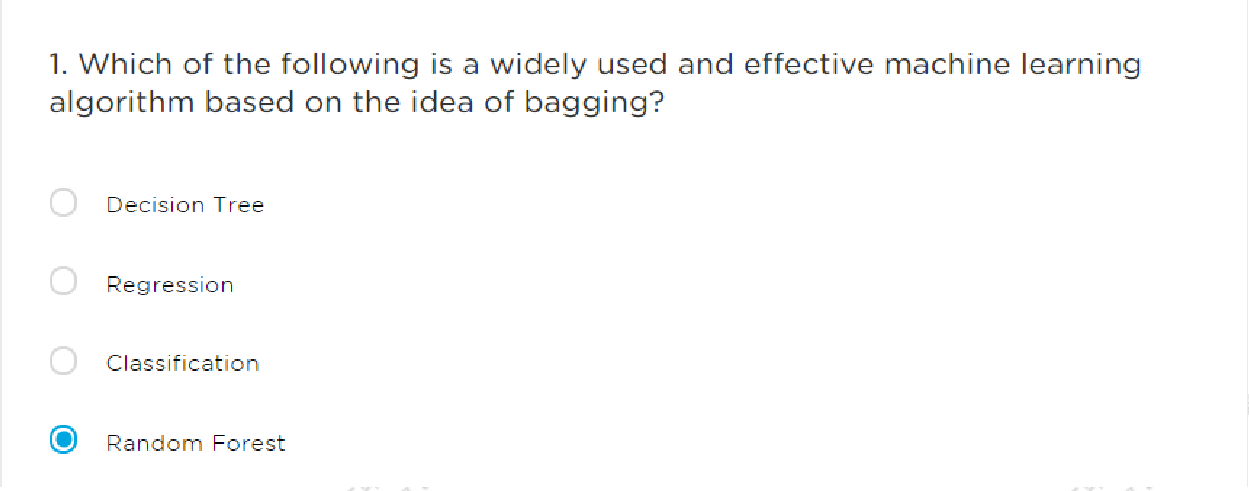


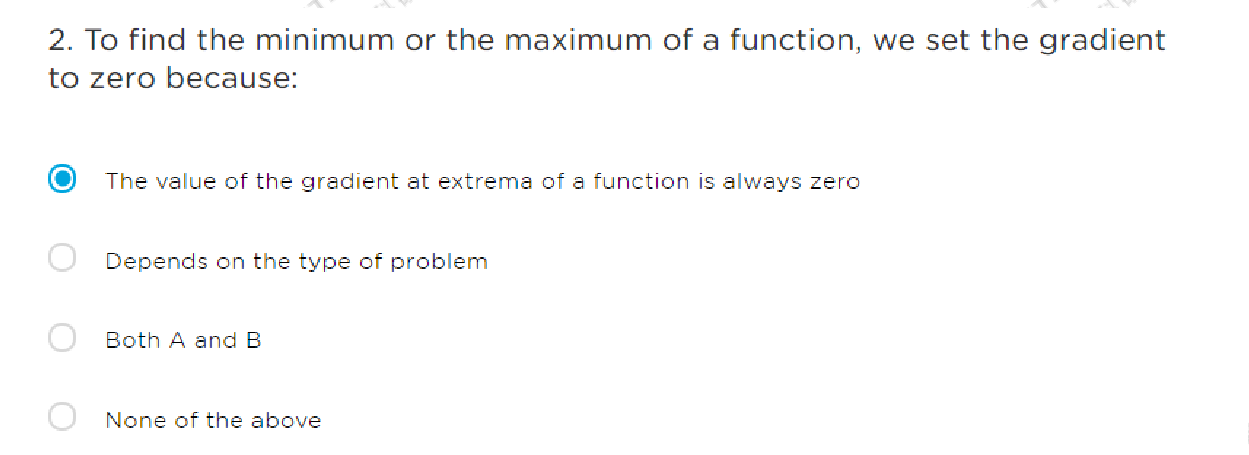
1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

○ Decision Tree

○ Regression

○ Classification

● Random Forest



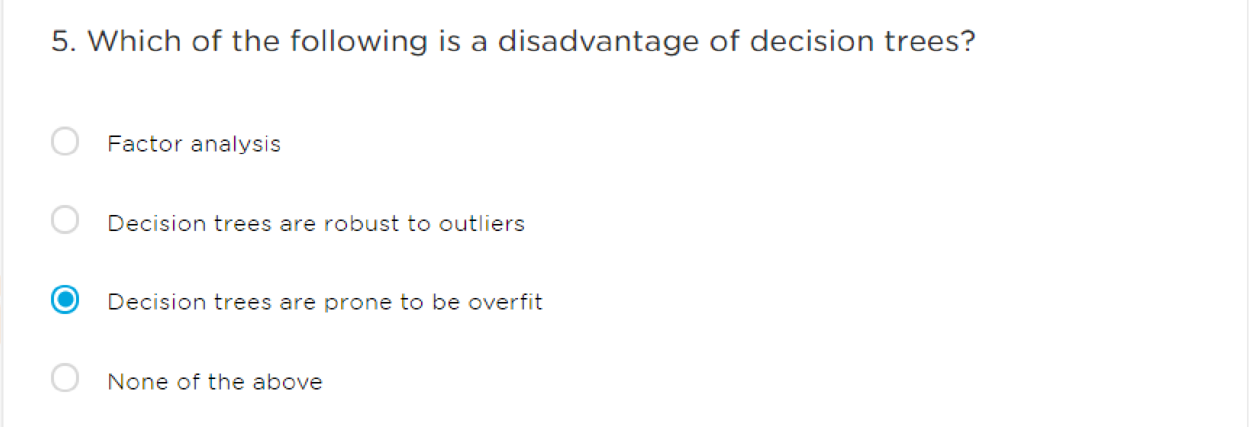
2. To find the minimum or the maximum of a function, we set the gradient to zero because:

● The value of the gradient at extrema of a function is always zero

○ Depends on the type of problem

○ Both A and B

○ None of the above



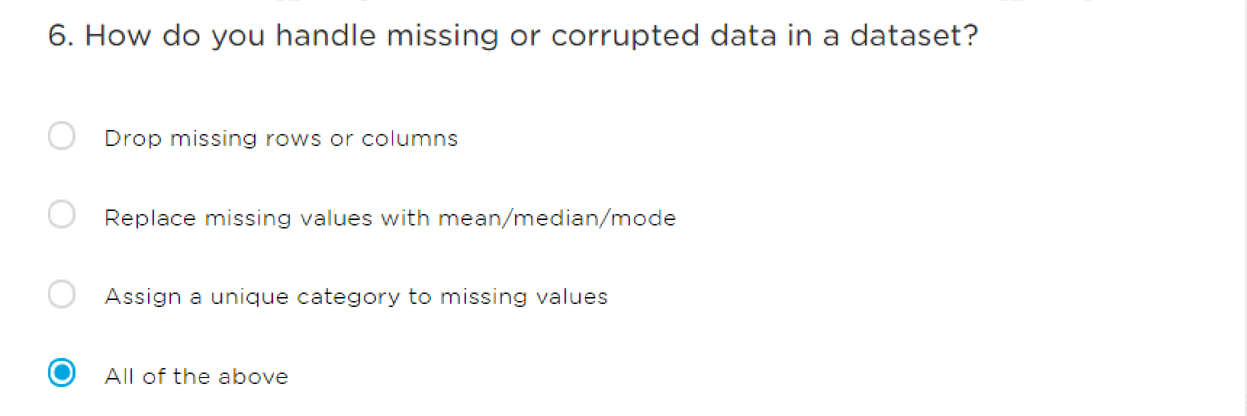
5. Which of the following is a disadvantage of decision trees?

○ Factor analysis

○ Decision trees are robust to outliers

● Decision trees are prone to be overfit

○ None of the above



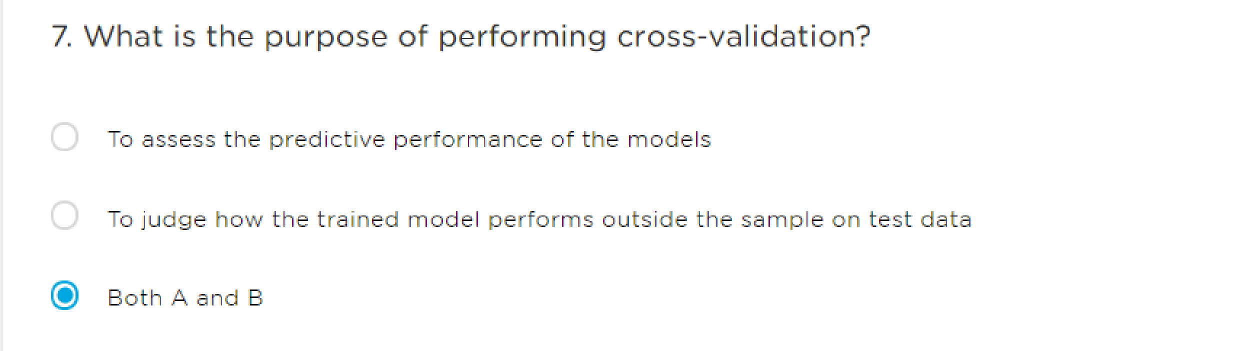
6. How do you handle missing or corrupted data in a dataset?

○ Drop missing rows or columns

○ Replace missing values with mean/median/mode

○ Assign a unique category to missing values

● All of the above

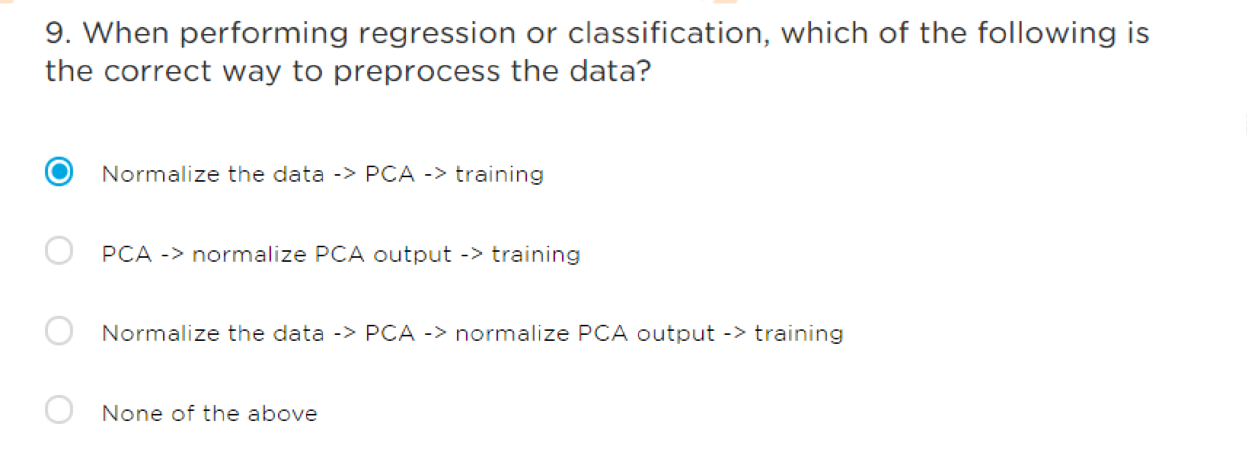


7. What is the purpose of performing cross-validation?

○ To assess the predictive performance of the models

○ To judge how the trained model performs outside the sample on test data

● Both A and B



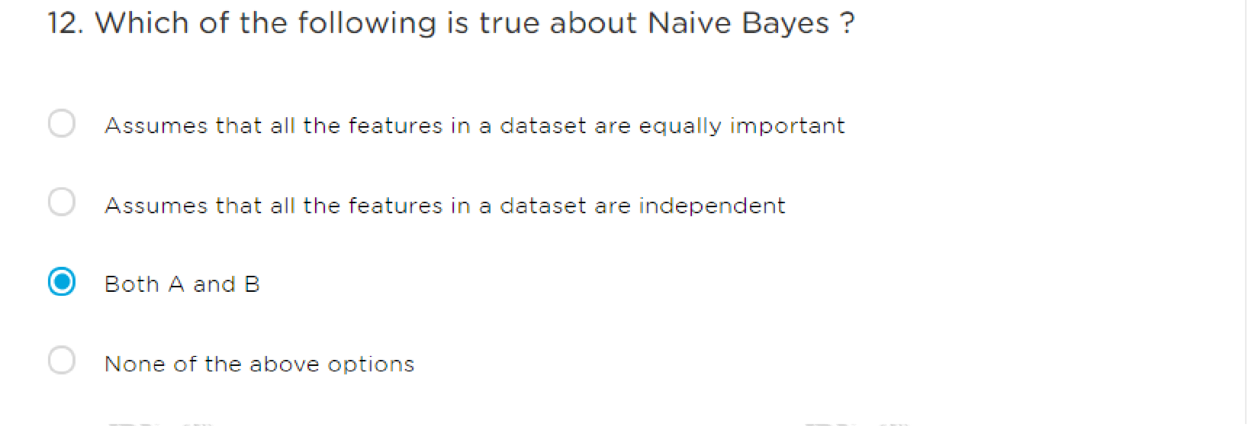
9. When performing regression or classification, which of the following is the correct way to preprocess the data?

● Normalize the data -> PCA -> training

○ PCA -> normalize PCA output -> training

○ Normalize the data -> PCA -> normalize PCA output -> training

○ None of the above



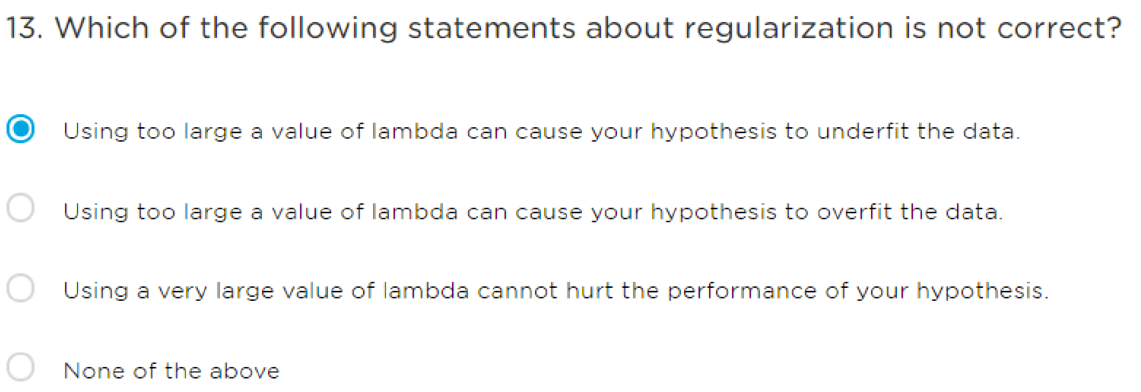
12. Which of the following is true about Naive Bayes ?

○ Assumes that all the features in a dataset are equally important

○ Assumes that all the features in a dataset are independent

● Both A and B

○ None of the above options



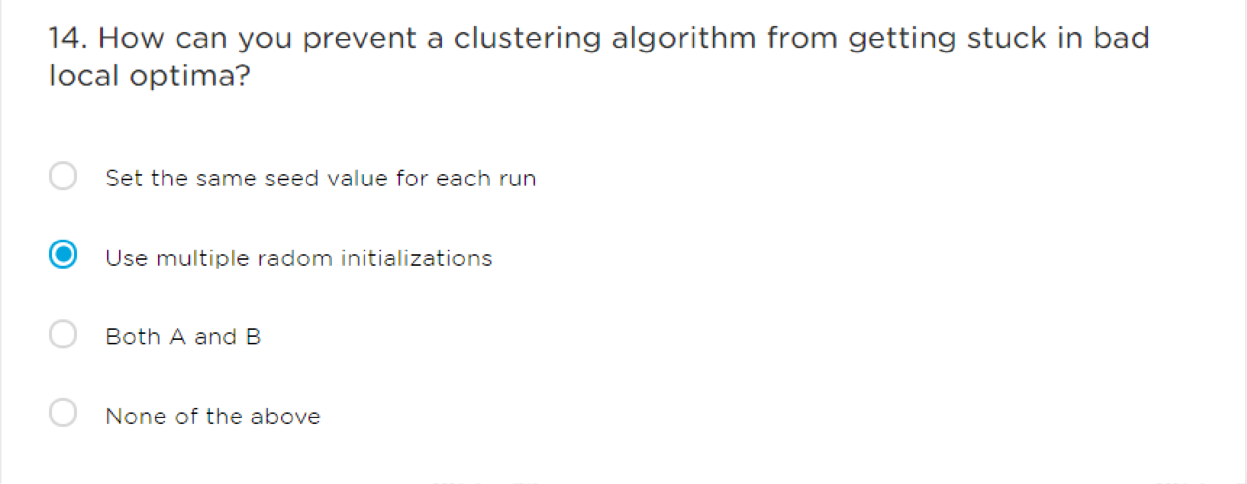
13. Which of the following statements about regularization is not correct?

● Using too large a value of lambda can cause your hypothesis to underfit the data.

○ Using too large a value of lambda can cause your hypothesis to overfit the data.

○ Using a very large value of lambda cannot hurt the performance of your hypothesis.

○ None of the above



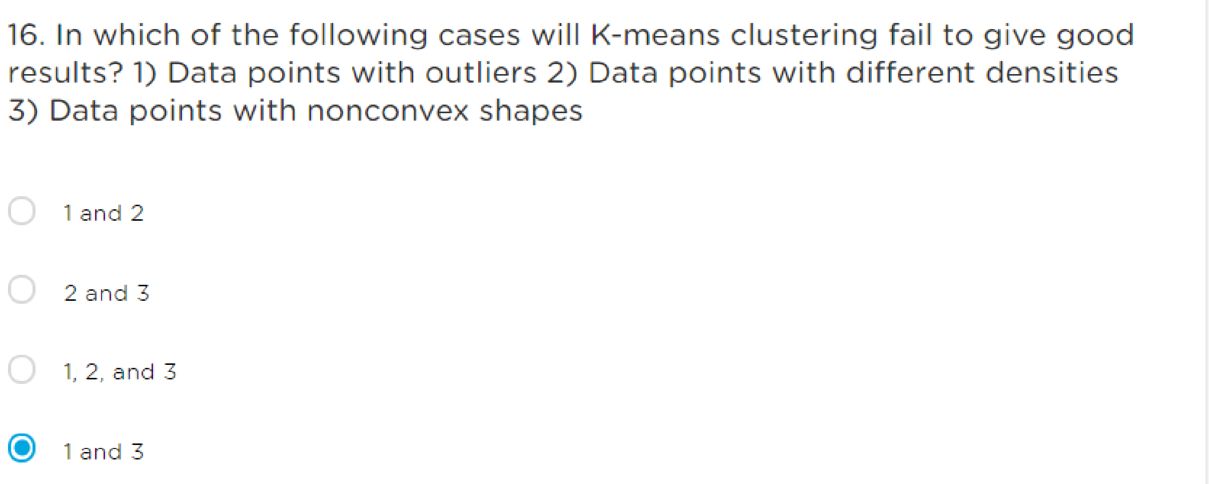
14. How can you prevent a clustering algorithm from getting stuck in bad local optima?

○ Set the same seed value for each run

● Use multiple radom initializations

○ Both A and B

○ None of the above



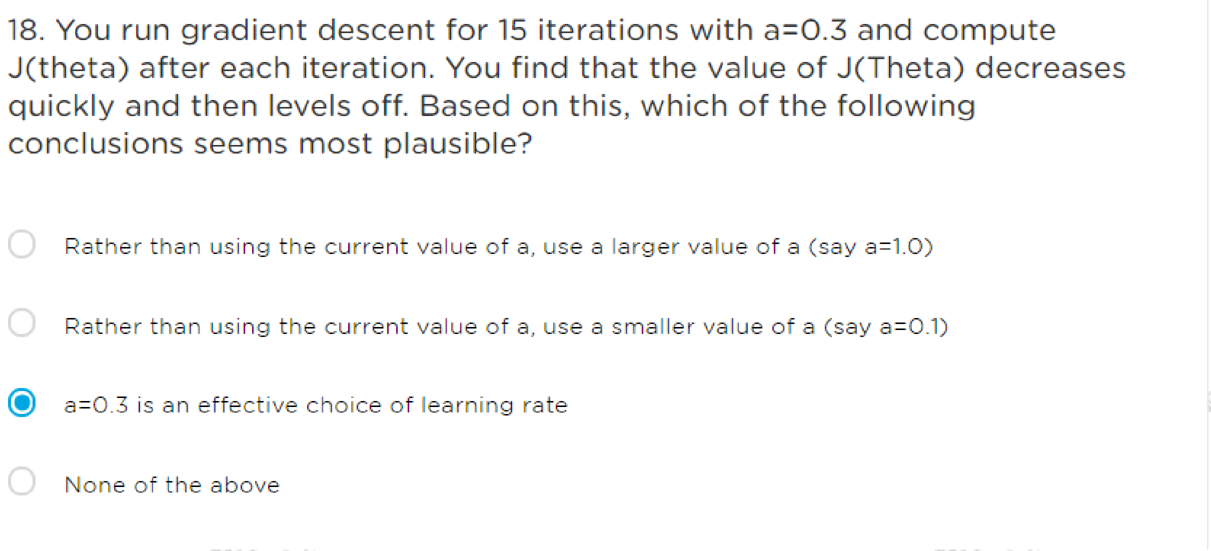
16. In which of the following cases will K-means clustering fail to give good results? 1) Data points with outliers 2) Data points with different densities 3) Data points with nonconvex shapes

○ 1 and 2

○ 2 and 3

○ 1, 2, and 3

● 1 and 3



18. You run gradient descent for 15 iterations with a=0.3 and compute J(theta) after each iteration. You find that the value of J(Theta) decreases quickly and then levels off. Based on this, which of the following conclusions seems most plausible?

○ Rather than using the current value of a, use a larger value of a (say a=1.0)

○ Rather than using the current value of a, use a smaller value of a (say a=0.1)

● a=0.3 is an effective choice of learning rate

○ None of the above

1) Which of the following is/are true about bagging trees?

1. In bagging trees, individual trees are independent of each other
2. Bagging is the method for improving the performance by aggregating the results of weak learners

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

Solution: C

Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

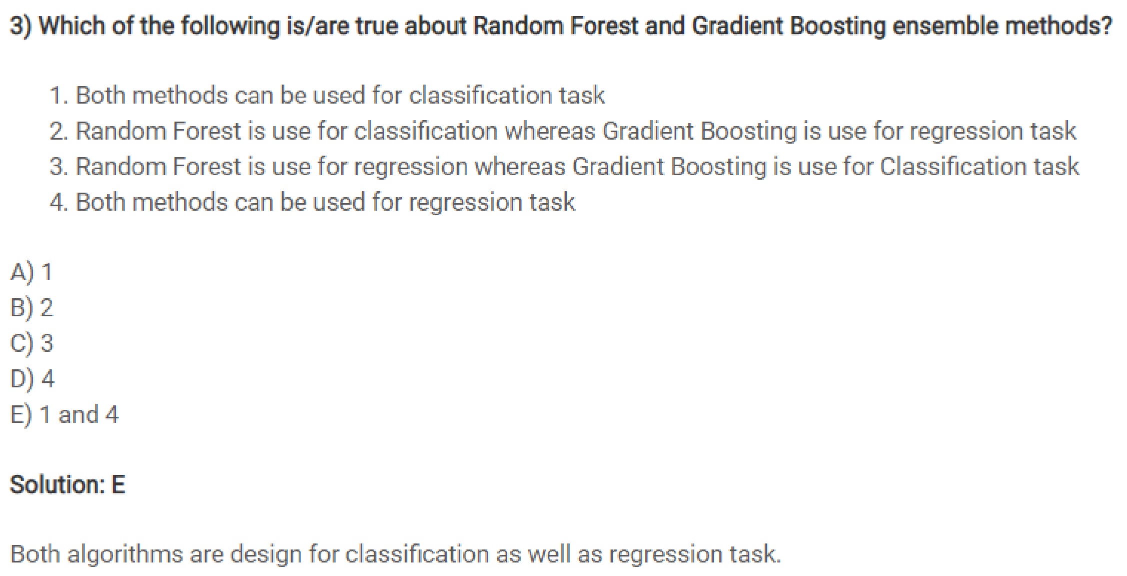
2) Which of the following is/are true about boosting trees?

1. In boosting trees, individual weak learners are independent of each other
2. It is the method for improving the performance by aggregating the results of weak learners

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

Solution: B

In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.



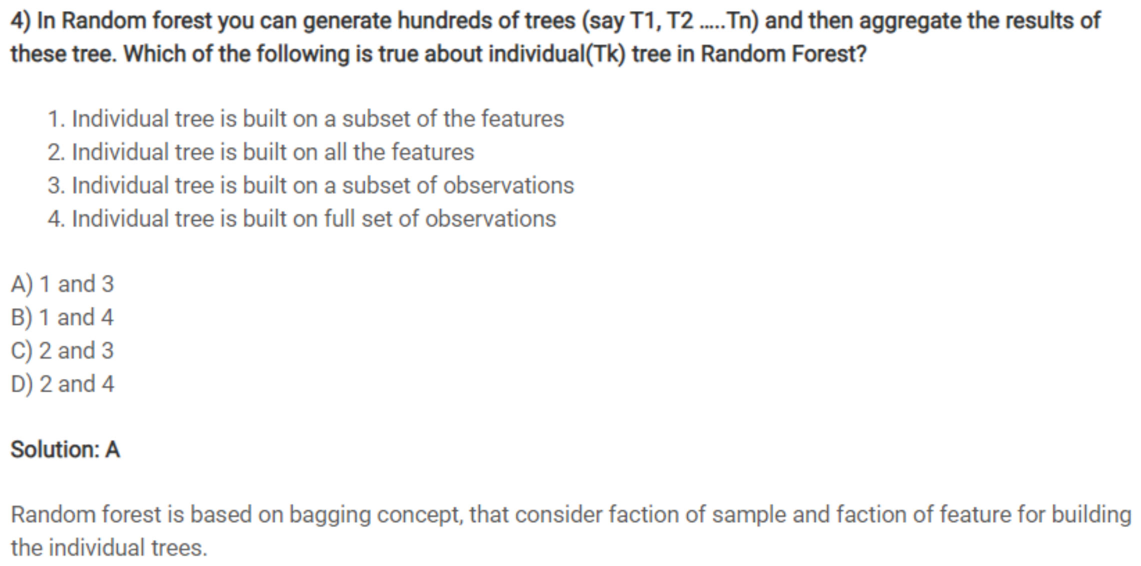
3) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

1. Both methods can be used for classification task
2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
4. Both methods can be used for regression task

A) 1
B) 2
C) 3
D) 4
E) 1 and 4

Solution: E

Both algorithms are design for classification as well as regression task.



4) In Random forest you can generate hundreds of trees (say T1, T2 .....Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

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

Solution: A

Random forest is based on bagging concept, that consider faction of sample and faction of feature for building the individual trees.

6) Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?

1. Gradient Boosting
2. Extra Trees
3. AdaBoost
4. Random Forest

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

Solution: D

Random Forest and Extra Trees don't have learning rate as a hyperparameter.

8) Which of the following is true about training and testing error in such case?

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training T1, T2 … Tn where T1 < T2…. Tn-1 < Tn.

A) The difference between training error and test error increases as number of observations increases
B) The difference between training error and test error decreases as number of observations increases
C) The difference between training error and test error will not change
D) None of These

Solution: B

As we have more and more data, training error increases and testing error de-creases. And they all converge to the true error.

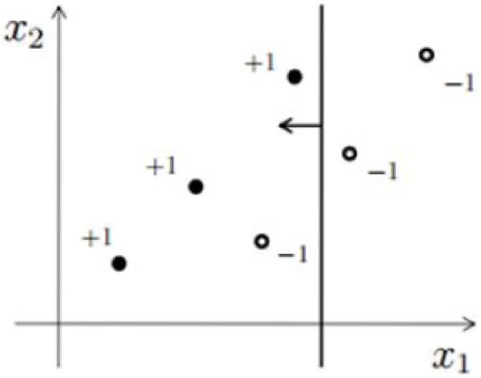
10) Which of the following algorithm are not an example of ensemble learning algorithm?

A) Random Forest
B) Adaboost
C) Extra Trees
D) Gradient Boosting
E) Decision Trees

Solution: E

Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm.

Consider the following figure for answering the next few questions. In the figure, X1 and X2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X1(say splitting point is x11) which is shown in the figure using vertical line. Every value less than x11 will be predicted as positive class and greater than x will be predicted as negative class.



12) How many data points are misclassified in above image?

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

Solution: A

Only one observation is misclassified, one negative class is showing at the left side of vertical line which will be predicting as a positive class.

13) Which of the following splitting point on feature x1 will classify the data correctly?

A) Greater than x11
B) Less than x11
C) Equal to x11
D) None of above

Solution: D

If you search any point on X1 you won't find any point that gives 100% accuracy.

14) If you consider only feature X2 for splitting. Can you now perfectly separate the positive class from negative class for any one split on X2?

A) Yes
B) No

Solution: B

It is also not possible.

15) Now consider only one splitting on both (one on X1 and one on X2) feature. You can split both features at any point. Would you be able to classify all data points correctly?

A) TRUE
B) FALSE

Solution: B

You won’t find such case because you can get minimum 1 misclassification.

1) [True or False] k-NN algorithm does more computation on test time rather than train time.

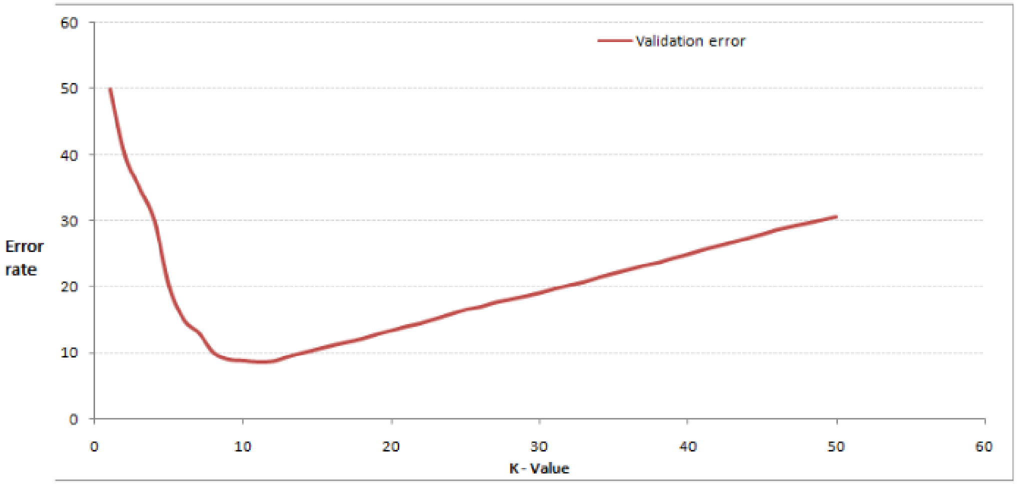
A) TRUE
B) FALSE

Solution: A

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.

2) In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.



A) 3
B) 10
C) 20
D 50

Solution: B

Validation error is the least when the value of k is 10. So it is best to use this value of k

3) Which of the following distance metric can not be used in k-NN?

A) Manhattan
B) Minkowski
C) Tanimoto
D) Jaccard
E) Mahalanobis
F) All can be used

Solution: F

All of these distance metric can be used as a distance metric for k-NN.

4) Which of the following option is true about k-NN algorithm?

A) It can be used for classification
B) It can be used for regression
C) It can be used in both classification and regression

Solution: C

We can also use k-NN for regression problems. In this case the prediction can be based on the mean or the median of the k-most similar instances.

5) Which of the following statement is true about k-NN algorithm?

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

A) 1 and 2
B) 1 and 3
C) Only 1
D) All of the above

Solution: D

The above mentioned statements are assumptions of kNN algorithm

6) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

A) K-NN
B) Linear Regression
C) Logistic Regression

Solution: A

k-NN algorithm can be used for imputing missing value of both categorical and continuous variables.

7) Which of the following is true about Manhattan distance?

A) It can be used for continuous variables
B) It can be used for categorical variables
C) It can be used for categorical as well as continuous
D) None of these

Solution: A

Manhattan Distance is designed for calculating the distance between real valued features.

8) Which of the following distance measure do we use in case of categorical variables in k-NN?

1. Hamming Distance
2. Euclidean Distance
3. Manhattan Distance

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

Solution: A

9) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?

A) 1
B) 2
C) 4
D) 8

Solution: A

sqrt((1-2)^2 + (3-3)^2) = sqrt(1^2 + 0^2) = 1

10) Which of the following will be Manhattan Distance between the two data point A(1,3) and B(2,3)?

A) 1
B) 2
C) 4
D) 8

Solution: A

sqrt(mod((1-2)) + mod((3-3))) = sqrt(1 + 0) = 1