TELANGANA UNIVERSITY
S.S.R. DEGREE COLLEGE, NIZAMABAD (C.C:5029)
IV SEMESTER INTERNAL ASSESSMENT I EXAMINATIONS
DATA SCIENCE (MACHINE LEARNING) QUESTION BANK

---

1. What is Machine learning? [c]
a) The selective acquisition of knowledge through the use of computer programs
b) The selective acquisition of knowledge through the use of manual programs
c) The autonomous acquisition of knowledge through the use of computer programs
d) The autonomous acquisition of knowledge through the use of manual programs

2. K-Nearest Neighbors (KNN) is classified as what type of machine learning algorithm?[a]
a) Instance-based learning
b) Parametric learning
c) Non-parametric learning
d) Model-based learning

3. Which of the following is not a supervised machine learning algorithm? [c]
a) K-means
b) Naïve Bayes
c) SVM for classification problems
d) Decision tree

4. What's the key benefit of using deep learning for tasks like recognizing images? [c]
a) They need less training data than other methods.
b) They're easier to explain and understand than other models.
c) They can learn complex details from the data on their own.
d) They work faster and are more efficient computationally.

5. Which algorithm is best suited for a binary classification problem? [b]
a) K-nearest Neighbors
b) Decision Trees
c) Random Forest
d) Linear Regression

6. What is the key difference between supervised and unsupervised learning? [a]
a) Supervised learning requires labeled data, while unsupervised learning does not.
b) Supervised learning predicts labels, while unsupervised learning discovers patterns.
c) Supervised learning is used for classification, while unsupervised learning is used for regression.
d) Supervised learning is always more accurate than unsupervised learning.

7. Which type of machine learning algorithm falls under the category of "unsupervised learning"?
[b]
a) Linear Regression
b) K-means Clustering
c) Decision Trees
d) Random Forest

8. Which of the following statements is true about AdaBoost? [c]
a) It is particularly prone to overfitting on noisy datasets
b) Complexity of the weak learner is important in AdaBoost
c) It is generally more prone to overfitting
d) It improves classification accuracy

9. Which one of the following models is a generative model used in machine learning? [b]

a) Support vector machines
b) Naïve Bayes
c) Logistic Regression
d) Linear Regression

10. An artificially intelligent car decreases its speed based on its distance from the car in front of it. Which algorithm is used? [a]

a) Naïve-Bayes
b) Decision Tree
c) Linear Regression
d) Logistic Regression

11. Which of the following statements is false about Ensemble learning? [b]

a) It is a supervised learning algorithm
b) It is an unsupervised learning algorithm
c) More random algorithms can be used to produce a stronger ensemble
d) Ensembles can be shown to have more flexibility in the functions they can represent

12. Which of the following statements is true about stochastic gradient descent? [a]

a) It processes one training example per iteration
b) It is not preferred, if the number of training examples is large
c) It processes all the training examples for each iteration of gradient descent
d) It is computationally very expensive, if the number of training examples is large

13. Decision tree uses the inductive learning machine learning approach. [b]
a) False
b) True

14. What elements describe the Candidate-Elimination algorithm? [b]
a) depends on the dataset
b) just a set of candidate hypotheses
c) just a set of instances
d) set of instances, set of candidate hypotheses

15. Which of the following statements is not true about boosting? [a]
a) It mainly increases the bias and the variance
b) It tries to generate complementary base-learners by training the next learner on the mistakes of the previous learners
c) It is a technique for solving two-class classification problems
d) It uses the mechanism of increasing the weights of misclassified data in preceding classifiers

16. Which of the following statements is not true about the Decision tree? [a]
a) It can be applied on binary classification problems only
b) It is a predictor that predicts the label associated with an instance by traveling from a root node of a tree to a leaf
c) At each node, the successor child is chosen on the basis of a splitting of the input space
d) The splitting is based on one of the features or on a predefined set of splitting rules

17. Decision tree uses the inductive learning machine learning approach. [a]
a) True
b) False
18. In a splitting rule at internal nodes of the tree based on thresholding the value of a single feature, it follows that a tree with k leaves can shatter a set of k instances. [b]
a) False
b) True

19. Minimum description length (MDL) principle is used to avoid overfitting in decision trees.[a]
a) True
b) False

20. Suppose in a decision tree, we are making some simplifying assumptions that each instance is a vector of d bits (X = {0, 1}d). Which of the following statements is not true about the above situation [d]
a) It thresholding the value of a single feature corresponds to a splitting rule of the form 1[xi=1] for some i = [d]
b) The hypothesis class becomes finite, but is still very large
c) Any classifier from {0, 1}d to {0, 1} can be represented by a decision tree with 2d leaves and depth of d + 1
d) Any classifier from {0, 1}d to {0, 1} can be represented by a decision tree with 2d+1 leaves and depth of d + 1

21: _____type of learning is characterized by an agent learning through interactions with an environment and receiving rewards?

Answer: Reinforcement learning

22:_____t is the primary goal of feature scaling in machine learning?

Answer: To normalize the feature values to a standard range

23: _____machine learning algorithm is sensitive to the scale of features and requires feature scaling?

Answer: Support Vector Machines (SVM)

24: In a k-fold cross-validation, _____ is the dataset divided?

Answer: It is divided into \(k-1\) training subsets and 1 validation subset

25: _____ technique is used to reduce the impact of noise and outliers in a dataset?

Answer: Regularization

26: _____ algorithm is used for finding the optimal clustering of data points?

Answer: K-Means clustering

27: _____ is the primary purpose of a decision tree's leaf nodes?

Answer: a) To make predictions

28: _____ machine learning approach is based on the assumption that similar data points are more likely to have the same labels?

Answer: a) Clustering

29: In a precision-recall curve, _____ axis represents precision?

Answer: b) Vertical axis

30: _____is the purpose of regularization in linear regression?

Answer: To avoid underfitting

31:_____ is the purpose of the validation set in machine learning?

Answer: To fine-tune hyperparameters

32: _____ type of machine learning algorithm aims to mimic the process of human learning?

Answer: Reinforcement learning

33: _____does the term "overfitting" refer to in machine learning?

Answer: When a model performs well on the training data but poorly on new data

34: _____ machine learning algorithm is suitable for solving regression problems?

Answer: Random Forest

35:b_____ technique is used to reduce the dimensionality of data while preserving as much information as possible?

Answer: Feature extraction

36: _____ is the purpose of the bias term in a linear regression model?

Answer: To shift the regression line up or down

37: _____ algorithm is used for finding frequent itemsets in transactional databases?

Answer: Apriori algorithm

38: In the context of machine learning, _____ is the term "bias-variance trade-off" referring to?

Answer: The trade-off between the complexity of a model and its ability to generalize

39: _____ algorithm is used for hierarchical clustering?

Answer: Agglomerative clustering

40: _____ method can be used to handle missing data in a dataset?

Answer: All of the above

41. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them.

This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables.

42. What is meant by Parametric and Non-parametric Models?

Parametric models refer to the models having a limited number of parameters. In case of parametric models, only the parameter of a model is needed to be known to make predictions regarding the new data.

Non-parametric models do not have any restrictions on the number of parameters, which makes new data predictions more flexible. In case of non-parametric models, the knowledge of model parameters and the state of the data needs to be known to make predictions.

43. Outlier Values can be Discovered from which Tools?

The various tools that can be used to discover outlier values are scatterplots, boxplots, Z-score, etc.

44. What is Support Vector Machine (SVM) in Machine Learning?

SVM is a Machine Learning algorithm that is majorly used for classification. It is used on top of the high dimensionality of the characteristic vector.

45. What is Cross-validation in Machine Learning?

Cross-validation allows a system to increase the performance of the given Machine Learning algorithm, which is fed a number of sample data from the dataset. This sampling process is done to

break the dataset into smaller parts that have the same number of rows, out of which a random part is selected as a test set and the rest of the parts are kept as train sets. Cross-validation consists of the following techniques:

- Holdout method
- K-fold cross-validation
- Stratified k-fold cross-validation
- Leave p-out cross-validation

46. What is Entropy in Machine Learning?

Entropy in Machine Learning measures the randomness in the data that needs to be processed. The more entropy in the given data, the more difficult it becomes to draw any useful conclusion from the data. For example, let us take the flipping of a coin. The result of this act is random as it does not favor heads or tails. Here, the result for any number of tosses cannot be predicted easily as there is no definite relationship between the action of flipping and the possible outcomes.

47. What is Epoch in Machine Learning?

Epoch in Machine Learning is used to indicate the count of passes in a given training dataset where the Machine Learning algorithm has done its job. Generally, when there is a large chunk of data, it is grouped into several batches. All these batches go through the given model, and this process is referred to as iteration. Now, if the batch size comprises the complete training dataset, then the count of iterations is the same as that of epochs.
In case there is more than one batch, d*e=i*b is the formula used, wherein d is the dataset, e is the number of epochs, i is the number of iterations, and b is the batch size.

48. What are Type I and Type II Errors?

Type I Error: Type I Error, false positive, is an error where the outcome of a test shows the nonacceptance of a true condition.

For example, suppose a person gets diagnosed with depression even when they are not suffering from the same, it is a case of false positive.

Type II Error: Type II Error, false negative, is an error where the outcome of a test shows the acceptance of a false condition.

For example, the CT scan of a person shows that they do not have a disease but in fact they do have the disease. Here, the test accepts the false condition that the person does not have the disease. This is a case of false negative.

49. How to handle Missing or Corrupted Data in a Dataset?

In Python pandas, there are two methods to locate lost or corrupted data and discard those values:

- isNull(): It can be used for detecting the missing values.
- dropna(): It can be used for removing columns or rows with null values.

fillna() can be used to fill the void values with placeholder values.

50. When to use mean and when to use median to handle a missing numeric value?

We choose the mean to impute missing values when the data distribution is normal and there are no significant outliers, as the mean is sensitive to both. In contrast, we use the median in cases of skewed distributions or when outliers are present, because the median is more robust to these factors and provides a better central tendency measure under these conditions.