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

---

1. Which of the following statements is false about k-Nearest Neighbor algorithm? [c]
a) It stores all available cases and classifies new cases based on a similarity measure
b) It has been used in statistical estimation and pattern recognition
c) It cannot be used for regression
d) The input consists of the k closest training examples in the feature space

2. Which of the following statements is not true about k-Nearest Neighbor classification?[d]
a) The output is a class membership
b) An object is classified by a plurality vote of its neighbors
c) If k = 1, then the object is simply assigned to the class of that single nearest neighbor
d) The output is the property value for the object

3. Suppose k = 3 and the data point A's 3-nearest-neighbours from the dataset are instances X, Y and Z. The table shows their classes and the distances computed. Then A's predicted class using majority voting will be 'Good'? [b]

Neighbor	Class	Distance
X	Good	0.2
Y	Bad	0.3
Z	Bad	0.5

advertisement
a) True
b) False

4. We have data from a survey and objective testing with two attributes A and B to classify whether a special paper tissue is good or not. Here are four training samples given in the table. Now the factory produces a new paper tissue that pass laboratory test with A = 3 and B = 7. If K = 3, then 'Good' is the classification of this new tissue? [a]

A	B	C = Classification
7	6	Bad
7	4	Bad
4	4	Good
2	4	Good

a) True
b) False

5. Suppose k = 3 and the data point A's 3-nearest-neighbours from the dataset are instances X, Y and Z. The table shows their classes and the distances computed. Then A's predicted class using inverse distance weighted voting will be 'Good'? [a]

Neighbor	Class	Distance
X	Good	0.1
Y	Bad	0.3
Z	Bad	0.5

a) True
b) False

6. Which of the following statements is not true about k Nearest Neighbor? [d]
a)It belongs to the supervised learning domain
b)It has an application in data mining and intrusion detection
c)It is Non-parametric
d) It is not an instance based learning algorithm

7. Which of the following statements is not supporting in defining k Nearest Neighbor as a lazy learning algorithm? [c]
a) It defers data processing until it receives a request to classify unlabeled data
b) It replies to a request for information by combining its stored training data
c) It stores all the intermediate results
d) It discards the constructed answer

8. Which of the following statements is not supporting kNN to be a lazy learner? [c]
a) When it gets the training data, it does not learn and make a model
b) When it gets the training data, it just stores the data
c) It derives a discriminative function from the training data
d) It uses the training data when it actually needs to do some prediction

9. Euclidian distance and Manhattan distance are the same in kNN algorithm to calculate the distance. [b]
a) True
b) False

10. What is the Manhattan distance between a data point (9, 7) and a new query instance (3, 4)? [b]
a) 7
b) 9
c) 3
d) 4

11. Naïve Bayes classifier algorithms are mainly used in text classification. [a]
a) True
b) False

12. What is the formula for Bayes' theorem? Where (A & B) and (H & E) are events and P(B), P(H) & P(E) ≠ 0. [d]
a) P(H|E) = [P(E|H) * P(E)] / P(H)
b) P(A|B) = [P(A|B) * P(A)] / P(B)
c) P(H|E) = [P(H|E) * P(H)] / P(E)
d) P(A|B) = [P(B|A) * P(A)] / P(B)

13. Which of the following statement is not true about Naïve Bayes classifier algorithm? [a]
a) It cannot be used for Binary as well as multi-class classifications
b) It is the most popular choice for text classification problems
c) It performs well in Multi-class prediction as compared to other algorithms
d) It is one of the fast and easy machine learning algorithms to predict a class of test datasets

14. What is the assumptions of Naïve Bayesian classifier? [c]
a) It assumes that features of a data are completely dependent on each other
b) It assumes that each input variable is dependent and the model is not generative
c) It assumes that each input attributes are independent of each other and the model is generative
d) It assumes that the data dimensions are dependent and the model is generative

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

16. Which one of the following terms is not used in the Bayes' Theorem? [b]
a) Prior
b) Unlikelihood
c) Posterior
d) Evidence

17. Is the assumption of the Naïve Bayes algorithm a limitation to use it? [a]
a) True
b) False

18. In which of the following case the Naïve Bayes' algorithm does not work well? [c]
a) When faster prediction is required
b) When the Naïve assumption holds true
c) When there is the case of Zero Frequency
d) When there is a multiclass prediction

19. There are two boxes. The first box contains 3 white and 2 red balls whereas the second contains 5 white and 4 red balls. A ball is drawn at random from one of the two boxes and is found to be white. Find the probability that the ball was drawn from the second box? [b]
a) 53/50
b) 50/104
c) 54/104
d) 54/44

20. Which one of the following models is a generative model used in machine learning? [c]
a) Linear Regression
b) Logistic Regression
c) Naïve Bayes
d) Support vector machines
2 1:_____ is the main goal of machine learning?

Answer: To enable computers to learn from data

2 2: In machine learning, _____ is a model?

Answer: A representation of the data

23: _____ of the following is a supervised learning task?

Answer: d) Classification

2 4: _____ evaluation metric is commonly used for classification tasks when class imbalance is present?
Answer: c F1-score
25:_____ is the purpose of the validation set in machine learning?
Answer: To fine-tune hyperparameters
2 6: _____ type of machine learning algorithm aims to mimic the process of human learning?
Answer: Reinforcement learning
27: _____does the term "overfitting" refer to in machine learning?
Answer: When a model performs well on the training data but poorly on new data
28: _____ machine learning algorithm is suitable for solving regression problems?

Answer: Random Forest

29:b_____ technique is used to reduce the dimensionality of data while preserving as much information as possible?
Answer: Feature extraction
30: _____ is the purpose of the bias term in a linear regression model?
Answer: To shift the regression line up or down
31: _____ algorithm is used for finding frequent itemsets in transactional databases?

Answer: Apriori algorithm

32: 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
33: _____ algorithm is used for hierarchical clustering?

Answer: Agglomerative clustering

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

Answer: All of the above

35: In a neural network, _____ are the layers between the input and output layers called?

Answer: Hidden layers

36:_____ optimization algorithm is commonly used to update the weights of neural networks during training?

Answer: Gradient Descent

37:_____ technique is used to combat the vanishing gradient problem in deep neural networks?

Answer:ReLU activation function

38: _____ machine learning algorithm is inspired by the functioning of the human brain's neural

Answer: Artificial Neural Networks

39: _____ensemble learning method combines multiple weak learners to create a strong learner?

Answer:Gradient Boosting

40: In the context of Support Vector Machines (SVM), _____ is the "kernel trick"?

Answer: A way to implicitly map data to higher-dimensional spaces

41. What is KNN Imputer?
We generally impute null values by the descriptive statistical measures of the data like mean, mode, or median but KNN Imputer is a more sophisticated method to fill the null values. A distance parameter is also used in this method which is also known as the k parameter. The work is somehow similar to the clustering algorithm. The missing value is imputed in reference to the neighborhood points of the missing values.
42. Explain the working procedure of the XGB model.
XGB model is an example of the ensemble technique of machine learning in this method weights are optimized in a sequential manner by passing them to the decision trees. After each pass, the weights become better and better as each tree tries to optimize the weights, and finally, we obtain the best weights for the problem at hand. Techniques like regularized gradient and mini-batch gradient descent have been used to implement this algorithm so, that it works in a very fast and optimized manner.
43. What is the purpose of splitting a given dataset into training and validation data?
The main purpose is to keep some data left over on which the model has not been trained so, that we can evaluate the performance of our machine learning model after training. Also, sometimes we use the validation dataset to choose among the multiple state-of-the-art machine learning models. Like we first train some models let’s say LogisticRegression, XGBoost, or any other than test their performance using validation data and choose the model which has less difference between the validation and the training accuracy.
44. Explain some methods to handle missing values in that data.
Some of the methods to handle missing values are as follows:

- Removing the rows with null values may lead to the loss of some important information.

- Removing the column having null values if it has very less valuable information. it may lead to the loss of some important information.
- Imputing null values with descriptive statistical measures like mean, mode, and median.
- Using methods like KNN Imputer to impute the null values in a more sophisticated way.

45. What is the difference between k-means and the KNN algorithm?

k-means algorithm is one of the popular unsupervised machine learning algorithms which is used for clustering purposes. But the KNN is a model which is generally used for the classification task and is a supervised machine learning algorithm. The k-means algorithm helps us to label the data by forming clusters within the dataset.

46. What is Linear Discriminant Analysis?

LDA is a supervised machine learning dimensionality reduction technique because it uses target variables also for dimensionality reduction. It is commonly used for classification problems.

The LDA mainly works on two objectives:

- Maximize the distance between the means of the two classes.
- Minimize the variation within each class.

46. How can we visualize high-dimensional data in 2-d?

One of the most common and effective methods is by using the t-SNE algorithm which is a short form for t-Distributed Stochastic Neighbor Embedding. This algorithm uses some non-linear complex methods to reduce the dimensionality of the given data. We can also use PCA or LDA to convert n-dimensional data to 2 – dimensional so, that we can plot it to get visuals for better analysis. But the difference between the PCA and t-SNE is that the former tries to preserve the variance of the dataset but the t-SNE tries to preserve the local similarities in the dataset.

47. What is the reason behind the curse of dimensionality?

As the dimensionality of the input data increases the amount of data required to generalize or learn the patterns present in the data increases. For the model, it becomes difficult to identify the pattern for every feature from the limited number of datasets or we can say that the weights are not optimized properly due to the high dimensionality of the data and the limited number of examples used to train the model. Due to this after a certain threshold for the dimensionality of the input data, we have to face the curse of dimensionality.

48. Whether the metric MAE or MSE or RMSE is more robust to the outliers.

Out of the above three metrics, MAE is robust to the outliers as compared to the MSE or RMSE. The main reason behind this is because of Squaring the error values. In the case of an outlier, the error value is already high and then we squared it which results in an explosion in the error values more than expected and creates misleading results for the gradient.

59. Why removing highly correlated features are considered a good practice?

When two features are highly correlated, they may provide similar information to the model, which may cause overfitting. If there are highly correlated features in the dataset then they unnecessarily increase the dimensionality of the feature space and sometimes create the problem of the curse of dimensionality. If the dimensionality of the feature space is high then the model training may take more time than expected, it will increase the complexity of the model and chances of error. This somehow also helps us to achieve data compression as the features have been removed without much loss of data.

50. What is the difference between the content-based and collaborative filtering algorithms of recommendation systems?

In a content-based recommendation system, similarities in the content and services are evaluated, and then by using these similarity measures from past data we recommend products to the user. But on the other hand in collaborative filtering, we recommend content and services based on the preferences of similar users. For example, if one user has taken A and B services in past and a new user has taken service A then service A will be recommended to him based on the other user’s preferences.