Introduction to Machine Learning -IITKGP

Assignment - 1

TYPE OF QUESTION: MCQ/MSQ

Number of questions: 15 Total mark: 2 * 15 = 30

---

1. Which of the following is/are 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. Predict the number of copies of a book that will be sold this month

   Correct Answers: a, c

   Explanation: In (c), the amount of rainfall is a continuous variable. But, we are predicting whether there will be abnormally heavy rainfall next year or not. So it is a Classification task. Similarly, the number of classes in gender identification (a) is discrete. So, it's a classification task. The output variable is a continuous class in other options, so these are regression tasks.

---

2. A feature F1 can take certain values: A, B, C, D, E, F, and represents the grade of students from a college. Which of the following statement is true in the following case?
   a. Feature F1 is an example of a nominal variable.
   b. Feature F1 is an example of an ordinal variable.
   c. It doesn't belong to any of the above categories.
   d. Both of these

   Correct Answer: b

   Explanation: Ordinal variables are the variables that have some order in their categories. For example, grade A should be considered a high grade than grade B.

---

3. Suppose I have 10,000 emails in my mailbox out of which 200 are spams. The spam detection system detects 150 emails as spams, out of which 50 are actually spam. What is the precision and recall of my spam detection system?

   a. Precision = 33.333%, Recall = 25%
   b. Precision = 25%, Recall = 33.33%
   c. Precision = 33.33%, Recall = 75%
   d. Precision = 75%, Recall = 33.33%

   Correct Answer: a

   Explanation:

   We know that,

   Precision = $\frac{Tp}{Tp+Fp}$

   $= \frac{50}{150}$

   $= 33.333\%$

   Recall = $\frac{Tp}{Tp+Fn}$

   $= \frac{50}{200}$

   $= 25\%$

---

4. Which of the following statements describes what is most likely TRUE when the amount of training data increases?

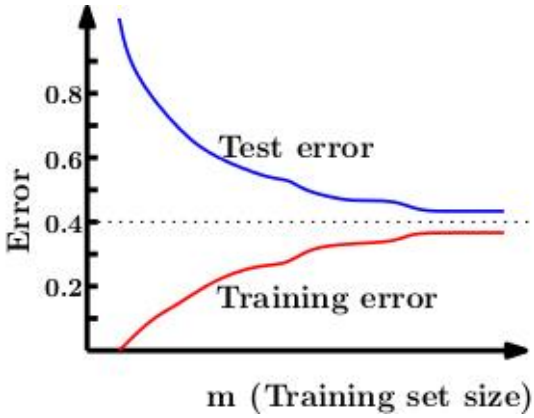
   a. Training error usually decreases and generalization error usually increases.
   b. Training error usually decreases and generalization error usually decreases.
   c. Training error usually increases and generalization error usually decreases.
   d. Training error usually increases and generalization error usually increases.

Correct Answer: a

Explanation: 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.

---

5. You trained a learning algorithm, and plot the learning curve. The following figure is obtained.



The algorithm is suffering from

a. High bias
b. High variance
c. Neither

Correct Answer: a

Explanation: In the plot, the training error is increased with the training set size. The true error is around 0.4 which is quite high. Thus, we can say that the bias is high.

---

6. I am the marketing consultant of a leading e-commerce website. I have been given a task of making a system that recommends products to users based on their activity on Facebook. I realize that user interests could be highly variable. Hence, I decide to

T1) Cluster the users into communities of like-minded people and

T2) Train separate models for each community to predict which product category (e.g., electronic gadgets, cosmetics, etc.) would be the most relevant to that community.

The task T1 is a/an _____ learning problem and T2 is a/an _____ problem.
Choose from the options:
a. Supervised and unsupervised
b. Unsupervised and supervised
c. Supervised and supervised
d. Unsupervised and unsupervised

Correct Answer: b

Explanation: From the definition of supervised and unsupervised learning

7. Select the correct equations.

TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative

i. Precision = $\frac{Tp}{Tp+Fp}$

ii. Recall = $\frac{Fp}{Tp+Fp}$

iii. Recall = $\frac{Tp}{Tp+Fn}$

iv. Accuracy= $\frac{Tp+Fn}{Tp+Fp+Tn+Fn}$

a. i, iii, iv
b. i and iii
c. ii and iv
d. i, ii, iii, iv

Correct Answer: a

Explanation: From the definition of Precision, Recall, and Accuracy

---

8. Which of the following tasks is NOT a suitable machine learning task(s)?
   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

Explanation: Finding the shortest path between a pair of nodes in a graph is not a suitable machine-learning task because it falls under the category of graph algorithms and can be efficiently solved using algorithms like Dijkstra's algorithm. Machine learning is typically used for tasks that involve pattern recognition, prediction, or classification based on data. In this case, the task of finding the shortest path in a graph is better suited for algorithmic or graph theory-based approaches rather than machine learning.

---

9. Which of the following is/are associated with overfitting in machine learning?

   a. High bias
   b. Low bias
   c. Low variance
   d. High variance
   e. Good performance on training data
   f. Poor performance on test data

Correct Answers: b, d, e, f

Explanation: Overfitting is characterized by good performance on the training data, as the model has essentially memorized the data. However, it leads to poor performance on the test data because the model fails to generalize well. Overfitting is associated with low bias and high variance, meaning the model is sensitive to noise or fluctuations in the training data.

---

10. Which of the following statements about cross-validation in machine learning is/are true?
    a. Cross-validation is used to evaluate a model's performance on the training data.
    b. Cross-validation guarantees that a model will generalize well to unseen data.
    c. Cross-validation is only applicable to classification problems and not regression problems.
    d. Cross-validation helps in estimating the model's performance on unseen data by simulating the test phase.

Correct Answer: d

Explanation: Cross-validation is a technique used in machine learning to assess the performance and generalization ability of a model. It involves dividing the available labeled data into multiple subsets or folds. The model is trained on a portion of the data (training set) and evaluated on the remaining portion (validation or test set). By repeating this process with different partitions of the data, cross-validation provides an estimate of the model's performance on unseen data.

---

11. What does k-fold cross-validation involve in machine learning?
    a. Splitting the dataset into k equal-sized training and test sets.
    b. Splitting the dataset into k unequal-sized training and test sets.
    c. Partitioning the dataset into k subsets, and iteratively using each subset as a validation set while the remaining k-1 subsets are used for training.
    d. Dividing the dataset into k subsets, where each subset represents a unique class label for classification tasks.

Correct Answer: c

Explanation: K-fold cross-validation involves dividing the dataset into k subsets or folds. The process then iterates k times, where each time, one of the k subsets is used as the validation set, while the remaining k-1 subsets are used for training the model. This ensures that each subset is used as the validation set exactly once, and the model is trained and evaluated k times, with each fold serving as the validation set once.

---

12. What does the term "feature space" refer to in machine learning?
    a. The space where the machine learning model is trained.
    b. The space where the machine learning model is deployed.
    c. The space which is formed by the input variables used in a machine learning model.
    d. The space where the output predictions are made by a machine learning model.

Correct Answer: c

Explanation: The feature space in machine learning refers to the space formed by the input variables or features used in a model. It represents the space where the data points reside.

---

13. Which of the following statements is/are true regarding supervised and unsupervised learning?
    a. Supervised learning can handle both labeled and unlabeled data.
    b. Unsupervised learning requires human experts to label the data.
    c. Supervised learning can be used for regression and classification tasks.
    d. Unsupervised learning aims to find hidden patterns in the data.

Correct Answers: c, d

Explanation:

Option “a” is incorrect. Supervised learning specifically requires labeled data, while unsupervised learning deals with unlabeled data.

Option “b” is incorrect. Unsupervised learning does not require human experts to label the data; it learns from the raw, unlabeled data.

Option “c” is correct. Supervised learning encompasses both regression, where the output variable is continuous, and classification, where the output variable is categorical.

Option “d” is correct. Unsupervised learning aims to find hidden patterns, structures, or relationships within the data without any prior knowledge of the output labels.

---

14. One of the ways to mitigate overfitting is
    a. By increasing the model complexity
    b. By reducing the amount of training data
    c. By adding more features to the model
    d. By decreasing the model complexity

Correct Answer: d

Explanation: Overfitting occurs when a machine learning model performs well on the training data but fails to generalize to new, unseen data. It usually happens when the model becomes too complex and starts to memorize the training examples instead of learning the underlying patterns. To mitigate overfitting, one of the effective approaches is to decrease the model complexity.

15. How many Boolean functions are possible with N features?

a. $\left(2^{2^N}\right)$
b. (2^N)
c. (N^2)
d. (4^N)

Correct Answer: a

Explanation: Any variable 'A' can have 2 values (i.e., 0 or 1)

For 'N' variables there are 2^N entries in the truth table.

And each output of any particular row in the truth table can be 0 or 1.

Hence, we have (2^{2^N}) different Boolean functions with N variables.