Introduction to Machine Learning -IITKGP

Assignment - 7

TYPE OF QUESTION: MCQ/MSQ

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

<u>Do Answer Questions 1-2 with the given data:</u>

1. Suppose you consider a concept as the conjunction of boolean literals. You are given the instance space X below:

Color	Shape	Size	Texture	Category
Red	Round	Big	Hard	Yes
Red	Round	Big	Hard	Yes
Blue	Flat	Small	Hard	No
Red	Round	Big	Soft	Yes
Green	Flat	Small	Hard	No

‘?’ means don’t care and ‘0’ indicates no value

Find the most specific concept using Find-S algorithm.

a. <Red, Round, Big, ?>
b. <Red, 0, Big, 0>
c. <Red, Round, ?, Soft>
d. <Red, ?, Big, ?>

Correct Answer: a

Detailed Solution:

We'll initialize the most specific hypothesis with the most general description, which is represented by the concept <0, 0, 0, 0>. Then, we'll iterate through the positive examples and update the hypothesis to become more specific while still covering all positive examples.

Initialize the most specific hypothesis:
hypothesis = <0, 0, 0, 0>

Iterate through the positive examples:
Example 1: <Red, Round, Big, Hard, Yes>
Since this is a positive example, we update the hypothesis to match the example:
hypothesis = <Red, Round, Big, Hard>

Example 2: <Red, Round, Big, Hard, Yes>
Since this example matches the current hypothesis, no change is needed.

Example 3: <Blue, Flat, Small, Hard, No>
This is a negative example, so it doesn't affect the hypothesis.

Example 4: <Red, Round, Big, Soft, Yes>
Since this example matches the current hypothesis except for "Soft," we update the hypothesis for that attribute:
hypothesis = <Red, Round, Big, ?>

Example 5: <Green, Flat, Small, Hard, No>
This is a negative example, so it doesn't affect the hypothesis.

After iterating through all positive examples, the most specific hypothesis that fits all positive examples is:
hypothesis = <Red, Round, Big, ?>

Among the given options, the closest one to the most specific hypothesis found is:
(a) <Red, Round, Big, ?>

So, the correct option is a. <Red, Round, Big, ?>

---

2. Find the number of instances possible in X using the values that can be seen in the table in Q1.
   a. 12
   b. 48
   c. 36
   d. 24

Correct Answer: b

Detailed Solution:

Color: Red, Blue, Green (3 values)
Shape: Round, Flat (2 values)
Size: Big, Small (2 values)
Texture: Hard, Soft (2 values)
Category: Yes, No (2 values)

Total number of possible instances = Number of values for Color * Number of values for Shape * Number of values for Size * Number of values for Texture * Number of values for Category
Total number of instances = 3 * 2 * 2 * 2 * 2 = 48
The correct answer is:
b. 48

Do Answer Questions 3-4 with the given data:

3. Suppose the instance space, X is the set of real numbers, ℝ and H be the set of intervals on the real number line. H is of the form a<x<b, where a and b may be real constants. Find VC(H). [VC stands for Vapnik-Chervonenkis Dimension]
   a. 2
   b. 3
   c. 5
   d. 4

   Correct Answer: a
   Detailed Solution:
   Consider a particular instance containing two distinct instances, S={3.1,5.7}. S can be shattered by H, since we can find intervals (1< x < 2), (1 < x < 4), (4 < x < 7) and (1 < x < 7). Hence, VC(H)=2. For the definition of Shattering, Follow Lecture Slides.

---

4. Can VC dimension of H be 3?
   a. Yes
   b. No

   Correct Answer: b
   Detailed Solution:
   Assume a set S = {x0, x1, x2}. Without loss of generality, we assume x0 < x1 < x2. Clearly, we cannot find a closed interval that includes x0 and x2 but not x1. Hence, VC(H) cannot be 3.

Do Answer questions 5-6 with the given data:

5. Suppose you have trained three classifiers, each of which returns either 1 or −1, and tested their accuracies to find the following:

Classifier	Accuracy
c1	0.6
c2	0.55
c3	0.45

   Let C be the classifier that returns a majority vote of the three classifiers. Assuming the errors of the ci are independent, what is the probability that C(x) will be correct on a new test example x?

   a. 0.1815
   b. 0.1215
   c. 0.5505
   d. 0.099

   Correct Answer: c

Detailed Solution:

Here are the cases for which C would be correct and their probabilities:

c1	**c2**	**c3**	**Probability**
Correct	Correct	Correct	0.6*0.55*0.45 = 0.1485
Correct	Incorrect	Incorrect	0.6*0.45*0.45 = 0.1215
Incorrect	Correct	Correct	0.4*0.55*0.45 = 0.099
Correct	Correct	Incorrect	0.6*0.55*0.55 = 0.1815

Total probability of C to be correct = 0.1485 + 0.1215 + 0.099 +0.1815 = 0.5505

---

6. Suppose you have run Adaboost on a training set for three boosting iterations. The results are classifiers h1, h2, and h3, with coefficients α1 = .2, α2 = −.3, and α3 = −.2. You find that the classifiers results on a test example x are h1(x) = 1, h2(x) = 1, and h3(x) = −1, What is the class returned by the Adaboost ensemble classifier H on test example x?

a. 1
b. -1

Correct Answer: a
Detailed Solution:
H(x) = sgn ((.2)(1) + (−.3)(1) + (−.2)(−1)) = sgn(.1) = 1. Refer to lecture slides for detailed explanation on Adaboost.

---

7. Bagging is done to ________________
a. increase bias
b. decrease bias
c. increase variance
d. decrease variance

Correct Answer: d
Detailed Solution:
Bagging is a way to decrease the variance of our prediction by generating additional data for training from our original dataset using combinations with repetitions to produce multisets of the same cardinality/size as our original data.

---

8. Weak learners are the ones used as classifiers in Boosting algorithms. They are called weak learners because ____________
a. Error rate greater than 0.5
b. Error rate less than 0.5
c. No error

Correct Answer: b

Detailed Solution:

Weak learner is a learner that will always do better than chance, when it tries to label the data. Doing better than chance means we are always going to have an error rate which is less than 1/2.

---

9. Dropout is used as a regularization technique in Neural Networks where many different models are trained on different subsets of the data. In ensemble learning, dropout techniques would be similar to ________________

a. Bagging
b. Boosting
c. None of the above

Correct Answer: a
Detailed Solution:
Dropout training is similar to bagging (Breiman, 1994), where many different models are trained on different subsets of the data. Dropout training differs from bagging in that each model is trained for only one step and all of the models share parameters.

---

10. 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

Correct Answer: c

Explanation:
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.

---

11. Considering the AdaBoost algorithm, which among the following statements is/are 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.

    Correct Answer: b, d

    Explanation: 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).

---

12. The VC dimension of hypothesis space H1 is larger than the VC dimension of hypothesis space H2. Which of the following can be inferred from this?
    a. The number of examples required for learning a hypothesis in H1 is larger than the number of examples required for H2.
    b. The number of examples required for learning a hypothesis in H1 is smaller than the number of examples required for H2.
    c. No relation to number of samples required for PAC learning.

    Correct Answer: a

    Explanation: From the definition of VC dimension, VC dimension is directly proportional to m (no of training samples)

---

13. For a particular learning task, if the required error parameter ϵ changes from 0.2 to 0.01, then how many more samples will be required for PAC learning?
    a. Same
    b. 2 times
    c. 20 times
    d. 200 times

Correct Answer: c

Detailed Solution:

We know that,

$$m \geq \frac{1}{\epsilon}(\log(|H|) + \log(\frac{1}{\delta}))$$

m is inversely proportional to ϵ.

$$m1 * \epsilon 1 = m2 * \epsilon 2$$

$$m2 = (m1 * \epsilon 1)/ \epsilon 2$$

$$m2 = (m1 * 0.2)/0.01$$

$$m2 = 20 * m1$$

---

14. In boosting, which data points are assigned higher weights during the training of subsequent models?

    a. Data points that are classified correctly by the previous models.
    b. Data points that are misclassified by the previous models.
    c. Data points that are randomly selected from the training data.
    d. Data points that are ignored during training.

Correct Answer: b

Explanation:

Boosting assigns higher weights to data points that are misclassified by previous models. This allows subsequent models to focus on the errors made by the earlier models and improve the overall accuracy of the ensemble.

---

15. In AdaBoost, how are the individual weak learners combined to form the final strong ensemble model's prediction?
    a. By taking the majority vote of all weak learners' predictions.
    b. By averaging the predictions of all weak learners.
    c. By weighting the predictions of weak learners based on their accuracy.
    d. By selecting the prediction of the weak learner with the highest accuracy.

Correct Answer: c

Explanation:

In AdaBoost, each weak learner is assigned a weight based on its accuracy in the ensemble. The final prediction of the ensemble model is determined by aggregating the weighted predictions of all weak learners. Weak learners that perform better have higher weights, indicating more influence on the final prediction. This mechanism allows AdaBoost to focus more on the predictions of more accurate learners.

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