INTRODUCTION, LIMITS OF LEARNING, GEOMETRY AND NEAREST NEIGHBORS

SYLLABUS

Introduction: What does it mean to Learn, Some Canonical Learning Problems, The Decision Tree Model of Learning, Formalizing the Learning Problem, ID3 Algorithm.

Limits of Learning: Data Generating Distributions, Inductive Bias, Not Everything is Learnable, Underfitting and Overfitting, Separation of Training and Test Data, Models, Parameters and Hyperparameters, Real World Applications of Machine Learning.

Geometry and Nearest Neighbors: From Data to Feature Vectors, k-Nearest Neighbors, Decision Boundaries, k-means Clustering, High Dimensions.

LEARNING OBJECTIVES

- Brief Introduction to Machine Learning and Canonical Problems
- Decision Tree Model of Learning
- ID3 Algorithm and Entropy
- Bayes Optimal Classifier and Inductive Bias
- Introduction to Underfitting and Overfitting Concepts
- Real World Applications of Machine Learning
- Feature Vectors and Decision Boundary
- K-Nearest Neighbor Algorithm
- K-means Clustering Algorithm.

INTRODUCTION

The Artificial Intelligent (AI) systems have the learning capability as humans have. But, the learning capability of AI systems is not the same as that of human learning capability i.e., the human capability of learning is higher than the AI systems. The AI systems possess some sort of mechanical learning capabilities, which are referred to as 'machine learning'. Various methods of machine learning are available. Some of them are, inductive learning, Artificial Neural Networks (ANN) and genetic algorithms.

Machine learning works well in many cases but it has few limitations. There are various reasons why a machine learning algorithm might fail on some learning tasks. Some of the reasons can be noise at feature and label level, limited features for learning etc.

K-nearest neighbor algorithm is considered as a supervised learning algorithm. It is used for classification and regression. The working principle of this algorithm is to label a newly entered data point into a group by using any of the distance metrics.

PART - A

SHORT QUESTIONS AND ANSWERS

Q1. What is machine learning?

Answer :

(Model Paper-II, Q2)

The artificial intelligent systems have the learning capability as humans have. But, the learning capability of AI systems is not the same as that of human learning capability i.e., the human capability of learning is higher than the AI systems. The AI systems possess some sort of mechanical learning capabilities, which are referred to as '*machine learning*'. Various methods of machine learning are available. Some of them are, inductive learning, Artificial Neural Networks (ANN) and genetic algorithms.

Q2. What is learning? Explain about different components of learning process.

Answer :

(Model Paper-I, Q1 | June/Aug.-22, Q1 (OR))

Learning

Machine learning field is mostly concerned about how to generate computer programs with improved experience. A computer program learns from experience E corresponding to some class of tasks T and performance measure P, if at all the performance at tasks in T as measured by P will be improved with experience E.

A Checkers Learning Problem

Task T: Playing checkers

Performance measure P: Percent of games won against opponents

Training experience E: Playing practice games against it self.

The definition of learning is capable of including the tasks which are called as "learning" tasks. It is even capable to encompass the computer programs which improvise from experience in straight forward way. The main aim is to define the class of problems precisely which encompass the interesting forms of learning in order to explore the algorithms which are used to solve and understand structure of learning problems and processes.

Components of Learning Process

The different components of learning process are,

1. Data Storage

Data storage is considered as an important component of the learning process. It is used to store and retrieve huge amounts of data.

2. Abstraction

Abstraction can be defined as process of obtaining knowledge from stored data. In general, it creates concepts from the data. So, in order to perform this process, known models are implemented/applied and also new models must be created.

3. Generalization

Generalization can be defined as a process of transforming knowledge of stored data into a form, that can be used for future purposes. The main objective of generalization is to identify inferences (or) properties of data that can be useful for future actions.

4. Evaluation

In learning process, evaluation is the last component which can be defined as a process of providing feedback to the user inorder to determine the utility of learned knowledge. This feedback helps to improve the learning process for future purpose.

Q3. Explain about binary and multiclass classification in detail.

Answer :

(Model Paper-II, Q1 | June/Aug.-22, Q2 (OR))

Binary Classification

Binary classification can be defined as a classification that tries to predict a binary value that represented either yes or no for an item.

Multiclass Classification

Multiclass classification can be defined as a classification that tries to predict one example from a number of classes for an item.

Q4. Write in brief about Decision Tree Model of Learning.

Answer :

(Model Paper-III, Q1 | July/Aug.-22, Q1 [MGU])

The decision tree can be considered as the traditional model of learning. It is similar to the concept of divide and conquer. This model can be used with a variety of learning problems. For instance, consider the binary classification problem which results only the output as either yes or no.

Consider an example to find whether a student likes a specific course or not. This can be done by throwing a few course related questions. They are as follows,

You: Is the course under consideration in Systems?

Me: Yes

You: Has this student opted for any other Systems courses?

Me: Yes

You: Has this student enjoyed the courses opted before?

Me: No

You: The result prediction is that the student won't like this course.

The above conversation can be represented in tree model as shown in the below figure,

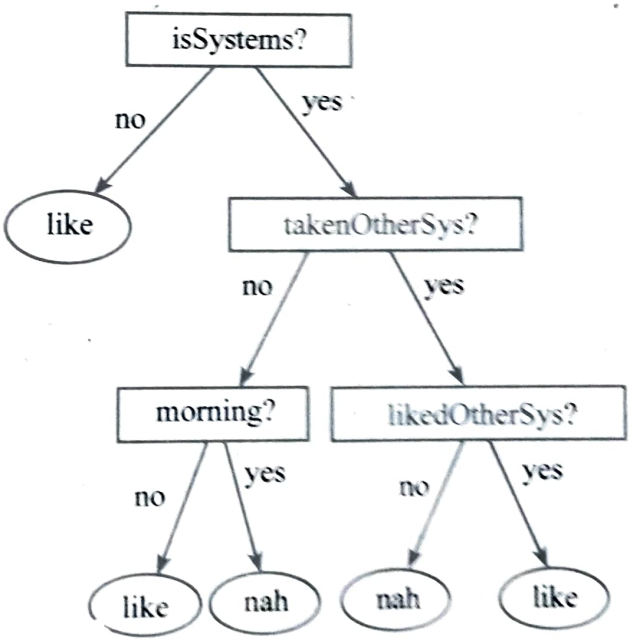


Figure: Decision Trees

Q5. What is entropy?

Answer :

The entropy can be given as,

Given probabilities $p_1, p_2, \ldots, p_s$ where $\sum_{i=1}^{s} p_i = 1$, entropy is defined as,

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} (p_i \log(1/p_i))$$

The value of entropy can be between 0 and 1 and can be when the probability of independent divisions are same.

Q6. Write the statement and proof of bayes optimal classifier.

Answer :

Model Paper-III, Q2

Statement

The Bayes Optimal Classifier (BOC) achieves minimal error, the error of any deterministic classifier

Assume that g is a classifier which is more efficient than $f^{(BO)}$. In such a way that there should be $g(x) \neq f^{(BO)}(x)$ for a variable x.

Now, the probability that $f^{(BO)}$ leads to an error on the variable x is,

$1 - D(x, f^{(BO)}(x))$

The probability that g makes an error on the variable x is,

$1 - D(x, g(x))$

According to the statement, we know that $f^{(BO)}$ was opted to maximise $D(x, f^{(BO)}(x))$, so this must be greater than $D(x, g(x))$.

Therefore, the probability that $f^{(BO)}$ results on this variable x is smaller than the probability that g results on x. This applies to any x for which $f^{(BO)}(x) \neq g(x)$ and hence, $f^{(BO)}$ achieves a smaller zero/one error than any g.

Q7. Define Inductive bias.

Answer :

(Model Paper-I, Q2 | June/Aug.-22, Q3 [OU])

Inductive bias can be defined as a set of assumptions that can justify the classifications that are assigned to future instances. This can be done along with training data. For a set of training examples there might be a number of trees which are compatible with these examples. Inductive bias can be explained based on its selection among the compatible hypotheses.

Q8. Explain hyperparameters with an example.

Answer :

Model Paper-II, Q3

Hyperparameters

Hyperparameters can be defined as the parameters which can control the values of other parameters in a specific model. These are also known as knobs whose values can be adjusted. These knobs may also modify the inductive bias of an algorithm.

Example

In a decision tree model, the maximum depth parameter is considered as the hyperparameter. In the decision tree training algorithm this parameter is set in such a way that it is adjusted between the overfitting and underfitting.

Q9. Define feature vectors and decision boundary.

Answer :

Model Paper-I, Q3

Feature Vectors

Feature vectors can be defined as the numerical representation of the input data. They represent the characteristics/features of data points.

Decision Boundary

A decision boundary can be defined as a mathematical boundary or a border that separates two regions. This can be done by using various optimization techniques like decision trees, neural networks, and support vector machines.

Q10. Define K-means clustering.

Answer :

Model Paper-III, Q3

The k-means is an iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is achieved. In this algorithm, the mean value of the objects in the cluster represents a cluster. It is the most popular and commonly used method. The algorithm is built on the concept of user specified input parameter (k). A set of 'n' objects are divided into 'k' clusters by the algorithm. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The cluster's centroid gives the measure of cluster's similarity.

PART - B

ESSAY QUESTIONS AND ANSWERS

1.1 INTRODUCTION

1.1.1 What does it Mean to Learn, Some Canonical Learning Problems

Q11. Define Machine Learning. Write about some of the canonical learning problems.

Answer :

Machine Learning

The artificial intelligent systems have the learning capability as humans have. But, the learning capability of AI systems is not the same as that of human learning capability i.e., the human capability of learning is higher than the AI systems. The AI systems possess some sort of mechanical learning capabilities, which are referred to as '*machine learning*'. Various methods of machine learning are available. Some of them are, inductive learning, Artificial Neural Networks (ANN) and genetic algorithms.

Canonical Learning Problems

Canonical problems can be defined as the challenges that occur while researchers are working with the concept of machine learning.

There are many canonical learning problems. These problems are differentiated by using the values that the problems are trying to predict. Some of the examples are as given in the below table,

Type of Value of Thing	Description	Example
Regression	It tries to predict the real value.	Predicting the today's value of a stock in the market based on the past values.
Binary Classification	It tries to predict a binary value that represented either yes or no.	Predicting whether a student like the SIA spectrum content or not.
Multiclass Classification	It tries to add an example to a number of classes.	Predicting the next course of action after the compilation errors about modules, packages, syntax and logical errors in a python program.
Ranking	It tries to group a particular set of value in the required order of relevance.	Predicting the order of web pages when a user executes a specific query.

1.1.2 The Decision Tree Model of Learning

Q12. Explain in detail about the decision tree model of learning.

Answer : **Model Paper-II, Q13(a)**

Decision Tree Model of Learning

The decision tree can be considered as the traditional model of learning. It is similar to the concept of divide and conquer. This model can be used with a variety of learning problems. For instance, consider the binary classification problem which results only the output as either yes or no.

Example

Consider an example to find whether a student likes a specific course or not. This can be done by throwing a few course related questions. They are as follows,

You: Is the course under consideration in Systems?

Me: Yes

You: Has this student opted for any other Systems courses?

Me: Yes

You: Has this student enjoyed the courses opted before?

Me: No

You: The result prediction is that the student won't like this course.

The above conversation can be represented in tree model as shown in the below figure,

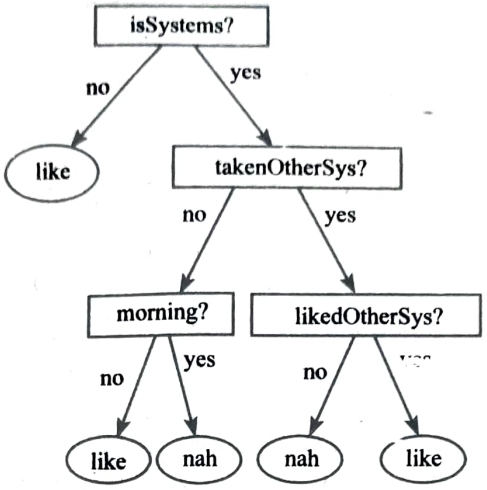


Figure: Decision Trees

In the above figure, the questions are provided in rectangles known as internal nodes and the guesses are provided in ovals known as child nodes.

In general, while performing the actual work with the testing data, the questions being asked are considered as features and the answers are considered as feature value. And, the rating is called a label.

The best way to find the useful guess out of various guesses can be determined by using the histogram for Labels. The histogram for labels with the above example can be given as shown in the below figure,

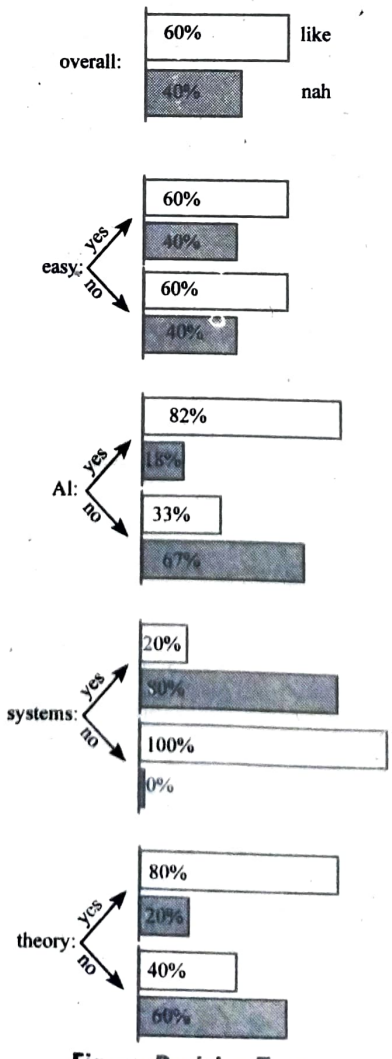


Figure: Decision Trees

Algorithm

The algorithm for decision model of learning problem can be given as,

```
DecisionTreeTraining(data, remaining_features)
guess ← most frequent answer in data                    // considered as default answer
if the labels in data are unambiguous then
    return Leaf(guess)                                  //base case: splitting further is not required
else
    if remaining_features is empty
    then
        return Leaf(guess)      //base case: splitting further is not required
    else                                                // Ask for more features
        for all f ∈ remaining_features do
            NO ← the subset of data on which f=no
            YES ← the subset of data on which f=yes
            score[f] ← # of majority vote answers in NO
                    + # of majority vote answers in YES // the accuracy we would get if we only queried on f
        end for
    f ← the feature with maximal score(f)
    NO ← the subset of data on which f=no
    YES ← the subset of data on which f=yes
    left ← DecisionTreeTrain(NO, remaining_features \ {f })
    right ← DecisionTreeTrain(YES, remaining_features \ {f })
    return Node(f, left, right)
    end if
DecisionTreeTesting(data, remaining_features)
if tree is of the form Leaf(guess) then
    return guess
else if tree is of the form Node(f, left, right) then
    if f = no in test point then
        return DecisionTreeTest(left, test point)
    else
        return DecisionTreeTest(right, test point)
    end if
end if
```

In the above algorithms, as soon as the leaf node reaches during the traversal, the test point provides the values associated with that leaf node.

1.1.3 Formalizing the Learning Problem

Q13. Explain about formalizing the learning problem.

Answer :

The various problems that occur while formalizing the learning problems are as follows,

1. The unseen test data must be used for measuring the performance of a specific learning algorithm.
2. The performance measurement must depend on the problem that is to be solved.
3. The existence of strong relationships among data that is being tested at training time and at testing time.

These problems can be solved by following below three steps,

Step-1: Assuming a Loss Function

Consider the loss function as $l(y, \hat{y})$, where y and $\hat{y}$ represent the truth and the prediction by the system respectively. And the overall function $l(y, \hat{y})$ gives the measure of error.

The loss functions for some of the canonical problems can be given as,

Regression. Squared loss $l(y, \hat{y}) = (y - \hat{y})^2$ or absolute loss $l(y, \hat{y}) = |y - \hat{y}|$.

Binary Classification: Zero/one loss $l(y, \hat{y}) = 0$ if $y = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{Otherwise} \end{cases}$

Multiclass Classification : Zero/one loss.

The loss function can be defined based on the type of problem.

Step-2: Estimate a Model

Consider a model which provides the training and testing data. In this case, consider the model as a probabilistic model of learning. It provides the distribution D over the input and output pairs (x,y). Here, x represents the input and y represents the output. Note that, the D provides low probability to unreasonable (x, y) pairs and high probability to reasonable (x,y) pairs. Here, the input/output pair can be unreasonable in two different ways. They are as follows,

- x can be a unusual or odd input
- y can be unusual for the given input pair, x.

Note that, there are no assumptions made on the distribution D. Thus, the data generating distribution and the data are unknown. Here, the complex portion is what type of random data that the distribution provides for testing.

Step-3: Compute a Function f that has Low Expected Error E over D with Respect to l

The mathematical representation of expected loss E over D with respect to loss function can be represented as,

$$\in \triangleq \mathbb{E}_{(x,y)\sim D}[l(y, f(x))] = \sum_{(x, y)} D(x, y)\, l(y, f(x))$$

Note that, the function f should be as minimum as possible. But, the main complexity is reducing the error value because the value of D is unknown. In this case, say D be the random training data given with N input/output pairs i.e., $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_N, y_N)$. Then, the learning function to determine the training error can be defined as,

$$\hat{\in} \triangleq \frac{1}{N} \sum_{n=1}^{N} l(y_n, f(x_n))$$

Note that, the error can be completely minimized to zero only when the realistic future data is available instead of the training data.

1.1.4 ID3 Algorithm

Q14. Explain ID3 algorithm and illustrate with suitable examples.

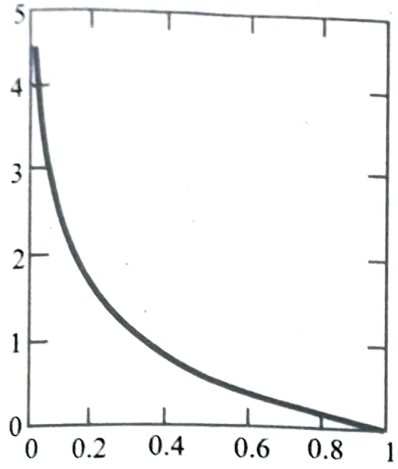
Answer :

(Model Paper-I, Q13(a) | June/Aug.-22, Q13(a) [OU])

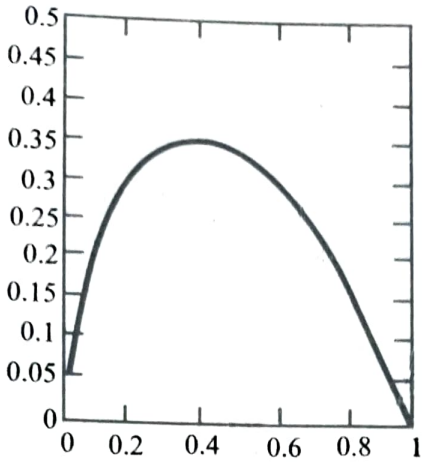
The ID3 (Iterative Dichotomiser 3) algorithm is a decision tree algorithm that is used to divide the data into various subsets depending on the features. The main objective of this algorithm is to reduce the number of comparisons between the data sets.

This algorithm is similar to the intuitive approach where adults play a game called "Twenty Questions". Here, the adult may ask the first question as "Is the thing alive?", whereas a child may ask is "Is it my daddy?". In this example, the question asked by an adult splits the search space into two large search domains (two subsets with almost the same probability of occurrence) while the question asked by the child splits the search space into two small search domains (i.e., one subset with small probability and the other subset with high probability of occurrence). The search space is divided in such a way that the attributes which may have probability of providing most information gain will be divided firstly.

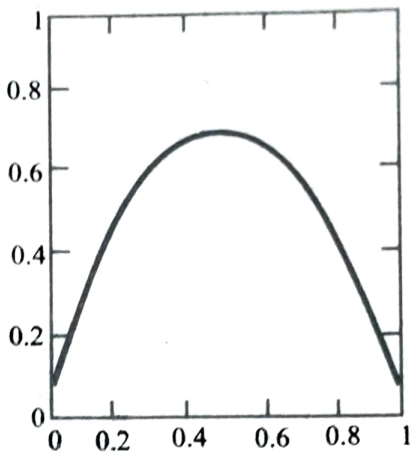
The information provided in the data sets can be quantified by using a concept called entropy. Entropy can be used to compute the quantity of randomness/surprise/uncertainty within the data. The entropy can be zero when all the data in datasets is categorized into a single class because there won't be any uncertainty. The ultimate goal of the decision tree approach is to classify the data in such a way that the final subsets with similar attribute values



(a) log(1/p)



(b) p log(1/p)



(c) H (p, 1 – p)

Figure: Entropy

In the above figure(a), the log(1/p) has the probability p with value between 0 and 1. Depending on the value of p, the amount of uncertainty can be calculated. For instance, if probability of occurrence(p) is 1, then the uncertainty will be 0(Since log(1)=0). If the value of p = 0, then the uncertainty value will be increased. Generally, the decision tree classification method classifies the data in such a way that the sum of the splitted data sets should be equal to 1. In the case of the "Twenty Questions" game, P(Daddy)>P(~Daddy) and P(Daddy)+P(~Daddy)=1. The complete information related to the divisions can be computed by calculating the average associated information with independent events.

The figure (b) defines the expected information on the basis of probability of events occurring. The overall information associated with the events can be computed by adding the information associated with the individual events.

The figure (c) defines the maximum entropy that occurs when the probabilities of two divisions are same.

Formal Definition of Entropy

The entropy can be given as,

Given probabilities $p_1, p_2, \ldots, p_s$ where $\sum_{i=1}^{s} p_i = 1$, entropy is defined as,

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} (p_i \log(1/p_i))$$

The value of entropy can be between 0 and 1 and can be when the probability of independent divisions are same.

Example

Consider the below table data for height classification,

Name	Gender	Height	Output1	Output2
Kristina	F	1.6 m	Short	Medium
Jim	M	2 m	Tall	Medium
Maggie	F	1.9 m	Medium	Tall
Martha	F	1.88 m	Medium	Tall
Stephanie	F	1.7 m	Short	Medium
Bob	M	1.85 m	Medium	Medium
Kathy	F	1.6 m	Short	Medium
Dave	M	1.7 m	Short	Medium
Wroth	M	2.2 m	Tall	Tall
Steven	M	2.1 m	Tall	Tall
Debbie	F	1.8 m	Medium	Medium
Todd	M	1.95 m	Medium	Medium
Kim	F	1.9 m	Medium	Tall
Amy	F	1.8 m	Medium	Medium
Wynette	F	1.75 m	Medium	Medium

In the given 15 rows of data, $\left(\frac{4}{15}\right)$ are short, $\left(\frac{8}{15}\right)$ are medium and $\left(\frac{3}{15}\right)$ are tall.

Therefore, the entropy can be calculated as,

$$= \left(\frac{4}{15}\right)\log\left(\frac{\frac{1}{4}}{\frac{15}{15}}\right) + \left(\frac{8}{15}\right)\log\left(\frac{\frac{1}{8}}{\frac{15}{15}}\right) + \left(\frac{3}{15}\right)\log\left(\frac{\frac{1}{3}}{\frac{15}{15}}\right)$$

$$= \left(\frac{4}{15}\right)\log\left(\frac{15}{4}\right) + \left(\frac{8}{15}\right)\log\left(\frac{15}{8}\right) + \left(\frac{3}{15}\right)\log\left(\frac{15}{3}\right)$$

$$= \left(\frac{4}{15}\right)(0.574) + \left(\frac{8}{15}\right)(0.273) + \left(\frac{3}{15}\right)(0.6989)$$

$$= 0.4384$$

If the gender is considered as a dividing attribute, the table data consists of 9 F tuples and 6 M tuples. Thus, the entropy for F tuple can be calculated as,

$$= \left(\frac{3}{9}\right)\log\left(\frac{\frac{1}{3}}{\frac{9}{9}}\right) + \left(\frac{6}{9}\right)\log\left(\frac{\frac{1}{6}}{\frac{9}{9}}\right)$$

$$= \left(\frac{3}{9}\right)(0.4771) + \left(\frac{6}{9}\right)(0.176)$$

$$= 0.2764$$

The entropy for type M can be calculated as,

$$= \left(\frac{1}{6}\right)\log\left(\frac{\frac{1}{1}}{\frac{6}{6}}\right) + \left(\frac{2}{6}\right)\log\left(\frac{\frac{1}{2}}{\frac{6}{6}}\right) + \left(\frac{3}{6}\right)\log\left(\frac{\frac{1}{3}}{\frac{6}{6}}\right)$$

$$= \left(\frac{1}{6}\right)(0.7781) + \left(\frac{2}{6}\right)(0.4771) + \left(\frac{3}{6}\right)(0.3010)$$

$$= 0.4392$$

By using the ID3 algorithm, the weighted sum of these last entropies can be given by,

$$= \left(\left(\frac{9}{15}\right) \times 0.2764\right) + \left(\left(\frac{6}{15}\right) \times 0.4392\right)$$

$$= 0.34152$$

The gain entropy can be calculated by using the gender attribute can be given as,

Gain entropy $= 0.4384 - 0.34152 = 0.0968$

Now, consider the height as a splitting attributes, then we can say that, there are,

2 tuples of 1.6 m

2 tuples of 1.7 m

1 tuple of 1.75 m

2 tuples of 1.8 m

1 tuple of 1.85 m

1 tuple of 1.88 m

2 tuples of 1.9 m

1 tuple of 1.95 m

1 tuple of 2 m

1 tuple of 2.1 m and

1 tuple of 22 m

There are 11 different tuples. Now, the ranges for the above tuples can be given as,

$(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty)$

For all the above ranges, the entropy will be 0 for all ranges except

$$(1.9, 2.0] = \left(0 + \frac{1}{2}(0.301) + \frac{1}{2}(0.301)\right) = 0.301$$

Therefore, the gain entropy for the height attributes can be given as,

$$\text{Gain Entropy}_{\text{height}} = 0.4384 - \frac{2}{15}(0.301) = 0.3893$$

After considering all these attributes, the gain entropy is still higher. Therefore, it requires further splitting as shown in below figure,

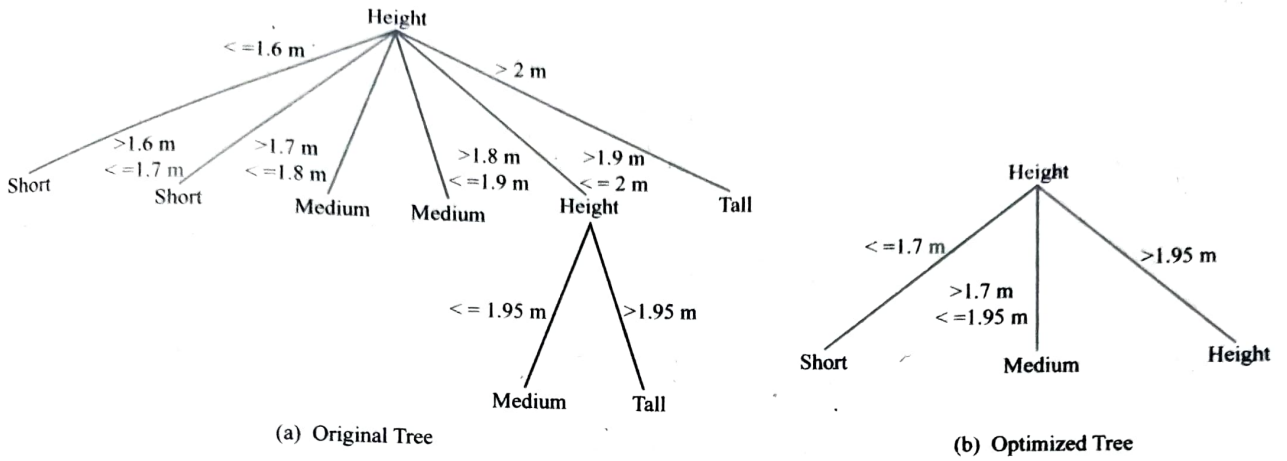


Figure: Classification Problem

Q15. Explain how to formalizing the learning problem and ID3 Algorithm with example.

Answer :

(Model Paper-III, Q13(a) | July/Aug.-22, Q7 [MGU])

Formulizing the Learning Problem

For answer refer Unit-I, Page No. 7, Q.No. 13.

ID3 Algorithm

For answer refer Unit-I, Page No. 8, Q.No. 14.

1.2 LIMITS OF LEARNING

1.2.1 Data Generating Distributions, Inductive Bias

Q16. Explain about bayes optimal classifier.

OR

Write about data generating distributions.

Answer :

Model Paper-III, Q13(b)

Bayes optimal classifier can be defined as a classifier that takes $\hat{x}$ as input and returns $\hat{y}$ which enlarges the COMPUTED($\hat{x}$, $\hat{y}$). Where, COMPUTED() is a python function that tasks x and y as inputs and results in the probability of x, y over the distribution D. This can be mathematically represented as,

$$f^{(BO)}(\hat{x}) = \arg\max_{\hat{y} \in y} D(\hat{x}, \hat{y})$$

The bayes classifier is said to be optimal in a case i.e., it achieves minimal zero/one error of any deterministic classifier.

Theorem

Statement

The Bayes Optimal Classifier $f^{(BO)}$ achieves minimal zero/one error of any deterministic classifier.

Proof

Assume that g is a classifier which is more efficient than $f^{(BO)}$. In such a way that there should be $g(x) \neq f^{(BO)}(x)$ for a variable x.

Now, the probability that $f^{(BO)}$ leads to an error on the variable x is,

$1 - D(x, f^{(BO)}(x))$

The probability that g makes an error on the variable x is,

$1 - D(x, g(x))$

According to the statement, we know that $f^{(BO)}$ was opted to maximise $D(x, f^{(BO)}(x))$, so this must be grea than $D(x, g(x))$.

Therefore, the probability that $f^{(BO)}$ results on this variable x is smaller than the probability that g results on x. applies to any x for which $f^{(BO)}(x) \neq g(x)$ and hence, $f^{(BO)}$ achieves a smaller zero/one error than any g.

Q17. Explain about inductive bias.

Answer :

Inductive Bias

Inductive bias can be defined as a set of assumptions that can justify the classifications that are assigne future instances. This can be done along with training data. For a set of training examples there might be a num of trees which are compatible with these examples. Inductive bias can be explained based on its selection am the compatible hypotheses.

Example

Consider the below images that represents the training data,

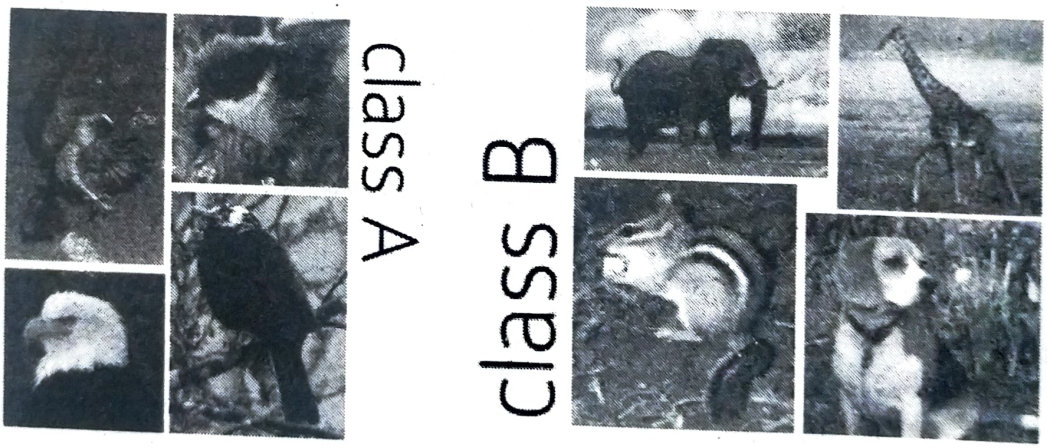


Figure (1): Training Data for Binary Classification with Labels Class A, B

Also, the below image represents the test data.

After observing the training data in figure (1), the test data in figure (2) can be labelled as either AABB or ABBA. But the accurate answer can not be revealed by the user because it requires around 100's of samples out of which 60-70 percent of people come up with ABBA and 30-40 percent of people come up with AABB. This is because some of the people classify the images in figure (1) as fly/non-fly and the remaining people as bird/non-bird classification. In machine learning, this point of differentiation of classification over another one is known as inductive bias.

In the decision tree learning algorithm, a variant called shallow decision tree has been introduced. It does not allow the user to query over a predefined depth d. In other words, once the user queries with d-many features, then he is not allowed to query more than that specific depth.

1.2.2 Not Everything is Learnable, Underfitting and Overfitting

Q18. What are the various reasons that proves machine learning fails in some cases or situations?

Answer :

The various reasons that prove machine learning fails on some specific types of tasks are as follows,

1. The main reason is that the training data may include noise at feature level as well as label level. These features lead to the incorrect measurements using sensors.

   Example

   Inorder to find the distance from a specific point to a tree, the robot may use a sensor which eventually returns an incorrect value. And, this is considered as the noise at feature level. Whereas, the audience writes a bad review for a movie and clicks on the wrong label for submission. This is considered as a noise at label level.

2. Limited features for learning about Machine learning algorithms.

   Example

   During the process of diagnosing the cancer, doctors may collect various data like X-rays, gene expressions, cancer history in the family and so on. Even with all this information, it would be difficult to diagnose whether a person has been diagnosed with cancer or not.

3. Not all the cases/situations require one correct answer. There may be some cases where it requires more than one answer.

   Example

   The development of a "safe web search" that deletes the offensive web page results when a user searches for some query. This can be done by collecting some samples and distributing them over different people to identify them as if they are offensive web pages or not. In this case, the result completely depends on the user's perspective. That means, the page that is marked as offensive by one person may or may not be marked as offensive by another function. Thus, these types of problems need to be isolated as a source of difficulty.

4. The inductive bias of machine learning algorithms is completely different from the current learning concepts.

   Example

   In a bird/non-bird classification problem, there is a requirement of few more training data examples to clarify whether this comes under bird/non-bird data classification or fly/non-fly classification. Even with more training data, it is difficult to classify them. The reason is that the inductive bias of the learner does not match with the target classification learning algorithm.

Q19. Write about underfitting and overfitting.

Answer :

Underfitting

Underfitting can be defined as a concept where there is a possibility of learning something and it is not happened. For instance, consider a student who is not prepared for an exam but still allowed to write an exam. Hence, he cannot perform well.

Overfitting can be defined as a concept where the focus is more on the training data but not able to summarise properly. For instance, consider a student who concentrates more on memorising the answers to the previously asked questions instead of understanding the concept.

Example

Initially, consider the construction of an empty decision tree on the given data which includes training data as 'yes' and 'no' (i.e., 12 yes and 8 no out of 20 samples). In this case the training error will be 100*(8/20)=40%.

In another case, consider the construction of a full decision tree. Here, all the leaves have at least one example. Thus the training error will be 100*(0/20)=0%. The main objective of the decision tree is not to provide training error as 0 but to fit to the unknown future data.

1.2.3 Separation of Training and Test Data, Models, Parameters and Hyperparameters

Q20. Explain about separation of training and testing data.

Answer :

The separation of training and testing data is the best way of proving your learning model in the real-time environment. In this approach, initially, separate some amount of data as test data and keep the remaining as training data. Now, perform the machine learning algorithms on the training, then only perform the user-defined learning model with the test data and report the results to higher authorities for review. This makes a good impression on your learning model.

For example, out of 1000 samples consider 800 as training data and remaining 200 as test data. Now, perform the existing machine learning algorithm on the 800 samples of training data. Once that is accomplished then apply your learning models on 200 test data. Finally, prepare a report with the result and submit it to authorities for review.

This kind of separation is called an 80/20 split. In some cases, when the data is large, most of the analysts use a 90/10 split. Note that, the cardinal rule of machine learning states that - 'Never and Ever touch the test data if that is not clear enough to perform any model on it'. If you touch the test data even if it is not clear enough, then the proposed model will not be considered as efficient for future unknown data.

Q21. Write about model, parameters and hyperparameters.

Answer :

Model

A model can be defined as an approach that is used to obtain the various estimating algorithms in machine learning. It can also be defined as a mathematical expression representing the relationship between the given input data and result.

Example

A decision tree for the classification decision of a student/course pair is considered as a model.

Parameters

Parameters can be defined as the values that can be used within a model to estimate or predict a decision.

Example

Some of the parameters in a decision tree model are,

- The various questions asked by the users
- The arrangement order proposed by the users, and
- The classification decisions at the leaves.

Generally, the values of these parameters vary to reduce the error percentage between the prediction by the mode and the real-output.

Hyperparameters can be defined as the parameters which can control the values of other parameters in a specific model. These are also known as knobs whose values can be adjusted. These knobs may also modify the inductive bias of an algorithm.

Example

In a decision tree model, the maximum depth parameter is considered as the hyperparameter. In the decision tree training algorithm this parameter is set in such a way that it is adjusted between the overfitting and underfitting.

1.2.4 Real World Applications of Machine Learning

Q22. What are the sequence of decisions that must be made to deploy a machine learning approach in the real world.

Answer :

The various sequence of decisions that must be made to deploy a machine learning approach in the real world are represented in the below table with an example(Increasing the revenue for a search engine),

Steps	Generic Decision	Example with Description
1.	Real-world goal	The main goal is to increase the revenue for the search engine.
2.	Real-world mechanism	The mechanism is to display or show best advertisements to users.
3.	Learning problem	Train a classifier that determines whether the user clicks on the ad or not. This needs some training data samples
4.	Data collection	The training data sample can be collected by connecting with the current system.
5.	Collected data	The displayed ad can be logged as soon as the user enters the search engine. If the user clicks on the ad, then it is represented by 1 otherwise 0.
6.	Data representation	The collected data will be converted into input/output pairs. Here, the bag of words are represented by bow and clicks are represented by + or – click.
7.	Select model family	Choose any model family and inductive bias. Here, depth less than or equal to 20 decision trees.
8.	Select training data	Select the sample data which can further be divided into training and development.
9.	Train model and Hyperparams	Choose a final decision tree having largest depth on the development data.
10.	Predict on test data	Perform actions and predict some results on the test data.
11.	Evaluate error	Evaluate the complete quality of our predictor as zero or classification error on the given test data.
12.	Deploy	Finally, deploy the system.

Note that, wrong execution of any one step throughout the process leads to unnecessary errors.

1.3 GEOMETRY AND NEAREST NEIGHBORS

1.3.1 From Data to Feature Vectors, k-Nearest Neighbors

Q23. Write in short about feature vectors.

Answer :

Feature Vectors

Feature vectors can be defined as the numerical representation of the input data. They represent the characteristics/features of data points.

Consider an example of a movie review that contains "super" two times, one is exclamatory mark and underlined text. This can be represented in the form of a feature vector as <2, 1, 1>.

The feature values can be interpreted by using the binary digits that are 0 and 1. In consideration of real valued features, the conversion from feature values to vectors is very straightforward. For example, the objective is to find the object in an image is orange, tomato, grape, or, blueberry. The user needs the color of the objects i.e., yellow, red, green and blue respectively to the object mentioned. This problem can be solved by turning into a categorical feature considers four values (i.e., yellow, red, green and blue) into binary valued functions(like IsYellow?, IsRed?, IsGreen?, IsBlue?). Each function has a binary output of either 1 or 0 which represents Yes or No respectively.

Thus, in this way one can map the data set into feature vector by using the below mapping,

1. Copy the real-valued features directly
2. Consider the binary valued features as 0(False) or 1(True)
3. Map V-categorical features to V-many binary indicators.

Q24. Explain K-Nearest Neighbour algorithm with its merits and demerits.

Answer :

(Model Paper-I, Q13(b) | June/Aug.-22, Q13(b) [Old])

K-Nearest Neighbour Algorithm

The K-nearest neighbor algorithm is considered as a supervised machine learning algorithm. It is used for regression and classification. The working principle of this algorithm is to label a newly entered data point into a group by using any of the distance metrics.

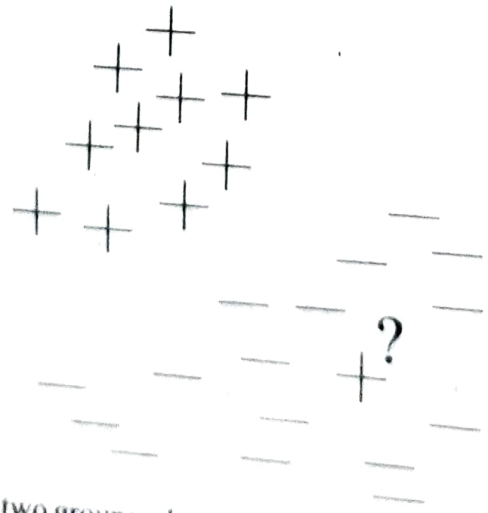
Generally, the examples in a high-dimensional space can be considered as vectors so that a variety of machine learning algorithms can be applied. For instance, the distance between two data points say (x_1, y_1) and (x_2, y_2) in a 2-Dimensional space can be measured by using the below formula,

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Whereas, the Euclidean distance between the vectors in a D-dimensional space can be measured by using the below formula,

$$d(a, b) = \left[\sum_{d=1}^{D} (a_d - b_d)^2\right]^{\frac{1}{2}}$$

Consider the below figure,



In the above figure, there are two groups where one represents the positive labels and another represents negative labels. There is also a newly entered data point represented by ?. Inorder to label the new data point, the K-NN concept is used. Consider any three data points near to the new data point. There are two negative labels and one positive label near the new entry point. As the probability is more for negative labels, the new point belongs to the negative labeled group.

Algorithm

The new terms used in the algorithm are as follows,

D represents the training dataset

N represents the examples of training samples

$(x_1, y_1),(x_2, y_2), \ldots ,(x_N, y_N)$ represents the example pairs

[ ] represents an empty set

$\oplus$.appends . to the list

$\hat{y}$ represents the prediction on $\hat{x}$.

KNN-Predict(D, K, $\hat{x}$)

$D \leftarrow [\,]$

for $n = 1$ to N do

$D \leftarrow D \oplus (d(x_n, \hat{x}), n)$ //store distance to training example n

end for

$D \leftarrow$ sort(D) //This arranges from lowest distance to highest

$\hat{y} \leftarrow 0$

for $k = 1$ to K do

(dist,n) $\leftarrow D_k$ //n this is the k^{th} closest data point

$\hat{y} \leftarrow \hat{y} + y_n$ //vote according to the label for the nth training point

end for

return sign($\hat{y}$) // return +1 if $\hat{y} > 0$ and −1 if $\hat{y} < 0$

The K-nearest neighbor algorithm is mostly used in various fields like image recognition, bioinformat and recommender systems. When it comes to large data sets, this algorithm is highly expensive and need good selection of hyper parameters.

Merits

1. It is very simple and easy to implement.
2. It can handle various problems of classification and regression.
3. It works well with representative data.
4. It is well suited for multiclass cases/problems.

Demerits

1. It utilizes high memory as it stores training data.
2. It is sensitive to outliers and irrelevant attributes.
3. Prediction rate is slow in case of large datasets.
4. It is expensive in terms of computation.

1.3.2 Decision Boundaries, k-means Clustering

Q25. Explain in detail about decision boundaries.

Answer :

Consider the below figure,

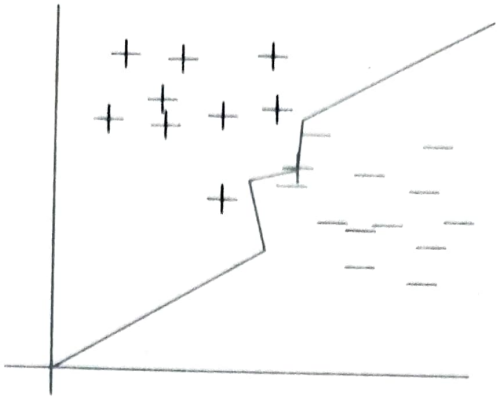


Figure: Decision Boundary for 1nn

In the above figure, the line that separates the + and – groups can be considered as the decision boundary.

The decision boundary is used to provide the visualization of the complexity of a learning model. Consider an example of decision trees for classifying *Ski* vs snowboard (sb) as shown in below figure,

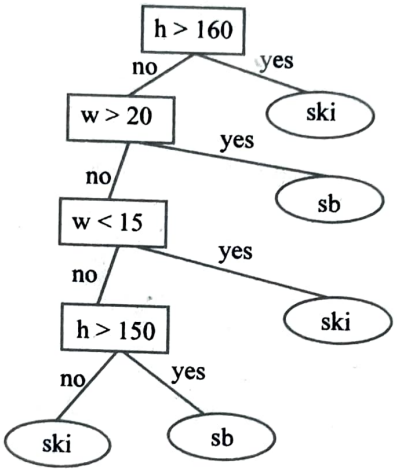


Figure: Decision Tree for Ski vs Snowboard

The decision boundary for the above figure is as shown in the below figure,

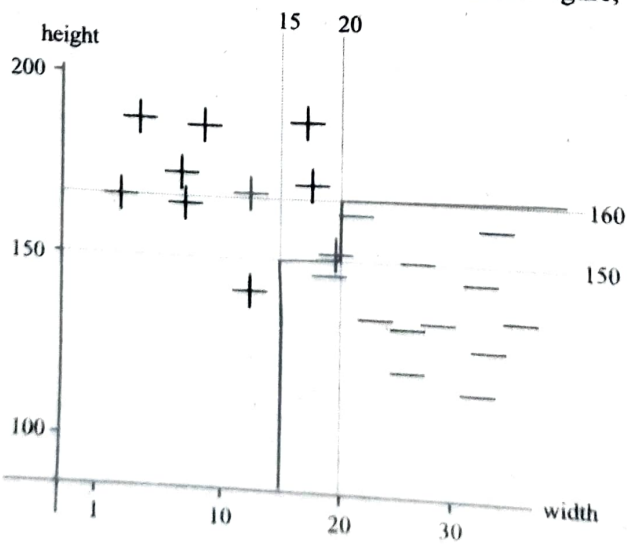


Figure: Decision Boundary for dt in Previous figure

Here, initially the space is divided based on the width of the first node. Then followed by its child. This way the tree can be classified. The decision boundaries for the given decision tree are axis-aligned cuts, shallow and very simple.

Q26. Explain K-Nearest Clustering and Decision Boundaries.

Answer :

(Model Paper-II, Q13(b) | July/Aug.-22, Q8 [MGU])

K-Nearest Clustering

For answer refer Unit-I, Page No. 16, Q.No. 24.

Decision Boundaries

For answer refer Unit-I, Page No. 17, Q.No. 25.

Q27. Discuss in detail about k-means clustering.

Answer :

The k-means is an iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is achieved. In this algorithm, the mean value of the objects in the cluster represents a cluster. It is the most popular and commonly used method. The algorithm is built on the concept of user specified input parameter (k). A set of 'n' objects are divided into 'k' clusters by the algorithm. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The cluster's centroid gives the measure of cluster's similarity.

Algorithm: K-Means(D, K)

for $k = 1$ to K do

  $\mu_k \leftarrow$ some random location // Initialize value to a random cluster

end for

repeat

  for $n = 1$ to N do

    $z_n \leftarrow \text{argmin}_k \|\mu_k - x_n\|$ // assign the n^{th} training sample to nearest center

  end for

  for $k = 1$ to K do

    $X_k \leftarrow \{ x_n : z_n = k \}$ // points assigned to cluster k

    $\mu_k \leftarrow \text{mean}(X_k)$ // re-estimate center of cluster k

  end for

until μs stop changing

return z // return cluster assignments

In the above algorithm, initially select 'k' objects randomly from D, as initial cluster centers. Depending upon the distance between the object and the cluster mean, each remaining object is assigned to the cluster to which it is most similar or near. Then, calculate new mean value of the object for each cluster. Repeat the same process until the centers converge.

1.3.3 High Dimensions

Q28. Explain why high dimensions are scary.

Answer :

The main reason behind being scared about high dimensions is the visualization. According to a study, the human can visualize till 4-dimensional space but not more than that. Additionally, there are two other issues while considering high dimensions collectively known as the curse of dimensionality. They are as follows,

1. Computational Difficulties

Computational difficulties arise when the prediction is difficult. For instance, consider the K-Nearest Neighbor concept. In this, the computational speed of prediction is very slow as it needs to look at all the training samples every time when the user wants to do a prediction for a data point. This issue can be achieved by creating the index data structure to the samples. And, split the data samples into grids of similar features so that one can directly check that particular grid training samples instead of each training sample.

But, again this technique will only apply till certain dimensional space. For instance, a 2-dimensional space can be split into 25 grid cells(i.e., 5*5). 3-dimensional space can be split into 125 grid cells(i.e., 5*5*5). Similarly, 20-dimensional space can be split into 95,367,431, 640,625 grid cells, which means only 95 trillion training data samples can be accepted. Thus, for high dimensions it is very hard to make predictions and the computations really lead to many problems.

2. Mathematical Difficulties

Mathematical difficulties arise as assumptions made while working with 2-dimensional and 3-dimensional spaces do not work well with high dimensional spaces.

For example, in a 2-dimensional space there are four spheres(with radius 1) where each one is adjacent to the other two spheres. It is as shown in the below figure,

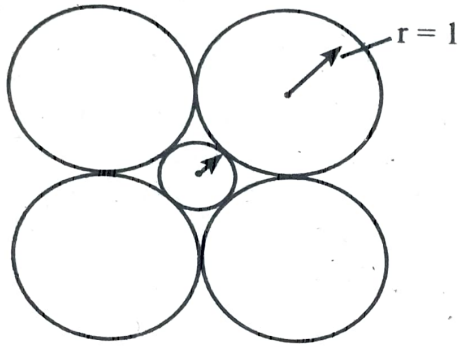


Figure: 2D Spheres in Spheres

The radius of the middle sphere can be easily calculated by using the pythagorean theorem.

$$1^2 + 1^2 = (1 + r)^2$$

So,

$$r = \sqrt{2} - 1.$$

Similarly for,

3-dimensional space, $r = \sqrt{3} - 1$.

4-dimensional space, $r = \sqrt{4} - 1$.

5-dimensional space, $r = \sqrt{5} - 1$.

D-dimensional space, $r = \sqrt{D} - 1$.

Therefore, as the dimensional space increases, the radius of the inner sphere will be increased and joins with the outer spheres.

Another issue is that the distance between the points in high dimensions increases as the dimensional space increases.

Multiple Choice

1. _____ can be defined as the challenges that occur while researchers are working with machine learning. [ ]

   (a) Learning problems (b) Canonical problems

   (c) Prediction problems (d) Decision problems

2. In decision tree, the questions are provided in rectangles called as _____. [ ]

   (a) Child nodes (b) Extreme nodes

   (c) Internal nodes (d) External nodes

3. The information provided in the datasets can be quantified by using _____. [ ]

   (a) Entropy (b) Probability

   (c) Regression (d) Classification

4. Naive Bayesian classifier is also called as _____. [ ]

   (a) Bayes classifier (b) Simple Bayesian classifier

   (c) Both (a) and (b) (d) None of these

5. _____ does not allow the user to query over a predefined depth d in decision tree learning algorithm. [ ]

   (a) Whole tree (b) Narrow decision tree

   (c) Shallow decision tree (d) Complex tree

6. _____ can be defined as the values that can be used within a model to estimate a decision. [ ]

   (a) Model (b) Hyperparameters

   (c) Variants (d) Parameters

7. Hyperparameters are also called as _____. [ ]

   (a) Models (b) Hypothesis

   (c) Knobs (d) None of these

8. _____ algorithm is used in fields like image recognition bioinformatics. [ ]

   (a) K-NN (b) K-means

   (c) Bayes (d) EM

9. _____ algorithm is built on the concept of user specified input parameter (k). [ ]

   (a) K-NN (b) K-means

   (c) Bayes (d) EM

10. For 3-dimensional space, the radius of the middle sphere (r) can be given as _____. [ ]

    (a) $\sqrt{2} - 1$ (b) $1 - \sqrt{3}$

    (c) $\sqrt{3r} - 1$ (d) $\sqrt{3} - 1$

II. Fill in the Blanks

1. _____ is the concept that computer program can learn and adapt to new data without intervention of human.
2. _____ tries to group a particular set of values in the required order of relevance.
3. Decision tree model of learning is similar to _____ concept.
4. _____ algorithm divides the data into various subsets depending on the features.
5. _____ can be defined as a set of assumptions which can justify classifications assigned to future instances.
6. _____ can be defined as a concept where there is a possibility of learning something.
7. _____ is an approach used to obtain the various estimating algorithms in machine learning.
8. The feature values can be interpreted by using _____ digits.
9. _____ is used to provide the visualization of the complexity of a learning model.
10. _____ difficulties aries while working with 2D and 3D spaces if they do not work with high dimensional space.

KEY

I. Multiple Choice

1. (b)
2. (c)
3. (a)
4. (b)
5. (c)
6. (d)
7. (c)
8. (a)
9. (b)
10. (d)

II. Fill in the Blanks

1. Machine learning
2. Ranking
3. Divide and Conquer
4. ID3
5. Inductive Bias
6. Underfitting
7. Model
8. Binary
9. Decision boundary
10. Mathematical

III. Very Short Questions and Answers

Q1. Define Canonical problem.

Answer :

Canonical problems can be defined as the challenges that occur while researchers are working with the concept of machine learning.

Q2. Define bayes optimal classifier.

Answer :

Bayes optimal classifier can be defined as a classifier that takes $\hat{x}$ as input and returns $\hat{y}$ which enlarges the COMPUTED($\hat{x}$, $\hat{y}$).

Q3. Define overfitting.

Answer :

Overfitting can be defined as a concept where the focus is more on the training data but not able to summarise properly.

Q4. What is a model?

Answer :

A model can be defined as an approach that is used to obtain the various estimating algorithms in machine learning. It can also be defined as a mathematical expression representing the relationship between the given input data and result.

Q5. Write about K-Nearest neighbor.

Answer :

The K-nearest neighbor algorithm is considered as a supervised machine learning algorithm. It is used for regression and classification. The working principle of this algorithm is to label a newly entered data point into a group by using any of the distance metrics.

SHORT QUESTIONS

Q1. What is learning? Explain about different components of learning process.

Answer : (Important Question | June/Aug.-22, Q1 [OU])

For answer refer Unit-I, Page No. 2, Q.No. 2.

Q2. Explain about binary and multiclass classification in detail.

Answer : (Important Question | June/Aug.-22, Q2 [OU])

For answer refer Unit-I, Page No. 2, Q.No. 3.

Q3. Write in brief about Decision Tree Model of Learning.

Answer : (Important Question | July/Aug.-22, Q1 [MGU])

For answer refer Unit-I, Page No. 3, Q.No. 4.

Q4. Define Inductive bias.

Answer : (Important Question | June/Aug.-22, Q3 [OU])

For answer refer Unit-I, Page No. 4, Q.No. 7.

Q5. Define feature vectors and decision boundary.

Answer : Important Question

For answer refer Unit-I, Page No. 4, Q.No. 9.

ESSAY QUESTIONS

Q6. Define Machine Learning. Write about some of the canonical learning problems.

Answer : Important Question

For answer refer Unit-I, Page No. 5, Q.No. 11.

Q7. Explain in detail about the decision tree model of learning.

Answer : Important Question

For answer refer Unit-I, Page No. 5, Q.No. 12.

Q8. Explain how to formalizing the learning problem and ID3 Algorithm with example.

Answer : (Important Question | July/Aug.-22, Q7 [MGU])

For answer refer Unit-I, Page No. 11, Q.No. 15.

Q9. Explain about inductive bias.

Answer : Important Question

For answer refer Unit-I, Page No. 12, Q.No. 17.

Q10. Write about underfitting and overfitting.

Answer : Important Question

For answer refer Unit-I, Page No. 13, Q.No. 19

Q11. Explain about separation of training and testing data.

Answer : Important Question

For answer refer Unit-I, Page No. 14, Q.No. 20

Q12. What are the sequence of decisions that must be made to deploy a machine learning approach in the real world.

Answer : Important Question

For answer refer Unit-I, Page No. 15, Q.No. 22

Q13. Explain K-Nearest Neighbour algorithm with its merits and demerits.

Answer : (Important Question | June/Aug.-22, Q13(b) [OU])

For answer refer Unit-I, Page No. 16, Q.No. 24

Q14. Explain in detail about decision boundaries.

Answer : Important Question

For answer refer Unit-I, Page No. 17, Q.No. 25

Q15. Explain why high dimensions are scary

Answer : Important Question

For answer refer Unit-I, Page No. 19, Q.No. 28