SET-1

Code No.: 20-CS–PC-322 | R20 | H.T.No. | | | R | 0 | | | | | |

CMR INSTITUTE OF TECHNOLOGY: HYDERABAD
UGC AUTONOMOUS
III–B.Tech.II–Semester-I- Mid Term Examinations – MAR – 2024
MACHINE LEARNING AND DATA SCIENCES
(CSE,CSE(AI&ML),CSE(DS),AI&ML,AI&DS)

[Time: 90 Minutes] **[Max. Marks: 25]**

Note: 1.This question paper contains two parts A and B.
2. Part A is compulsory which carries 10 marks. Answer all questions in Part A.
3. Part B consists of 3 questions. Answer all 3 questions. Each question carries 5 marks and may have sub questions.
4. Illustrate your answers with NEAT sketches wherever necessary.

PART-A **5X2M=10M**

1 a) Relate the concepts of correlation and causation.

Correlation means there is a relationship or pattern between the values of two variables. A scatter plot displays data about two variables as a set of points in the xy-plane and is a useful tool for determining if there is a correlation between the variables.

There are three possible results :

1. Positive Correlation

2. Negative Correlation

3. No correlation

Causation means that one event causes another event to occur. Causation can only be determined from an appropriately designed experiment causation is a term that is used to indicate a causal relationship between two variables; one variable is dependent on the other. In a causal relationship, there is an independent variable ('the cause') and a dependent variable ('the effect')

1 b) Simplify vector in machine learning. What are the different operation performed on them.

A vector is a tuple of one or more values called scalars. Vectors are built from components, which are ordinary numbers. You can think of a vector as a list of numbers, and vector algebra as operations performed on the numbers in the list. Vectors are often represented using a lowercase character such as "v"; for example:

$$v = (v1, v2, v3)$$

different operations are :

- Vector Arithmetic
- Vector Subtraction
- Vector Multiplication
- Vector Division
- Vector Dot Product
- Vector-Scalar Multiplication

1 C) How do you find K-NN with example ?

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.
Example: Suppose, we have an image of a creature that lookssimilar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure.

1 d) Compare and contrast classification and regression.

Regression Algorithm	**Classification Algorithm**
• In Regression, the output is a continuous or numerical value.	• In Classification, the output is a discrete or categorical value.
• Regression model maps the input variable(x) with the continuous output variable(y).	• Classification model maps the input variable(x) with the discrete output variable(y).
• In Regression, we find the best fit line that can predict the output accurately.	• In Classification, we find the decision boundary that can divide the dataset into different classes.
• Regression algorithms solve regression problems such as house price prediction, cryptocurrency price prediction, etc.	• Classification algorithms solve classification problems such as face detection, speech recognition, etc.
• Regression algorithms can be further divided into Linear and Non-linear Regression.	• Classification algorithms can be divided into Binary classifiers and Multi-class classifiers.

1e) Why neural networks are used in DEEP learning?

- Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks
- Deep-learning networks are distinguished from the more common place single hidden-layer neural networks by their depth; that is, the number of node layers through which data must pass in a multistep process of pattern recognition.
- Earlier versions of neural networks such as the first perceptions were shallow, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning
- In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output.

PART-B **3X5M = 15M**

2 I) How do you find the distribution of a continuous variable ?

A probability distribution in which the random variable X can take on any value (is continuous). Because there are infinite values that X could assume, the probability of X taking on any one specific value is zero. Therefore we often speak in ranges of values (p(X>0) = .50). The normal distribution is one example of a continuous distribution. The probability that X falls between two values (a and b) equals the integral (area under the curve) from a to b:

Probability Density Function

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

- A probability distribution is formed from all possible outcomes of a random process (for a random variable X) and the probability associated with each outcome. Probability distributions may either be discrete (distinct/separate outcomes, such as number of children) or continuous (a continuum of outcomes, such as height).

- A probability density function is defined such that the likelihood of a value of X between a and b equals the integral (area under the curve) between a and b. This probability is always positive. Further, we know that the area under the curve from negative infinity to positive infinity is one.

- The normal probability distribution, one of the fundamental continuous distributions of statistics, is actually a family of distributions (an infinite number of distributions with differing means (μ) and standard deviations (σ). Because the normal distribution is a continuous distribution, we can not calculate exact probability for an outcome, but instead we calculate a probability for a range of outcomes (for example the probability that a random variable X is greater than 10).

Example 1: (Discrete Outcomes)

- Coin flip: $\Omega = \{H, T\}$.
- Throw a dice: $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.
- Paper / scissor / stone: $\Omega = \{\text{paper}, \text{scissor}, \text{stone}\}$.
- Draw an even integer: $\Omega = \{2, 4, 6, 8, \ldots\}$.

In the last example, we see that a sample space can be infinite.

Example 2: (Continuous Outcomes)

- Waiting time for a bus in West Lafayette: $\Omega = \{t \mid 0 \leq t \leq 30 \text{ minutes}\}$.
- Phase angle of a voltage: $\Omega = \{\theta \mid 0 \leq \theta \leq 2\pi\}$.
- Frequency of a pitch: $\Omega = \{f \mid 0 \leq f \leq f_{\max}\}$.

Practice Exercise. There are 8 processors on a computer. A computer job scheduler chooses one processor randomly. What is the sample space? If the computer job schedular can choose two processors at once, what is the sample space then?

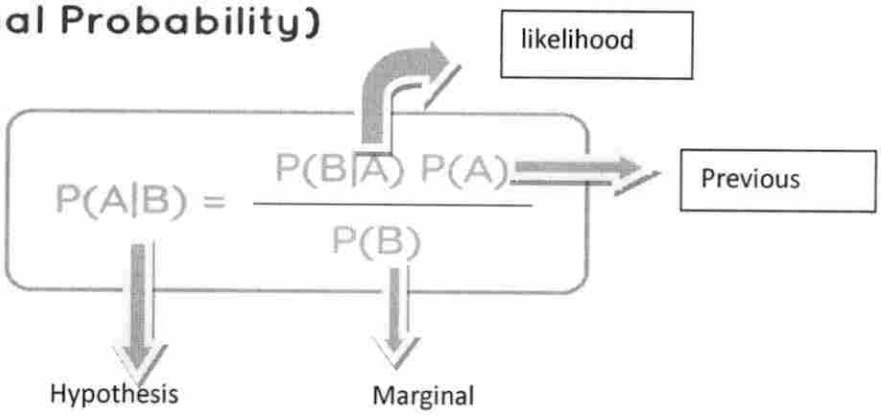
Solution. The sample space of the first case is $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The sample space of the second case is $\Omega = \{(1, 2), (1, 3), (1, 4), \ldots, (7, 8)\}$.

2 ii) What is Bayes Theorem ? Explain with examples ?

- Bayes theorem is derived using definition of conditional probability . The bayes theorem formula includes two conditional probabilities Determines the conditional probability of an event A given that event B has already occurred. Bayes theorem is also known as the Bayes Rule or Bayes Law

- Bayes theorm gives probability of an event based on prior knowledge of condition

Bayes' Formula (Conditional Probability)



- An event B can be explained by a set of exhaustive and mutually exclusive hypothesis A1,A2,….,An.
- Given 'a prior' probabilities P(A1),P(A2),…,P(An) corresponding to a total absence of knowledge regarding the occurrence of B and conditional probabilities
- P(B/A1),P(B/A2),…,P(B/An)
- the 'a posterior' probability P(A / B) j of some event Aj is given by

$$P(A_j / B) = \frac{P(A_j).P(B / A_j)}{\sum_{i=1}^{n} P(A_i)P(B / A_i)}$$

(OR)

3.) What is the relationship between hypothesis testing and confident interval? Is confidence interval a statistical inference ? justify your answer .

- **hypothesis testing** uses point estimate to decide which of two hypotheses (guesses) about parameter is correct.
- Suppose X continuous from a population with mean μ and standard deviation σ.

What is the value of μ?

- We select a random sample from that population and try to make inference about μ.

Null hypothesis (H0): – An explicit statement about an unknown parameter the validity of which you wish to test, e.g., μ=μ0

- Alternative hypothesis (H1): – An alternative statement about the unknown parameter used to compare your null with, e.g., • μ ¹ μ0 (two-sided test)

  μ < μ0 (one-sided test)

  μ > μ0 (one-sided test)

Confidence Interval (CI): the range with the most probable values for the unknown parameter with a (1-α)% level of confidence Confidence intervals and hypothesis testing are closely related because both methods use the same underlying methodology. Additionally, there is a close connection between significance levels and confidence levels. Indeed, there is such a strong link between them that hypothesis tests and the corresponding confidence intervals always agree about statistical significance.

The relationship between the confidence level and the significance level for a hypothesis test is as follows:

Confidence level = 1 – Significance level (alpha)

For example, if your significance level is 0.05, the equivalent confidence level is 95%.

Both of the following conditions represent statistically significant results:

- The P-value in a hypothesis test is smaller than the significance level.
- The confidence interval excludes the null hypothesis value.

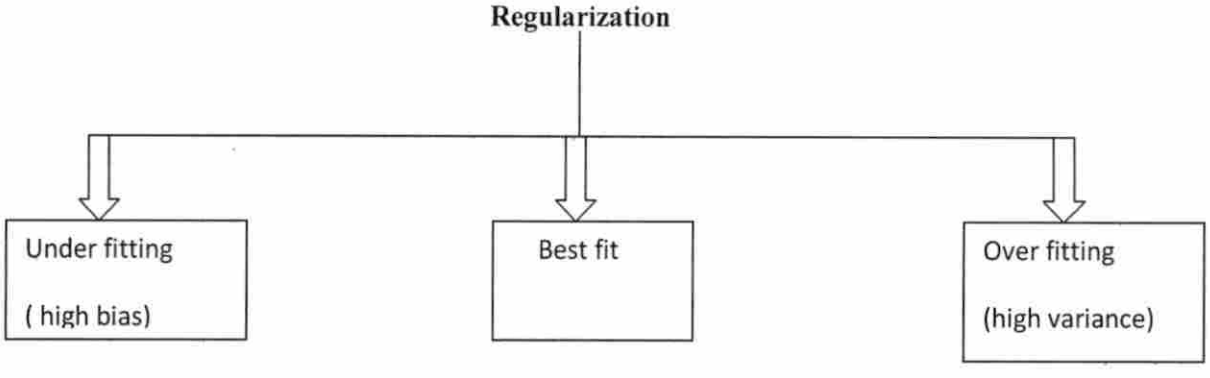
Further, it is always true that when the P-value is less than your significance level, the interval excludes the value of the null hypothesis.

Types of hypothesis testing:

1. T Test
2. Z Test
3. F- Test
4. ANOVA
5. Chi-Square Test

4) What do you mean by regularization ? Explain the different regularization methods ?

- This model need to find optimal point data model. the purpose for this model find best fit
- Reduce the complexity & apply cost function to control the parameter try to fit the datapoint in Regression line
-



- This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + ...+ \beta pXp$$

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

If there is noise in the training data, then the estimated coefficients won't generalize well to the future data.

Ridge Regression

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Above image shows ridge regression, where the RSS is modified by adding the shrinkage quantity. Now, the coefficients are estimated by minimizing this function. Here, λ is the tuning parameter that decides how much we want to penalize the flexibility of our model
- This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept β0, This intercept is a measure of the mean value of the response when xi1 = xi2 = …= xip = 0.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}.$$

Lasso

- Lasso is another variation, in which the above function is minimized. Its clear that this variation differs from ridge regression only in penalizing the high coefficients. It uses |βj|(modulus)instead of squares of β, as its penalty. In statistics, this is known as the L1 norm.
- The ridge regression can be thought of as solving an equation, where summation of squares of coefficients is less than or equal to s. And the Lasso can be thought of as an equation where summation of modulus of coefficients is less than or equal to s. Here, s is a constant that exists for each value of shrinkage factor λ. These equations are also referred to as constraint functions.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Elastic Net Regression

- Linear regression refers to a model that assumes a linear relationship between input variables and the target variable. With a single input variable, this relationship is a line, and with higher dimensions, this relationship can be thought of as a hyperplane that connects the input variables to the target variable.
- The coefficients of the model are found via an optimization process that seeks to minimize the sum squared error between the predictions (yhat) and the expected target values (y).

loss = sum i=0 to n (y_i – yhat_i)^2

- This is particularly true for problems with few observations (samples) or more samples (n) than input predictors (p) or variables (so-called p >> n problems).
- One popular penalty is to penalize a model based on the sum of the squared coefficient values. This is called an L2 penalty.

l2_penalty = sum j=0 to p beta_j^2

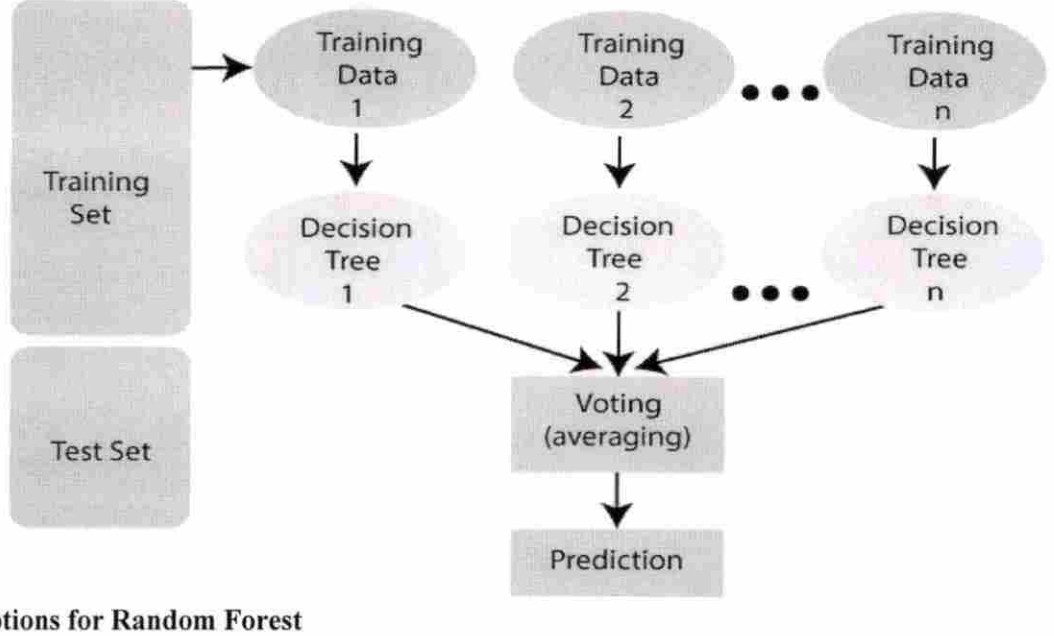
- Another popular penalty is to penalize a model based on the sum of the absolute coefficient values. This is called the L1 penalty.

l1_penalty = sum j=0 to p abs(beta_j)

(OR)

5) Explain Random Forest algorithm with example .

- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset".
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting



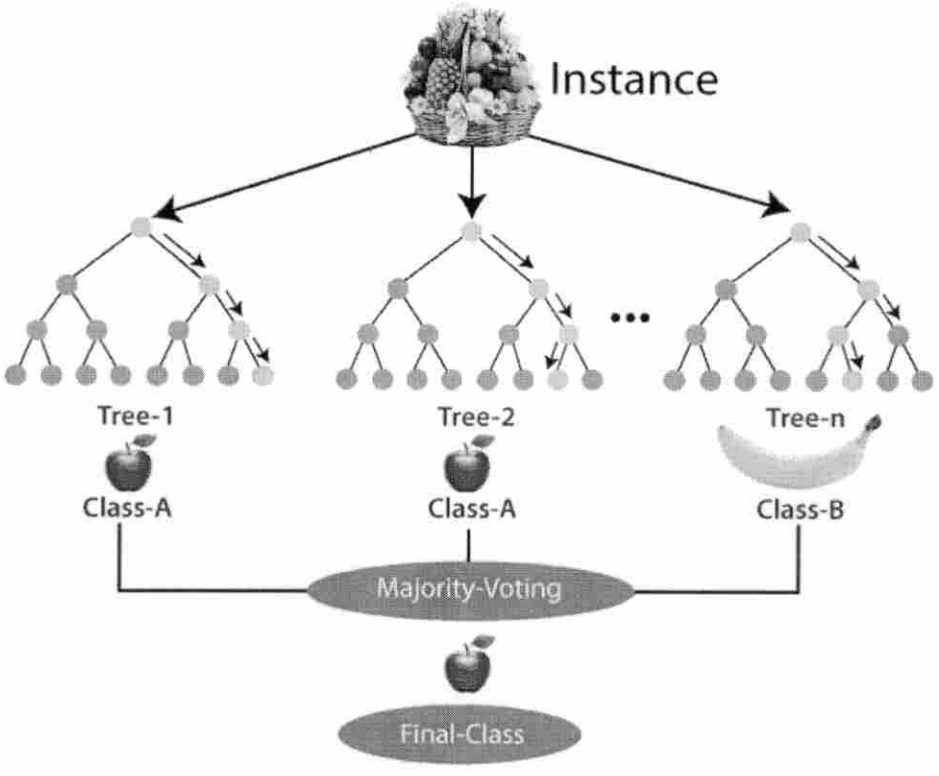
Assumptions for Random Forest

- Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not.
- Therefore, below are two assumptions for a better Random forest classifier:
- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.
- Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

Working process

- Step-1: Select random K data points from the training set.
- Step-2: Build the decision trees associated with the selected data points (Subsets).
- Step-3: Choose the number N for decision trees that you want to build.
- Step-4: Repeat Step 1 & 2.
- Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:



Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensional
- It enhances the accuracy of the model and prevents the over fitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks

6) Demonstrate Find-S algorithm. Mention the difference between general hypothesis and specific hypothesis with example .

- The find-S algorithm is a basic concept learning algorithm in machine learning.
- The find-S algorithm finds the most specific hypothesis that fits all the positive examples.
- We have to note here that the algorithm considers only those positive training example.
- The find-S algorithm starts with the most specific hypothesis and generalizes this hypothesis each time it fails to classify an observed positive training data.
- Hence, the Find-S algorithm moves from the most specific hypothesis to the most general hypothesis.

Representation :

- ? indicates that any value is acceptable for the attribute.
- specify a single required value (e.g., Cold) for the attribute.
- Φ indicates that no value is acceptable.
- The most general hypothesis is represented by: {?, ?, ?, ?, ?, ?}
- The most specific hypothesis is represented by: {ϕ, ϕ, ϕ, ϕ, ϕ, ϕ}

Steps Involved In Find-S:

- Start with the most specific hypothesis. h = {ϕ, ϕ, ϕ, ϕ, ϕ, ϕ}
- Take the next example and if it is negative, then no changes occur to the hypothesis.
- If the example is positive and we find that our initial hypothesis is too specific then we update our current hypothesis to a general condition.
- Keep repeating the above steps till all the training examples are complete.
- After we have completed all the training examples we will have the final hypothesis when can use to classify the new examples.

EXAMPLE	COLOR	TOUGHNESS	FUNGUS	APPEARANCE	POISONOUS
1.	GREEN	HARD	NO	WRINKELD	YES
2.	GREEN	HARD	YES	SMOOTH	NO
3.	BROWN	SOFT	NO	WRINKLED	NO
4.	ORANGE	HARD	NO	WRINKLED	YES
5.	GREEN	SOFT	YES	SMOOTH	YES
6.	GREEN	HARD	YES	WRINKLED	YES
7.	ORANGE	HARD	NO	WRINKLED	YES

ATTRIBUTES ON WHICH THE CONCEPT DEPENDS ON **CONCEPT**

First, we consider the hypothesis to be a more specific hypothesis. Hence, our hypothesis would be : h = {ϕ, ϕ, ϕ, ϕ, ϕ, ϕ}

Consider example 1 : The data in example 1 is { GREEN, HARD, NO, WRINKLED }. We see that our initial hypothesis is more specific and we have to generalize it for this example. Hence, the hypothesis becomes **:**

h = { GREEN, HARD, NO, WRINKLED }

Consider example 2 : Here we see that this example has a negative outcome. Hence we neglect this example and our hypothesis remains the same.

h = { GREEN, HARD, NO, WRINKLED }

Consider example 3 : Here we see that this example has a negative outcome. Hence we neglect this example and our hypothesis remains the same.

h = { GREEN, HARD, NO, WRINKLED }

Consider example 4 : The data present in example 4 is { ORANGE, HARD, NO, WRINKLED }. We compare every single attribute with the initial data and if any mismatch is found we replace that particular attribute with a general case (" ? "). After doing the process the hypothesis becomes :

h = { ?, HARD, NO, WRINKLED }

Consider example 5 : The data present in example 5 is { GREEN, SOFT, YES, SMOOTH }. We compare every single attribute with the initial data and if any mismatch is found we replace that particular attribute with a general case (" ? "). After doing the process the hypothesis becomes : h = { ?, ?, ?, ? }

Final Hyposthesis: h = { ?, ?, ?, ? }

(OR)

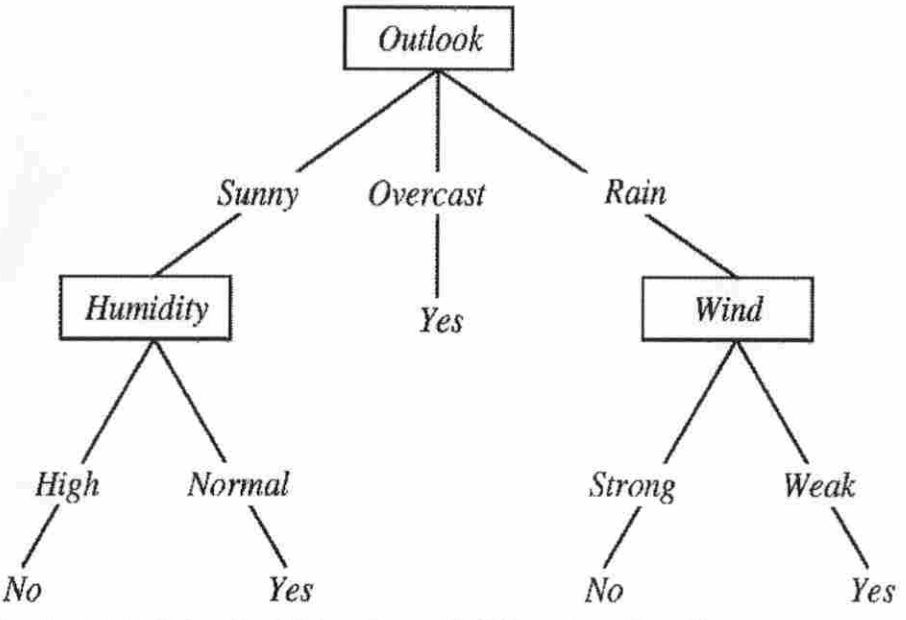
7) Explain the process of learning by Rule Induction with an example ?

- Rule induction is one of the most important techniques of machine learning.
- Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data mining at the same time.
- Usually rules are expressions of the form

if (attribute − 1, value − 1)
and (attribute − 2, value − 2)
and ⋯
and (attribute − n, value − n)
then (decision, value)

EX: Rule from Desicion tree

- R1: **IF** Outlook=Sunny **AND** Humidity=High **THEN** Play=NO
- R2: **IF** Outlook=Sunny **AND** Humidity=Normal **THEN** Play=NO
- R3: **IF** Outlook = Overcast **THEN** Play=Yes
- R2: **IF** Outlook=rain **AND** Wind=Strong **THEN** Play=NO
- R2: **IF** Outlook=Rain **AND** Wind =Weak **THEN** Play=Yes



Rule Induction Rule Induction Using Sequential Covering Algorithm:

- Sequential Covering Algorithm can be used to extract IF-THEN rules form the training data.
- The sequential learning Algorithm where rules are learned for one class at a time.
- When learning a rule from a class Ci, we want the rule to cover all the tuples from class C only and no tuple form any other class.

Algorithm: Sequential Covering

Input:

- D, a data set class-labeled tuples,
- Att_vals, the set of all attributes and their possible values.
- Output: A Set of IF-THEN rules.

Method:

- Rule_set={ }; // initial set of rules learned is empty
- for each class c do
- repeat
- Rule = Learn_One_Rule(D, Att_valls, c);
- remove tuples covered by Rule form D;
- until termination condition;
- Rule_set=Rule_set+Rule; // add a new rule to rule-set
- end for
- return Rule_Set;

1. Dr. S. Alagumuthukrishnan CSE
2. J. Vasantha priyanka CSE(DS) 23/3/24
3. Indumathi. V (CSE-DS) 23/3/24.
4. Tahseen Jahan (CSE(DS)) 23-2-24
5. V. Devi Priya (CSE) 23/3/2024
6. B. Sindhuja (CSE) 23/3/2024
7. A. Madhavi
8. S Malli Babu