MCKV Institute of Engineering

Paper Code: PE-IT501B

Machine Learning

Time Allotted: 1 Hour *Full Marks:* 30

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternatives for any ***five*** of the following: 5×1

i. Which of the following is a predictive model?
   A. Clustering
   B. Regression
   C. Summarization
   D. Association rules

ii. You have a dataset of different flowers containing their petal.lengths and color. Your model has to predict the type of flower for given petal lengths and color. This is a-
   A. Regression task
   B. Classification task
   C. Clustering task
   D. None

iii. With the help of a confusion matrix, we can compute-
   A. Recall
   B. Precision
   C. Accuracy
   D. All of the above

iv. Which of the following is a lazy learning algorithm?
   A. SVM
   B. KNN
   C. Decision tree
   D. All of the above

v. What does 'k' stand for in the KNN algorithm?
   A. Number of neighbors
   B. Number of output classes
   C. Number of input features

D. None

vi. How does a decision tree work?

A. Minimizes the information gain and maximizes the entropy

B. Maximizes the information gain and minimizes the entropy

C. Minimizes the information gain and minimizes the entropy

D. Maximizes the information gain and maximizes the entropy

Group – B

(Short Answer Type Questions)

Answer any ***two*** of the following 2×5

2. Explain the various stages involved in designing a learning system.

3. Differentiate between Supervised, Unsupervised and Reinforcement Learning with example.

4. Discuss the principle of Naïve Bayes algorithm.

Group – C

(Long Answer Type Questions)

Answer any ***one*** of the following 1×15

5+10=15

5. a. Explain KNN classification Algorithm step by step.

b. A training dataset (Dataset1) related to tissue paper quality has been given.**Find out the class label of the special test paper tissue with parameter values (X1=3 and X2=7).** Solve the problem using KNN Algorithm. Any distance metric can be used of your choice.

Dataset1

Objects	X1 (Acid Durability)	X2(Strength) Kg/square meter	Y(Class)
1	7	7	Bad
2	7	4	Bad
3	3	4	Good
4	1	4	Good

5+10=15

6. a. Discuss the characteristics of Decision Tree classification algorithm.

b. Find out the information gain of the attributes A1 and A2 for the below dataset.

Objects	A1	A2	Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

MCKV Institute of Engineering

Paper Code: PE-IT501B

Machine Learning

Time Allotted: **3 Hours** ***Full Marks:*** **70**

The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable.

Group – A

(Multiple Choice Type Questions)

1.Choose the correct alternatives for any ***ten*** of the following: 10×1=10

i) FIND-S algorithm ignores?

a) Positive b) Negative c) Both d) None

ii) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Here feature type is_____

a) Ordinal b) Categorical c) Boolean d) None

iii) With the help of a confusion matrix, we can compute-

a) Recall b) Precision c) Accuracy d) All of these

iv) You are given reviews of few Netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of_____

a) unsupervised learning b) semi supervised learning

c) supervised learning d) reinforcement learning

v) Which learning requires Self-Assessment to identify patterns within data?

a) supervised learning b) unsupervised learning

c) semi supervised learning d) reinforced learning

vi. Some telecommunication company wants to segment their customers into distinct groups, this is an example of_____

a) supervised learning b) unsupervised learning

c) data extraction d) reinforcement learning

vii) How does a decision tree work?

a) Minimizes the information gain and maximizes the entropy

b) Maximizes the information gain and minimizes the entropy

c) Minimizes the information gain and minimizes the entropy

d) Maximizes the information gain and maximizes the entropy

viii) Which of the following is required by K-means clustering?

a)defined distance metric b) number of clusters

c) initial guess as to cluster centroids d) all of the mentioned

ix) Which of the following clustering requires merging approach?

a) Partitional b) Hierarchical c) Naive Bayes d) None of the mentioned

x) Which of the following machine learning algorithm is based upon the idea of bagging?

a) Random Forest b) Decision tree c) Classification d) Regression

xi)Which of the following is a performance measure for regression?

a) Accuracy b) Recall c) MSE d) Error rate

xii)The probability that a particular hypothesis holds for a data set based on the Prior is called

a) Independent probabilities b) Posterior probabilities

c) Interior probabilities d) Dependent probabilities

Group – B

(Short Answer Type Questions)

Answer any ***three*** of the following 3×5=15

2. a) What is supervised learning? Give any two examples.

b) What is over-fitting? When does it happen? (1+1)+(1+2)

3.a) "KNN may works as a 'Regression' problem not only 'Classification' problem" – Justify with proper example.

b) How we can choose a good value for k, the number of neighbors in KNN algorithm?

c) Why KNN algorithm called lazy learner? (2+2+1)

4. What are the roles of the Activation functions in Neural Networks and name two popular activation function. 5

5. Explain Semi-supervised learning and reinforcement learning with examples. 5

6. Discuss the differences between k-means and weighted k- means algorithm. 5

Group – C

(Long Answer Type Questions)

Answer any ***three*** of the following 3×15=45

7. a) Explain the impact of features on machine learning.

b) What do you meant by least square method? Explain least square method in the context of linear regression.

c) "Support Vector Machine is maximum margin classifier" - Comment and Justify correctness or incorrectness of the statement.

d) Explain how performance of learning models can be increased in Bagging? 3+5+2+5

8. a) Describe the Find-S algorithm. Explain by taking Enjoy Sport concept and training instances given below.

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

b) Define Consistent Hypothesis and Version Space. (4+6)+5

9. a) Draw the decision tree using ID 3 algorithm over the following dataset:

Gender	Car ownership	Travel cost	Income Level	Transportation (Class)
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

b) Write down the differences between Bagging and Boosting Techniques. 10+5

10.a) Show the final result of hierarchical clustering (using Agglomerative Clustering approach) with single link by drawing a Dendrogram.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

b) Write down the differences between hard clustering and soft clustering with examples.

c) What is hyper plane? 6+6+3

11.a) What is meant by ensembling?

b) What are the advantages and disadvantages of ensembling learning?

c) Compare bagging and boosting.

d) How random forest ensembling model works? 2+ 4+ 3+6

MCKV Institute of Engineering

Paper Code: PE-IT501B

Machine Learning

Time Allotted: 1 Hour *Full Marks:* 30

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternatives for any ***five*** of the following: 5×1

i. Which of the following is required by K-means clustering?
a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) all of the mentioned

ii. Which of the following combination is incorrect?
a) Continuous – euclidean distance
b) Continuous – correlation similarity
c) Binary – manhattan distance
d) None of the mentioned

iii. Which of the following clustering requires merging approach?
a) Partitional
b) Hierarchical
c) Naive Bayes
d) None of the mentioned

iv. What is the minimum no. of variables/ features required to perform clustering?
a) 0
b) 1
c) 2
d) 3

v. Which of the following algorithm is most sensitive to outliers?
a) K-means clustering algorithm
b) K-medians clustering algorithm
c) K-modes clustering algorithm
d) K-medoids clustering algorithm

vi. The most widely used metrics and tools to assess a classification model are:
a) Confusion matrix

b) Cost-sensitive accuracy
c) Area under the ROC curve
d) All of the above

Group – B

(Short Answer Type Questions)

Answer any ***two*** of the following 2×5

2. Compare K-means clustering with Hierarchical Clustering Techniques.

3. Explain K-Medoids clustering algorithm with an example.

4. Distinguish between overfitting and underfitting. How it can affect model generalization?

Group – C

(Long Answer Type Questions)

Answer any ***one*** of the following 1×15

15

5. Illustrate K- means clustering algorithm with an example.

15

6. Show the final result of hierarchical clustering (Use Agglomerative Approach) with complete link by drawing a Dendrogram.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

MCKV Institute of Engineering

Paper Code: PE-IT501B

Paper Name : Machine Learning

Time Allotted: 1 Hour ***Full Marks: 30***

The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable.

Group - A

(Multiple Choice Type Questions)

1. Choose the correct alternatives for any ***five*** of the following: 5×1=5

i. Data used to build a data mining model on_____.
(a) validation data
(b) training data
(c) test data
(d) hidden data

ii. Machine learning techniques differ from statistical techniques in that machine learning methods
(a) typically assume an underlying distribution for the data.
(b) are better able to deal with missing and noisy data.
(c) are not able to explain their behavior.
(d) have trouble with large-sized datasets.

iii. Classification problems are distinguished from estimation problems in that
(a) classification problems require the output attribute to be numeric.
(b) classification problems require the output attribute to be categorical.
(c) classification problems do not allow an output attribute.
(d) classification problems are designed to predict future outcome.

iv. Another name for an output attribute.
(a) predictive variable
(b) independent variable
(c) estimated variable
(d) dependent variable

v. Analysis of ML algorithm needs
(a) Statistical learning theory
(b) Computational learning theory
(c) Both of above
(d) None of above

vi. FIND-S algorithm ignores which data?
(a) Positive
(b) Negative
(c) Both
(d) None

vii. Foot size of human, for a ML analysis, is a/an

(a) Nominal variable
(b) Ordinal variable
(c) Discrete variable
(d) Continuous variable

Group - B

(Short Answer Type Questions)

Answer any ***two*** of the following 2×5=10

2. Define null hypothesis. What is the alternate hypothesis? What is p-value? [2+2+1]
[Module2/CO1/Understand-IOCQ]

3. What is over fitting? How it can be avoided? [1+4] [Module1/CO1/Understand-IOCQ]

4. What is hypothesis? What is the hypothesis space? Explain by a diagram. [2+3]
[Module2/CO2/Analyse-LOCQ]

Group - C

(Long Answer Type Questions)

Answer any ***one*** of the following 1×15=15

5.

(a) Write the KNN(K nearest neighbor) algorithm. What are the criteria on basis of which the value of K is assumed? [3+2]
[Module4/CO2/Understand-LOCQ]

(c) Classify the candidate {GPA: 7.8, No of Projects done: 4} to the following dataset of 10 students using KNN classifier. Assume value of K as 3. [10]

Sl.	GPA	No of Projects	Award
1.	9.5	5	YES
2.	8.0	4	YES
3.	7.2	1	NO
4.	6.5	5	YES
5.	9.5	4	YES
6.	3.2	1	NO
7.	6.6	1	NO
8.	5.4	1	NO
9.	8.9	3	YES
10.	7.2	4	YES

[Module4/CO3/Apply-HOCQ]

6.

(a) Write and explain FIND-S algorithm.

(b) Define the most specific and most generic hypothesis with reason.

[Module2/CO2/Understand-IOCQ]

(c) Explain Find-S with the help of the following training data :

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

[3+2+10]

[Module2/CO2/Apply-HOCQ]

MCKV Institute of Engineering

Paper Code: PE-IT501B

Machine Learning

Time Allotted: 1 Hour *Full Marks:* 30

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternatives for any *five* of the following: 5×1

i. Which of the following are the applications of clustering?

A. Identifying patterns of crime in different regions of a city and managing police enforcement based on frequency and type of crime

B. Identifying consumer segments and their properties to position products appropriately

C. Looking at social media behaviour to find out the types of online communities that exist

D. All of the above

ii. Which of the following approaches can be used in Hierarchical Clustering?

Divisive Clustering

Agglomerative Clustering

A. None of the above

B. Both of the above

iii. Which method should be preferred in the case of K-mode or K-means clustering when both categorical and numerical variables are present in the dataset?

A. None of the above

B. K-mode

C. K-means

D. Either would work

iv. Regarding K-Means algorithm Select the correct statement among the following:

A. K-means algorithm can be applied to both categorical and numerical variables.

B. The results of k-means algorithm get impacted by outliers and range of the attributes.

C. The clusters formed by k-means algorithm do not depend on the initial selection of cluster centers.

D. K-means clustering automatically selects the most optimum value of k

v. Select the problem sets where k-means clustering can be applied.

A. Given an e-com company's customer details - the products they purchased and the amount spent. The company wants to group it's customers based on their buying behaviour.

B. Predict whether a new customer would respond to a bank's new product basis his historical information

C. Weather forecast for next week, given the dataset of weather information for last five years

vi. K-Means algorithm: Arrange the steps of k-means algorithm in the order in which they occur:

1. Randomly selecting the cluster centroids
2. Updating the cluster centroids iteratively
3. Assigning the cluster points to their nearest center

A. 2-1-3
B. 1-2-3
C. 1-3-2

Group – B

(Short Answer Type Questions)

Answer any *two* of the following 2×5

2. Explain the differences between soft clustering and hard clustering with example.

3. Write down the differences between agglomerative and divisive approach of hierarchical clustering.

4. Explain PAM algorithm.

Group – C

(Long Answer Type Questions)

Answer any *one* of the following 1×15

5. Explain step by step K means clustering algorithm with example.

6. Explain Hierarchical clustering algorithm with example.

MCKV Institute of Engineering

Paper Code : PE-IT501B

Paper Name : Machine Learning

Time Allotted: 1 Hour | *Full Marks: 30*

The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable.

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternatives for any ***five*** of the following: 5×1=5

i. For $y = b_0 + b_1x$, y is

(a) Dependent variable (b) Independent variable

(c) intercept (d) slope

ii. For $y = b_0 + b_1x$, x is

(a) Dependent variable (b) Independent variable

(c) intercept (d) slope

iii. For $y = b_0 + b_1x$, b_1 is

(a) Dependent variable (b) Independent variable

(c) intercept (d) slope

iv. For $y = b_0 + b_1x$, b_0 is

(a) Dependent variable (b) Independent variable

(c) intercept (d) slope

v. Human blood group, for a ML analysis, is a/an

(a) Nominal variable (b) Ordinal variable

(c) Discrete variable (d) Continuous variable

vi. Population, for a ML analysis, is a/an

(a) Nominal variable (b) Ordinal variable

(c) Discrete variable (d) Continuous variable

Group – B

(Short Answer Type Questions)

Answer any ***two*** of the following 2×5=10

2. What is simple linear regression? Where is it most suitable to use? [CO1(Remember/IOCQ)] [3+2]

3. What is MAE? What is RMSE? Where MAE is more suitable over R2 score and why? [CO2(Understand/LOCQ)] [1×1×3]

4. Explain significance of testing and training of a ML model with a diagram. [CO2(Understand/LOCQ)]

Group – C

(Long Answer Type Questions)

Answer any ***one*** of the following 1×15=15

5.

(i) What is a hypothesis? What is meant by hypothesis space? Illustrate with a diagram.

(ii) What is hypothesis testing? What are the parameters of hypothesis testing?

Explain with the help of the following premises and dataset assuming suitable H_0 and H_A.

Premise1 : P1 and P1 are two distinct species of herbs.

Premise2: Fertilizer F_1, if applied to both P_1 and P_2 samples influences Higher growth.

Sl#	Plant species	F_1	Growth
1	P_1	Yes	Higher
2	P_2	Yes	Lower
3	P_2	Yes	Higher
4	P_2	No	Higher
5	P_1	Yes	Lower

[(3+2)+10]

[CO2(Apply/HOCQ)]

6. Differentiate Classification and Regression. Apply OLS algorithm of simple linear regression to the following dataset and derive the equation of the regression line. [5+10]

[X = 89, 43, 36, 36, 95, 10, 66, 34, 38, 20]

[Y = 21, 46, 3, 35, 67, 95, 53, 72, 58, 10]

[CO3(Apply/HOCQ)]