Foundations of Data Science

Unit - I

1. “Deployment of a data mining solution can be much less technical”. Justify. [6-2021]
2. What is data science? Explain the fundamental data mining tasks with suitable example for each. [8-2021, 8-2023]
3. Explain briefly Data Processing and Big Data. [6-2021]
4. Give the crisp model of data mining. Explain in detail its stages. [7-2021]
5. Outline and describe any four benefits of data analytic thinking in Business. [7-2021]
6. Explain the difference between Supervised and Unsupervised learning. [6-2021]
7. List and explain any four data mining tasks. [6-2023]
8. Distinguish between Bigdata 1.0 and Bigdata 2.0. [6-2023, 4-2022]
9. With a neat diagram, explain how data science is placed in the context of various data related process. [8-2023, 10-2022]
10. Illustrate CRISP model for data mining and its stages. [6-2023]
11. Mention and explain any four data mining tasks. [8-2022]
12. Give the CRISP model for data mining. Explain its stages. [8-2022]
13. Describe the data discovery and data preparation performed during the data analytics life cycle for predicting customer churn. [10-2022]
14. Distinguish between i) Machine learning and KDD ii) Predictive and descriptive modeling [6-2023]

Unit – II

1. What is probability estimation? Give all the steps to construct the probability estimation tree with smoothed probabilities. [10-2021, 8-2023]
2. Construct the probability estimation tree for the below given training set. [10-2021]

Sl. No.	A	B	C	Class
1	+	*	25	Y
2	+	%	35	N
3	-	*	40	N
4	-	*	35	Y
5	-	*	25	Y
6	+	*	20	N
7	+	%	30	Y

1. Construct the probability estimation tree with smoothed probabilities for the following data set. [10-2021, 12-2023]

Sl. No.	House owner	Marital status	Balance	Class: Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

1. Determine how attribute selection helps in constructing the decision tree. [6-2021]

2. Explain in detail the difference between predictive and descriptive modelling. [4-2021]
3. Demonstrate with example how attribute selection helps in constructing the decision tree. [8-2023]
4. Construct probability estimation tree with smoothed probabilities for the following dataset. [12-2023]

Sl.No.	House owner	Marital status	Balance	Class defaulted
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Divorced	95k	Yes
6	No	Married	60k	No
7	Yes	Divorced	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

1. Construct a probability estimation tree with probability for the following data set . [12-2023]

Name	Body temperature	Gives birth	Four legged	Hibernates	Target class
Human	Warm-blooded	Yes	No	No	Yes
Pigeon	Warm-blooded	No	No	No	No
Elephant	Cold-blooded	Yes	Yes	No	Yes
Leopard	Cold-blooded	Yes	No	No	No

Name	Body temperature	Gives birth	Four legged	Hibernates	Target class
Turtle	Cold-blooded	No	Yes	No	No
Penguin	Cold-blooded	No	No	No	No
Eel	Cold-blooded	No	No	No	No
Dolphin	Warm-blooded	Yes	No	No	Yes
Spinyanteater	Warm-blooded	No	Yes	Yes	Yes
Gila monster	Cold-blooded	No	Yes	Yes	No

1. Define entropy and information gain. Construct a decision tree for the following data set. [10-2022]

S.No.	Age	Balance	Class
1	20	40K	Yes
2	25	50K	Yes
3	30	30K	Yes
4	35	110K	No
5	40	120K	No
6	40	80K	Yes
7	30	115K	No
8	35	80K	?

Classify the 8th instance based on the decision tree constructed. 10. Differentiate between the following: i) Predictive and descriptive modeling ii) Supervised and unsupervised segmentation iii) Decision tree and probability estimation tree iv) Induction and deduction. [10-2022] 11. Discuss the different measures for selecting the attribute with the best split. Give suitable example. [6-2022] 12.

What is probability estimation? Outline steps to construct probability estimation tree with an example. [6-2022]

Unit – III

1. What are Support Vector Machines? With a neat diagram explain in detail the objective function of a Support Vector Machine. [10-2021]
2. For the below given training set develop a linear discriminant model for classification and classify the last Instance. [10-2021]

Sl. No.	Attr. A	Attr. B	Class
1	20	50K	Yes
2	24	50K	Yes
3	30	30K	Yes
4	38	100K	No
5	43	120K	No
6	41	90K	Yes
7	35	85K	??

1. Illustrate Repression using mathematical functions. [7-2021]
2. Explain with example how you can get class probability estimation using logistic repression. [5-2021, 7-2022]
3. Develop a linear discriminant function for the following sample data set and predict the class value for the 8th tuple. Also compute the value of f(x). [8-2021]

Sl. No.	Age	Balance	Class
1	20	40K	Yes
2	55	110K	No
3	30	30K	Yes

Sl. No.	Age	Balance	Class
4	25	35K	Yes
5	45	120K	No
6	40	80K	No
7	25	50K	Yes
8	50	95K	??

1. Illustrate support vector machines and objectives function with a neat diagram. [6-2023]
2. Describe the following with mathematical functions. i) SVM ii) Regression [6-2023]
3. Write the equation of a linear model and generate a linear discriminant function for the following data and predict the target value for the sixth and seventh tuple? [8-2023]

Sl No.	Age	Income	Class
1	20	40k	Yes
2	25	50k	Yes
3	30	30k	Yes
4	35	110k	No
5	40	120k	No
6	40	80k	?
7	30	115k	?

1. Differentiate classification trees with linear classifiers. [6-2023, 6-2022]
2. Interpret a comparison between logistic regression and tree induction. [6-2023]
3. For the given training set fit a linear discriminant model for classification and classify the last instance. [8-2023]

Sl.No.	Attribute A	Attribute B	Class
1	20	50k	Yes
2	24	50k	Yes
3	30	30k	Yes
4	38	100k	No
5	43	120k	No
6	41	90k	Yes
7	35	115k	No
8	45	102k	?

1. Consider the training examples shown below in Table(1) for a binary classification problem. [8-2022]

Instance	a1	a2	a3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

i) What is the entropy of this collection of training examples with respect to the positive class? ii) What are the information gains of a1 and a2 relative to these

training examples? 13. For the following data propose a linear discriminant function and predict the class for 7th tuple. [8-2022]

S.No.	Age	Balance	Class
1	25	40K	Yes
2	20	50K	Yes
3	35	30K	Yes
4	35	110K	No
5	40	120K	No
6	50	80K	No
7	60	115K	?
8			

1. List the objectives of support vector machine? How they are handled? Explain with suitable diagrams. [6-2022]
2. Discuss over fitting in mathematical function. Give suitable example. [6-2022]
3. Illustrate linear discriminant function using Iris data set. [7-2022]

Unit – IV

1. Discuss on the issues that arise in nearest neighbour method. [8-2021]
2. Differentiate between majority voting and similarity moderated voting. [5-2021]
3. Explain how clustering is carried out around centroids. [7-2021]
4. How K-means clustering is done? Explain with an example. [7-2021, 8-2023]
5. List few business problems or tasks based on Similarity. [4-2021]
6. Construct a single link dendogram for the given data set. Show all the intermediate distance matrices. [9-2021]

Sl. No.	A	B	C
1	20	32	10
2	25	41	12
3	30	42	13
4	30	41	15
5	25	40	16
6	28	28	18

1. Discuss on the issue that arise in nearest neighbor methods. [8-2023]
2. Find the distance between 1 and all others in the following table and identify the k-nearest neighbors. [12-2023]

Customer ID	Gender	Car type	Shirt size	Class
1	M	Family	S	C o
2	M	Sports	M	C o
3	M	Sports	M	C o
4	M	Sports	L	C o
5	M	Sports	XL	C o
6	M	Sports	XL	C o
7	F	Sports	S	C o
8	F	Sports	S	C o
9	F	Sports	M	C o
10	F	Luxury	L	C o
11	M	Family	L	C 1
12	M	Family	XL	C 1

Customer ID	Gender	Car type	Shirt size	Class
13	M	Family	M	C 1
14	M	Luxury	XL	C 1
15	F	Luxury	S	C 1
16	F	Luxury	S	C 1
17	F	Luxury	M	C 1
18	F	Luxury	M	C 1
19	F	Luxury	M	C 1
20	F	Luxury	L	C 1

1. Consider the following data set, classify Punit based on following methods. i) Majority voting ii) Similarity moderated voting [12-2023]

Student	Age	Marks	Credits	Placed
Arjun	25	75	8.3	Yes
Bhim	45	23	6.8	No
Charan	30	29	5.6	Yes
Suraj	35	60	7.3	No
Shyam	22	68	4.9	Yes
Punit	42	50	3.8	??

1. What is clustering? Explain the K-means algorithm for clustering. [10-2022]
2. Construct a single link Dendrogram for the Euclidean distance matrix given below. Show all the intermediate distance matrices. [10-2022]

	A	B	C	D	E
A	0	1	2	2	3

	A	B	C	D	E
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

1. Discuss the issues with Nearest-Neighbour methods and explain how to overcome them. [6-2022]
2. Consider the following data set classify David based on similarity moderated voting method. [8-2022]

Customer	Age	Income	Cards	Response
David	37	50K	2	?
John	35	35K	3	Yes
Rachael	22	50K	2	No
Ruth	63	200K	1	No
Jefferson	59	170K	1	No
Norah	25	40K	4	Yes

1. Explain how Hierarchical clustering is carried out. Mention its use. [6-2022]

Unit – V

1. How do we evaluate the binary classifier? Explain. [7-2021]
2. Explain how do you measure the term frequency and inverse document frequency with a suitable example. [6-2021, 10-2022]
3. How line prediction and social recommendation are interrelated? [7-2021, 6-2023]
4. What are the basic measures for text retrieval? Discuss in detail the text retrieval methods. [8-2021]

5. Define Bag of Words. Give the steps to calculate the term frequency of a word document. [6-2021, 6-2023]
6. Discuss in detail the applications of data mining for the retail industry. [6-2021]
7. Analyze the basic measures for text retrieval and its methods in detail. [8-2023]
8. Outline a short note on: i) Link prediction and social recommendation ii) Evaluating classifiers [10-2023]
9. What is meant by association analysis? Compute, i) Support ii) Confidence iii) Lift iv) Leverage if the 10000 transactions analyzed, the data shows that 6000 of the customer transactions include computer games, while 7500 include videos and 4000 include both computer games and videos. For the rule, Boys (X, “Computer games”) ⇒ boys (X, “Video”). [10-2023]
10. Define the following: i) Support ii) Confidence iii) Lift iv) Leverage v) Soft clustering. [10-2022]
11. Explain the major frequent mining approaches based on the kinds of data taken as input? Explain any two. [10-2022]
12. What is meant by association analysis? If 10000 transactions analyzed, the data shows that 6000 of the customer transactions include computer games, while 7500 include videos and 4000 include both computer games and videos. For the rule Buys(X, ‘Computer Games’) => Buys(X, Video) Compute support, confidence, lift, Leverage. [10-2022]