Machine Learning (B20EF0503)

Semester 5 Academic Year 2023-2024

Dept of Computer Science and Engineering

Question bank

Unit -1

1. **Define Machine learning. Explain the below terms with a suitable example**
   a. **Analytics Business Table**
   b. **Training Dataset**
   c. **Testing Dataset**
2. Explain Overfitting and Underfitting with a suitable example.
3. Define Data Quality Issues. Explain in detail about the data quality issues.
4. Consider the below dataset. Apply range normalization to the continuous feature 'XYZ' to the new range [0 ,1].

XYZ	20	33	55	99	135

5. Extracting insights from data is the job of data analytics. In this context, define predictive Data Analytics. List any four applications of predictive data analytics and explain about each.
6. Illustrate the structure of data quality report for categorical and continuous features with a suitable example.
7. With diagram describe the six key phases of the predictive data analytics project lifecycle that are defined by the CRISP-DM.
8. Discuss the technique of Normalization in detail and normalize 400, 500, 600, 800, 1200 in the range (0,1) using range normalization and Standard score.
9. Outline the different types of data related to the features in the Analytics Base Table.
10. List out and describe about the most common data quality issues in detail.
11. What Is Predictive Data Analytics? Explain with applications and its tools
12. Explain data preparation with its types
13. Explain forward Sequential Selection and Backward Sequential Selection with diagram
14. . With a neat diagram, explain the working of machine learning with a neat diagram.
15. Discuss briefly the following with handling strategies:
    (i) Missing values
    (ii) Irregular cardinality
    (iii) Outliers
16. Consider the input values [6,10,26,48,24,86,62,94,116, 144, 122,166] generate an equal width binning using 3 bins and equal frequency binning using 3 bins and
17. Identify the role of preprocessing in Machine Learning. Explain any three pre-processing steps.
18. Calculate Covariance for the given dataset.
    Height: 143, 125, 139, 160, 170
    Weight: 55, 60, 43, 62,84
19. Differentiate bias and variance with suitable explanation.

Unit-2

1. One of the best known decision tree induction algorithm is Iterative Dichotomizer 3 (ID3) algorithm. Build the Decision Tree using ID3 algorithm for the data set given below:

ID	Stream	Slope	Elevation	Vegetation
1	False	Steep	High	Chapparal
2	True	Moderate	Low	Riparian
3	True	Steep	Medium	Riparian
4	False	Steep	Medium	Chapparal
5	False	Flat	High	Conifer
6	True	Steep	Highest	Conifer
7	True	Steep	High	Chapparal

2. Similarity metric measures the similarity between two instances a and b in a dataset. Explain different similarity measures that can replace Euclidean distance.

3. Apply ID3 algorithm to construct a decision tree for the below email spam prediction dataset and also make a prediction for new query instance.

Suspicious words	**Unknown eSender**	**Contains Images**	**Class**
true	false	true	Spam
true	true	false	Spam
true	true	false	Spam
false	true	true	ham
false	false	false	ham
false	false	false	ham
Query	False	True	True	?

4. Explain decision tree with the demonstration of How do we decide which is the best decision tree to use?

5. Explain the Shannon's Entropy Model

6. Entropy for a type of suit being picked from a pack of (4 different suits)52 cards. Calculating the entropy of a set of 52 playing cards if we only distinguish between cards based on their suit (heart, club, diamond or spade)

a) Entropy and Gini index is a measure of randomness/distributions present in the dataset. Apply the following measure for the given dataset.
   1. Entropy
   2. Gini Index

Height	**Weight**	**Class**
9	150	Elephant
10	170	Elephant
6	80	Cow
5	90	Cow
4.5	110	Cow
11	165	Elephant

7. Explain Gini index

8. One of the best known decision tree induction algorithm is Iterative Dichotomizer 3 (ID3) algorithm. Build the Decision Tree using ID3 algorithm for the data set given below:

Sno	Height	Performance in Class	Class	Playing Cricket
1	5	Above Avearge	X	N
2	5.6	Below Average	X	N
3	4.7	Above Avearge	X	N
4	5.6	Above Avearge	X	Y
5	5.8	Above Avearge	X	N
6	5	Above Avearge	XI	Y
7	4.11	Above Avearge	XI	Y
8	5.9	Above Avearge	XI	Y
9	5.1	Above Avearge	X	N
10	5.6	Above Avearge	XI	Y
11	5.9	Below Average	X	Y
12	5.2	Above Avearge	XI	N
13	6	Below Average	XI	N
14	6.1	Above Avearge	XI	Y
15	5.9	Above Avearge	XI	Y
16	5.3	Below Average	X	N
17	5.8	Below Average	X	N
18	5.7	Above Avearge	XI	Y
19	4.9	Above Avearge	X	N
20	5.9	Below Average	XI	Y

9. Illustrate handling continuous descriptive features
10. Explain how to Predict the Continuous Targets
11. What is model ensemble? Explain with characteristics and standard approaches
12. Explain the **Nearest Neighbouring Algorithm**
13. Apply KNN on the dataset of the company which produces tissues for biological science laboratory. Predict the acceptability of the new type of tissue with acid durability being 3 and strength being 7. Consider the case for K=2.

Name	**Acid Durability**	**Strength**	**Class**
Type 1	7	7	Bad
Type 2	7	4	Bad
Type 3	3	4	Good
Type 4	1	4	Good

14. Write K-d tree Algorithm

15. Predict Continuous Targets Using K nearest approach
    Need to make prediction of price for the item with age =2 and rating =5 with k=3 ,considering the target value being continuous
    Take your own data set based on the question

16. Explain Cosine Similarity

17. Suggest the company whether the new product Prod5, with Attrib1=3 and Attrib2=7 will be accepted or rejected using similarity based learning algorithm by considering 2 nearest neighbors.

Product	Attrib1	Attrib2	Status
Prod1	7	7	Reject
Prod2	7	4	Reject
Prod3	3	4	Good
Prod4	1	4	Good

18. Apply KNN on the dataset of the company which produces tissues for biological science laboratory. Predict the acceptability of the new type of tissue with durability = 3 and strength =7 . Consider the case for K=3.

Name	**Durability**	**Strength**	**Class**
Type 1	6	6	**Bad**
Type 2	6	4	**Bad**
Type 3	3	3	**Good**
Type 4	2	4	**Good**

19 Consider the below training dataset and compute distance between the query to all the instances of the dataset using Euclidian distance metrics.

X	**Y**	**Z**	**Class**
10	40	34	Yes
11	55	39	Yes
60	67	80	No
59	58	99	No

Query	12	46	33