|  |
| --- |
| P.V.P SIDDHARTHA INSTITUTE OF TECHNOLOGY |
| BRANCH :  | Computer Science & Engineering | REGULATION :  | PVP20 |
| Course:  | B.Tech | Course Name:  | DATA SCIENCE |
| Course Code:  | 20CS4501A | Year and Semester:  | III-I |
| **QUESTION BANK**  |

|  |
| --- |
| **UNIT - I** |
| Q. NO. | QUESTION | CO | LEVEL | MARKS |
| 1 | Explain the phases of Data Science.  | 1 | 2 | 14 |
| 2 | What is Exploratory Data Analysis? Explain any two types of visualization.  | 1 | 2 | 14 |
| 3 | Explain various Hyper parameter optimization techniques with suitable examples.  | 1 | 2 | 14 |
| 4 | Briefly explain the role of Data Science in various fields.  | 1 | 2 | 14 |
| 5 | Explain Data Science phases and lifecycle and write the names of tools for Data Science.  | 1 | 2 | 14 |
| 6 | Explain about roles and stages in data science project. | 1 | 2 | 14 |
| 7 | Explain about exploring and managing data in data science.  | 1 | 2 | 14 |
| 8 | Explain the various processes for preparing a dataset to perform a data science task.  | 1 | 2 | 14 |
| 9 | Define Hyperparameter Optimization and discuss various strategies for optimizing hyperparameter methods.  | 1 | 2 | 14 |

|  |
| --- |
| **UNIT-II** |
| Q. NO. | QUESTION | CO | LEVEL | MARKS |
| 1 | Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72 (i) Use min-max normalization to transform the value of 45 for age onto the range [0,1] (ii) Use Z-Score normalization to transform the value 45 for age where the standard deviation of age is 20.64 years.  | 2 | 3 | 14 |
| 2 | 1. Differentiate between data reduction and dimensionality reduction for data discretization.
2. Explain the role of attributes in classification. 54
 | 2 | 2 | 14 |
| 3 | Normalize the following group of data by using the following techniques. 200, 300, 400, 600, 1000 1. min-max normalization technique
2. z-score normalization
3. Decimal scaling.
4. Write your observations on the above techniques.
 | 2 | 3 | 14 |
| 4 | How to implement Data Transformation and Data Discretization? Explain with examples. | 2 | 2 | 14 |
| 5 | Given data = {2, 3, 4, 5, 6, 7; 1, 5, 3, 6, 7, 8}. Use PCA Algorithm to compute the principal component.  | 2 | 3 | 14 |
| 6 | In real-world data, tuples with missing values for some attributes are a common occurrence. Apply various pre-processing methods for handling this problem.  | 2 | 3 | 14 |
| 7 | The following are the sorted data price (in rupees) of certain items in the supermarket. 4, 8, 15, 21, 21, 24, 25, 28, 34, 36, 39, 42, 51, 57, 60 Smooth the data by using the following smoothing techniques. Consider the bin size as 3. i) Bin means ii) Bin medians iii) Bin boundaries  | 2 | 3 | 14 |
| 8 | 1. Why data transformation is important and when do we need it.
2. When do we use splitting table technique?
3. How does adding redundant column can become cause for loss of information? Justify?
 | 2 | 2 | 14 |
| 9 | Evaluate any two data reduction techniques with examples. What is the format for reporting results of each?  | 2 | 3 | 14 |

|  |
| --- |
| **UNIT-III** |
| Q. NO. | QUESTION | CO | LEVEL | MARKS |
| 1 | Explain the Mean and Variance of Binomial Distribution and its properties. | 3 | 2 | 14 |
| 2 | There is a class of 25 students, and the mean score of their test is 60 out of 100, with standard deviation 4 marks from the mean. While other students of the school have a mean score of 50 on the same test. What will be the t-score for calculating the probability that school students scored not less than 60 on their tests? | 3 | 3 | 14 |
| 3 | Find the mean and standard deviation of a normal distribution in which 7% are under 35, and 89% are under 63. | 3 | 3 | 14 |
| 4 | In a Normal Distribution 31% of the items are under 45 and 8% are over 64. Find the Mean and Variance of the Distribution. | 3 | 3 | 14 |
| 5 | From the following data, find whether there is any significant linking in the habit of taking soft drinks among categories of employees by using chi-square test.

|  |  |
| --- | --- |
| Soft drinks | Employees |
| Clerks | Teachers | Officers |
| Pepsi | 10 | 25 | 65 |
| Thumps up | 15 | 30 | 65 |
| Fanta | 50 | 60 | 30 |

 | 3 | 3 | 14 |
| 6 | Explain about various methods of Data Collection involved in Data Science. | 3 | 2 | 14 |
| 7 | In a sample of 1000 cases, the mean of a particular test is 14, and the standard deviation is 2.5. assuming the distribution to be normal, Determine 1. how many students score between 12 and 15?
2. how many scores above 18?
3. how many scores below 18?
 | 3 | 3 | 14 |
| 8 | Differentiate Stratified sampling and Cluster sampling techniques with examples. | 3 | 2 | 14 |
| 9 | A pair of dice is thrown 360 times and frequencies of each sum are indicated below Would you say that the dice are fair on the basis of the Chi-Square test at 0.01 Level of Significance.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sum | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Frequency | 8 | 24 | 35 | 37 | 44 | 65 | 51 | 42 | 26 | 14 | 14 |

 | 3 | 3 | 14 |

4

|  |
| --- |
| **UNIT-IV** |
| Q. NO. | QUESTION | CO | LEVEL | MARKS |
| 1 | If the logit score (linear predictor) is given by –2.4 + 1.5 X1 + 2 X2, find the estimated P(Y = 1) for each of the following combination of the IDVs:X1:  0       1.5        2       3       -2      -2.5X2:  1         0       1.5     -1        2       2.5   | 4 | 3 | 14 |
| 2 | Which specific regressors seem essential in multiple regression? How will you address this question? Discuss. | 4 | 2 | 14 |
| 3 | Suppose you have the following data with one real-value input variable & one real-value output variable. What is leave-one out cross validation mean square error in case of linear regression (Y = bX+c)?

|  |  |
| --- | --- |
| **X (independent variable)** | **Y (dependent variable)** |
| 0 | 2 |
| 2 | 2 |
| 3 | 1 |

 | 4 | 3 | 14 |
| 4 | 1. Discuss the need for fitting the model in multiple regression.
2. Explain Logistic Regression with an example.
 | 4 | 2 | 14 |
| 5 | 1. Obtain the likelihood equation for estimating the parameters of a logistic regression model.
2. Define multiple linear regression model. Explain the least squares method to estimate parameters in multiple linear regression models.
 | 4 | 2 |  |
| 6 | Explain Linear Discriminative analysis in detail with an example. | 4 | 2 | 14 |
| 7 | Apply linear regression using the method of least squares to the following data and predict the crop yield for rain fall of 5 cm.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rain fall(in cms)** | 10.5 | 8.8 | 13.4 | 12.5 | 18.8 | 10.3 | 7.0 | 15.6 | 16 |
| **Paddy yield****(quintal per acre)** | 30.3 | 46.2 | 58.8 | 59.0 | 82.4 | 49.2 | 31.9 | 76.0 | 78.8 |

 | 4 | 3 | 14 |
| 8 | Explain how can over fitting and under fitting issues are handled in Regression modeling. | 4 | 2 | 14 |
| 9 | Build a linear regression model with the following data and test for overall fit. Also, test for the individual significance of X1 and of X2.Y:  12.8    13.9    15.2     18.3     14.5     12.4X1:    2          3        5          5          4          1X2:    4          2        5          1          2          3 | 4 | 3 | 14 |
| 10 | Apply logistic regression to demonstrate binary classification. | 4 | 3 | 14 |

|  |
| --- |
| **UNIT - V** |
| Q. NO. | QUESTION | CO | LEVEL | MARKS |
| 1 | Explain the bias/ variance dilemma about the model complexity. | 1 | 2 | 14 |
| 2 | Explain k-fold cross validation and how it can be implemented for building a model. | 4 | 3 | 14 |
| 3 | Imagine that you find out that your model has low bias and high variance. Which algorithm would be best suited to this problem? What's the reason? | 4 | 3 | 14 |
| 4 | 1. Develop the estimate of in-sample error derivation.
2. Explain Minimum description length principle for model building
 | 1 | 2 | 14 |
| 5 | Briefly explain how you would calculate a cross-validated estimate of prediction error in aLinear regression. Is this estimate likely more minor or more significant than the in-sample error? | 1 | 2 | 14 |
| 6 | What is the holdout approach? What is the limitation of this approach? Name four alternative approaches for it. | 1 | 2 | 14 |
| 7 | Explain bias and variance in machine learning and how bias-variance decomposition is used for deciding the model complexity? | 1 | 2 | 14 |
| 8 | Given data set STr = {(xi, yi), i=1,….6}, xi**∈**ℝ a feature scalaryi∈{-1,+1} a class label.Data points in the data set are(x1,y1)=(2,-1) (x2,y2)=(7,-1) (x3,y3)=(4,+1) (x4,y4)=(1,-1) (x5,y5)=(3,+1) (x6,y6)=(6,+1)Suppose you are training a Linear Classifier f(x;a,b) = sign(ax+b) with 2-fold Cross Validation where sign(z) = $\left\{\begin{array}{c}+1,z\geq 0\\-1, z<0\end{array}\right.$Split STr into S1={(x1,y1) (x2,y2) (x3,y3)} and  S2={(x4,y4) (x5,y5) (x6,y6)} After training the classifier f on S1, we have a1=-1, b1=5 and then try to validate the classifier on S2.After training the classifier f on S2, we have a2=2, b2=-3 and then try to validate the classifier on S1Calculate the average training error in the 2-fold cross-validation.  | 4 | 3 | 14 |
| 9 | Given data set TTr = {(xi, yi), i=1,….4}, xi**∈**ℝ a data pointyi∈{-1,+1} a corresponding label.The Data points are(x1,y1)=(5,-1) (x2,y2)=(6,-1) (x3,y3)=(1,+1) (x4,y4)=(4,-1) Suppose you are training a Linear Classifier f(x;a,b) = sign(bx+a) with 2-fold Cross Validation where sign(z) = $\left\{\begin{array}{c}+1,z\geq 0\\-1, z<0\end{array}\right.$Split STr into T1={(x1,y1) (x2,y2)} and  T2={(x3,y3) (x4,y4)} Subsequently training the classifier f(x;a,b) on T1, we have a1=-3, b1=4 and then try to validate the classifier on T2.Training the classifier f(x;a,b) on T2, we have a2=1, b2=-5 and then try to validate the classifier on T1Calculate the average validation error (i.e. the cross-validation error) in the 2-fold cross-validation. | 4 | 3 | 14 |