|  |
| --- |
| **P.V.P Siddhartha Institute of xTechnology** |
| **Department of Computer Science and Engineering** |
| **Course: B.Tech** | **Year: III** | **Semester: I** | **Descriptive: II** | **A.Y: 2023-24** |
| **Subject Code:****20CS4501A** | **Subject Name: Data Science** | **Regulation: PVP20** |
| **Duration: 1 hr 30 min** | **Maximum Marks:15 Marks** | **Date: 02-11-23** |
| **Answer all the Questions. Each Question carries 5Marks 3×5M=15M** |
| **Q.No** |  | **Marks** | **CO** | **Level** |
| **1.** | **a)** | Differentiate Stratified sampling and Cluster sampling techniques with examples. | **2M** | **CO1** | **L2** |
| **b)** | 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
 | **3M** | **CO3** | **L3** |
|  |
| **2.** | **a)** | Apply Linear Discriminative analysis for 2D data set into a 1-dimensional plane.X1 = {(4,1) (2,4) (2,3) (3,6) (4,4)}X2 = { (9,10) (6,8) (9,5) (8,7) (10,8) | **5M** | **CO4** | **L3** |
|  |
| **3.** | **a)** | What is the holdout approach? What is the limitation of this approach? Name four alternative approaches for it. | **2M** | **CO1** | **L2** |
| **b)** | 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. | **3M** | **CO4** | **L3** |

|  |  |  |
| --- | --- | --- |
| **P.V.P Siddhartha Institute of Technology(Autonomous)** | **Invigilator Signature with date:** | **Marks Obtained:** |
| **Department of Computer Science and Engineering** |
| **Course: B.Tech** | **Year: III** | **Semester: I** | **Objective: II** |
| **Regulation: PVP20** | **Maximum Marks:10 Marks** | **Session: F.N.** |
| **A.Y: 2023-24** | **Date: 02-11-23** | **Duration: 20 min** |
| **Subject Code: 20CS3403** | **Course Name: Data Science** |
| **Registered Number:** | **Name:** |
| **Answer all the Questions. Each Question carries ½ Mark 20×½ M =10M** |
| **S.No** | **Question** | **CO** | **Level** | **Answer** |
| 1. | Which of the following standard Probability density functions apply to discrete Random Variables? | CO1 | L2 |  |
| a) Gaussian Distribution | b) Poisson Distribution | c) Rayleigh Distribution | d) Exponential Distribution |
| 2. | What is the area under a conditional Cumulative density function? | CO1 | L2 |  |
| a) 0 | b) Infinity | c) 1 | d) Changes with CDF |
| 3. | A table with all possible values of a random variable and its corresponding probabilities is called \_\_\_\_\_\_\_\_\_\_\_ | CO1 | L2 |  |
| a) Probability Mass Function | b) Probability Density Function | c) Cumulative distribution function | d) Probability Distribution |
| 4. | Find λ in Poisson’s distribution if the probabilities of getting a head in biased coin toss as 3/4 and 6 coins are tossed. | CO1 | L2 |  |
| a) 4.5 | b) 3.5 |  c) 5.5 | d) 6.6 |
| 5. | The mean and variance of Poisson’s distribution are the same. | CO1 | L2 |  |
| a) False | b) True |  |  |
| 6. |  It is suitable to use Binomial Distribution only for \_\_\_\_\_\_\_\_\_\_\_ | CO1 | L2 |  |
| a) Large values of ‘n’ | b) Fractional values of ‘n’ | c) Small values of ‘n’ | d) Any value of ‘n’ |
| 7. | Which one is a classification algorithm?  | CO1 | L2 |  |
| a) Lasso regression | b) Linear regression | c) Polynomial regression | d) Logistic regression |
| 8. |  What characterizes unlabeled examples in machine learning | CO1 | L2 |  |
| a) there is plenty of confusing knowledge | b) there is no confusing knowledge | c) there is prior knowledge | d) there is no prior knowledge |
| 9. | Identify the learning algorithm for “facial identities for facial expressions”. | CO1 | L2 |  |
| a) prediction | b) Recognition patterns | c) Recognizing anomalies | d) generating patterns |
| 10. | Among the following options, identify the one which is false regarding Multiple linear regression. | CO1 | L2 |  |
| a) use for prediction | b) use for interpretation | c) discovers casual relationships | d) relates inputs to outputs |
| 11. | Analysis of ML algorithm needs | CO1 | L2 |  |
| a) statistical learning theory | b) computational learning theory | c) only B | d) both A & B |
| 12. | Classification is- | CO1 |  L2 |  |
| a) Supervised learning | b) Reinforcement learning | c) Unsupervised learning | d) semi-supervised Learning |
| 13. | What is unsupervised learning? | CO1 | L2 |  |
| a) no. of groups may be known | b) neither features nor no. of groups is known | c) Features of the group explicitly stated | d) None |
| 14. | When is it appropriate to use Logistic Regression? | CO1 | L2 |  |
| a) When the independent variables are not binary | b) When the dependent variable is not binary | c) When the dependent variable is binary | d) When the independent variables are binary |
| 15. | With the help of a confusion matrix, we can compute the-  | CO1 | L2 |  |
| a) Recall | b) Precision | c) Accuracy | d) All of the above |
| 16. | We have been given a dataset with n records in which we have an input attribute as x and an output attribute as y. Suppose we use a linear regression method to model this data. We randomly split the data into training and testing sets to test our linear regressor. What do you expect will happen with bias and variance as you increase the size of training data? | CO1 | L2 |  |
| a) Bias increases and Variance decreases | b) Bias decreases and Variance increases | c) Bias decreases and Variance decreases | d) Bias increases, and Variance increases |
| 17. | What is overfitting? | CO1 | L2 |  |
| a) Poor results in training and poor result in test | b) Great result in training and poor result in test | c) Poor results in training and poor result in test | d) Great result in training and a great result in the test |
| 18. | What is the term on which the machine learning algorithms build a model based on sample data? | CO1 | L2 |  |
| a) Data Training | b) Transfer data  | c) Training data  | d) none |
| 19. | Which of the following is the correct use of Cross-validation? | CO1 | L2 |  |
| a) selecting parameters in the prediction function | b) Comparing predictors | c) selecting variables to include in a model | d) All of the mentioned |
| 20. | Among the following options identify the one which is false regarding regression. | CO1 | L2 |  |
| a) discovers casual relationships | b) prediction | c) interpretation | d) relates inputs to outputs |