**Department of Master of Computer Applications**

**RV College of Engineering®, Bengaluru-560059**

***An Autonomous Institution Affiliated to VTU, Belagavi***

|  |  |
| --- | --- |
| ***Course Title*: Computer Organization and Architecture** | ***Course Code*: 18 MCA343** |
| ***Total Contact Hours*: 39 Hrs.** | ***Credits:03*** |
| ***SEE Marks*: 100** | ***CIE Marks*: 100** |
| ***Semester: III*** | ***Academic Year: 2019-20(Odd)*** |
| ***Course Co-odinator*: Dr. Andhe Dharani** | ***Date*: 27/08/2019** |

**Unit I**

1. What is machine learning? Explain any two business applications of machine learning
2. What are the possible ethical issues of machine learning applications
3. What is human learning? Give any two examples
4. What are the types of human learning? Are there equivalent forms of machine learning
5. Explain the different forms of machine learning with a few examples
6. What do you mean by a well-posed learning problem? Explain important features that are required to well-define a learning problem
7. Can all problems be solved using machine learning problems? Explain your response with real time examples
8. Explain the process of abstraction with an example
9. How do machines learn? Explain the process of machine learning
10. What are the different types of supervised learning? Explain them with a sample application in each area
11. What are the different types of unsupervised learning? Explain them with a sample application in each area
12. Write the difference between
    1. Abstraction and Generalization
    2. Supervised and unsupervised learning
    3. Classification and Regression
13. Explain in detail the process of machine learning
14. What are the main activities involved when you are preparing to start with modeling in machine learning?
15. Differentiate
    1. Categorical Vs Numeric attribute
    2. Dimensionality reduction Vs Feature extraction
16. Why do we need to explore data? Is there a difference in the way of exploring qualitative data vis-à-vis qualitative data?
17. What is meant by data pre-processing?
18. Explain qualitative and quantitative data in details. Differentiate between the two
19. Prepare a simple data set along with some sample records in it. Have at least one attribute of the different data types used in machine learning
20. What are the different causes of data issues in machine learning? What are the fallouts?
21. When there are variables with certain values missing, will that impact the learning activity? If so, how can that be addressed?
22. Explain how cross-tabs can be used to understand relationship between two variables
23. With examples, explain the different ways of exploring categorical data
24. When there are variables with certain values missing, will that impact the learning activity? If so, how can that be addressed?
25. Explain in details, the different strategies of addressing missing data values.
26. What are the different techniques for data pre-processing? Explain, in brief, dimensionality reduction and feature selection.
27. Write the difference between
    1. Nominal and Ordinal data
    2. Box plot and Histogram
    3. Mean and median

**Unit II**

1. What is model in context of machine leaning? How can you train a model?
2. Explain “No Free Lunch” theorem in context of machine leaning.
3. Explain with neat diagram hold out method and bootstrap sampling
4. Explain the process of K-fold cross- validation with neat diagram.
5. What is the main purpose of a descriptive model? State some real-world problems solved using descriptive models.
6. Differentiate
   1. Predictive Vs. Descriptive models
   2. Model underfitting Vs. Overfitting
   3. Cross-Validation Vs. Bootstrapping
7. Express target function in context of a real-life example. How is the fitness of a target function measured?
8. Can the performance of a learning model be improved? If yes, explain how.
9. How would you calculate the success of an unsupervised learning model? What are the most popular measures of performance of an unsupervised learning model?
10. Is there a way to use a classification model for a numerical values. How is it different from predictive modeling for categorical values?
11. While predicting malignancy of tumour of a set of patients using a classification model, the following are the data recorded:
    1. Correct predictions – 15 malignant, 75 benign
    2. Incorrect predictions – 3 malignant, 7 benign

Calculate the error rate, Kappa value, sensitivity, precision and F-measure of the model

1. Explain, in detail the different aspects of feature engineering
2. Discuss with example the need for feature selection, what are the different approaches of feature selection
3. Explain the process of
   1. Encoding categorical (nominal and ordinal) variables
   2. Transforming numeric features to categorical features
4. When can a feature be termed as redundant? What are the measures to determine the potentially redundant features
5. Explain the filter and wrapper approaches of feature selection. What are the merits and demerrits of these approaches
6. With a real time example, explain the main underlying concept of feature extraction. What are the most popular algorithms for the extraction
7. Why is cosine similarity a suitable measure in context of text categorization? Two rows in a document-term matrix have values – (2,3,2,0,2,3,3,0,1) and (2,1,0,0,3,2,1,3,1). Find the cosine similarity.
8. Find the hamming distance between 10001011 and 11001111
9. Compare the Jaccard index and similarity matching coefficient of two features having values (1,1,0,0,1,0,1,1) and (1,0,0,1,1,0,0,1)
10. Give few practical examples of high-dimensional data set. What is the challenge while applying machine learning technique on high-dimensional data set? How can that be addressed?
11. Write the difference between
    1. Sequential forward selection vs. Sequential backward elimination
    2. Filter vs. Wrapper method of feature selection
    3. Jaccard Coefficient vs. SMC

**Unit III**

1. Explain the concept of Prior, Posterior and Likelihood with an example
2. How Bayes’ theorem supports the concept learning principle?
3. Explain Naïve Bayes classifier with an example of its use in practical life
4. In an airport security checking system, the passangers are checked to find out any intruder. Let I with i ϵ {0,1} be the random variable which indicates whether somebody is an intruder (i=1) or not (i=0) and A with a ϵ {0,1} be the variable indicating alarm. An alarm will be raised if an intruder is identified with probability P(A=1| I = 1) =0.98 and a non-intruder with probability P(A=1| I = 0) = 0.001, which implies the error factor. In the population of passengers, the probability of someone is intruder is P(I = 1) =0.00001. What is the probability that an alarm is raised when a person actually is an intruder?
5. An antibiotic resistance test (random variable T) has 1% false positives(i.e. 1% of those not resistance to an antibiotic show positive result in the test) and 5% false negatives (i.e. 5% of those actually resistant to an antibiotic test negative). Let us assume that 2% of those tested are resistant to antibiotics. Determine the probability that somebody who tests positive is actually resistant (random variable D).
6. For preparation of the exam, a student knows that one question is to be solved in the exam which is either of types A,B or C. The probabilities of A, B or C appearing in the exam are 30%, 20% and 50% respectively. During the preparation, the student solved 9 of 10 problems of type A, 2 of 10 problems of type B and 6 of 10 problems of type C
   1. What is the probability that the student will solve the problem of the exam?
   2. Given that the student solved the problem, what is the probability that it was of type A?
7. A CCTV is installed in a bank to monitor the incoming customers and take a photograph. Though there are continuous flows of customers, we create bins of timeframe of 5 min each. In each time frame of 5 min, there may be a customer moving into the bank with 5% probability or there is no customer (again, for simplicity, we assume that either there is 1 customer or none, not the case of multiple customers). If there is a customer, it will be detected by the CCTV with a probability of 99%. If there is no customer, the camera will take a false photography by detecting other thing’s movement with a probability of 10%.
   1. How many customers enter the bank on average per day (10 jpirs)?
   2. How many false photographs (there is a photograph taken even though there is no customer) and how many missed photographs(there is no photograph even though there is a customer) are there on average per day?
   3. If there is photography, what is the probability that there is indeed a customer?
8. Define the concept of consistent learners.
9. Write two strengths and weaknesses of Bayes Classifier
10. In brief, explain the SVM model.
11. Define support Vectors in SVM model.
12. What are the advantages and disadvantages of kNN model
13. Write and explain the kNN model
14. Discuss the error rate and validation error in the kNN algorithm
15. Discuss how to calculate the distance between the test data and the training data for kNN
16. Explain various options of searching a decision tree
17. What are the advantages and disadvantages of decision tree method
18. Discuss the decision tree algorithm in detail
19. What is inductive bias in a decision tree? How to avoid overfitting
20. A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.005 of the entire population have this cancer. Find out the probability of considering positivity of the test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp | Humidity | Wind | Tennis? |
| *D*1 | Sunny | Hot | High | Weak | *No* |
| *D*2 | Sunny | Hot | High | Strong | *No* |
| *D*3 | Overcast | Hot | High | Weak | *Yes* |
| *D*4 | Rain | Mild | High | Weak | *Yes* |
| *D*5 | Rain | Cool | Normal | Weak | *Yes* |
| *D*6 | Rain | Cool | Normal | Strong | *No* |
| *D*7 | Overcast | Cool | Normal | Strong | *Yes* |
| *D*8 | Sunny | Mild | High | Weak | *No* |
| *D*9 | Sunny | Cool | Normal | Weak | *Yes* |
| *D*10 | Rain | Mild | Normal | Weak | *Yes* |
| *D*11 | Sunny | Mild | Normal | Strong | *Yes* |
| *D*12 | Overcast | Mild | High | Strong | *Yes* |
| *D*13 | Overcast | Hot | Normal | Weak | *Yes* |
| *D*14 | Rain | Mild | High | Strong | *No* |

Calculate the Entropy, Information gain and draw the final decision tree

1. Discuss the random forest model in detail
2. Consider the following dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| N | Color | Type | Origin | Stolen? |
| 1 | red | sports | domestic | yes |
| 2 | red | sports | domestic | no |
| 3 | Red | sports | domestic | yes |
| 4 | Yellow | sports | domestic | no |
| 5 | Yellow | sports | imported | yes |
| 6 | Yellow | SUV | imported | no |
| 7 | Yellow | SUV | imported | yes |
| 8 | Yellow | SUV | domestic | no |
| 9 | Red | SUV | imported | no |
| 10 | Red | sports | domestic | yes |

Construct frequency tables, and ***Classify (red, SUV, domestic)*** using Naïve Bayes classifier, justify whether the car might be stolen or not

1. Consider the following set of training examples.

a) What is the entropy of this collection of training example with respect to the target

function classification?

b) What is the information gain of A2 relative to these training examples?

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | Classification | A1 | A2 |
| 1 | + | T | T |
| 2 | + | T | T |
| 3 | - | T | F |
| 4 | + | F | F |
| 5 | - | F | T |
| 6 | - | F | T |

**Unit IV**

1. Define simple linear regression using a graph explaining slope and intercept
2. Find the slope of the graph where lower point on the line is represented as (-3,-1) and the higher point on the line is represented as (2,2)
3. What are the conditions of a positive and negative slope in linear regression
4. Explain ordinary least square with formula for a and b
5. Explain the OLS algorithm with steps.
6. With an example, explain the multiple linear regression
7. Explain the assumptions in regression analysis and BLUE concept
8. Explain in detail
   1. Polynomial regression
   2. Logistic regression
9. Discuss maximum likelihood estimation in detail
10. With an example analyze when to use agglomerative clustering or Divisive clustering
11. What are the board three categories of clustering techniques? Explain each one in brief
12. Detail out the PAM algorithm with stepwise explanation
13. Describe the main difference in the approach of k-means and k-mediods algorithms with a neat diagram
14. You are given a set of one-dimensional data points: {5,10,15,20,25,30,35}. Assume that k = 2 and first set of random centroid is selected as {15,32} and then it is refined with {12, 30}
    1. Create two clusters with each set of centroid mentioned above following the k-means approach
    2. Calculate the SSE for each set of centroid
15. Explain the apriori algorithm for association rule learning with an example
16. How the distance between clusters is measured in hierarchical clustering? Explain the use of this measure in making decision on when to stop the iteration
17. Discuss the strengths and weaknesses of the k-means algorithm?
18. During a research work, you found 7 observations as described with the data points below. You want to create 3 clusters from these observations using K-means algorithm. After first iteration, the clusters C1,C2,C3 has following observations:

C1 : {(2,2),(4,4),(6,6,)}

C2 : {(0,4),(4,0)}

C3 : {(5,5) ,(9,9)}

If you want to run a second iteration then what will be the cluster centroids? What will be the SSE of this clustering

1. How the distance between clusters is measured in hierarchical clustering? Explain the use of this measure in making decision on when to stop the iteration
2. In a software project, the team is trying to identify the similarity of software defects identified during testing. They wanted to create 5 clusters of similar defects based on the text analytics of the defect descriptions. Once the 5 clusters of defects are identified, any new defect created is to be classified as one of the types identified through clustering. Explain this approach through a neat diagram. Assume 20 defect data points which are clustered among 5 clusters and k-means algorithm was used

**Unit V**

1. Give an application example where global outliers, contextual outliers, and collective outliers are all interesting. What are the attributes, and what are the contextual and behavioral attributes?
2. How is the relationship among objects modeled in collective outlier detection?
3. Explain the challenges of outlier detection
4. Explain the Distance-Based Outlier Detection algorithm
5. Explain the CELL method for distance-based outlier detection
6. Suppose a city’s average temperature values in July in the last 10 years are, in value-ascending order, 24.0oC, 28.9oC, 28.9oC, 29.0oC, 29.1oC, 29.1oC, 29.2oC, 29.2oC, 29.3oC, and 29.4oC, using use the maximum likelihood method to find the parameters and then find the outliers
7. To understand why angle-based outlier detection is a heuristic method, give an example where it does not work well. Can you come up with a method to overcome this issue?
8. Explain with an example how angle-based outlier detection (ABOD) is computationally better.
9. Explain the challenges to be met for outlier detection methods for high-dimensional data
10. With sample data sets, and neat diagrams explain the different shapes of histogram. What are bins
11. Explain how bivariate relationships can be explored using scatter plot. Can outliers be detected using scatter plot?
12. What is an IQR? How is it measured?
13. Adapt a simple semi-supervised method for outlier detection. Discuss the scenario where you have (a) only some labeled examples of normal objects, and (b) only some labeled examples of outliers.
14. Using an equal-depth histogram, design a way to assign an object an outlier score
15. Explain, in detail, the different components of a box plot? When will the lower whisker be longer than the upper whisker? How can outliers be detected using box plot