**1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?**
a. Decision Tree
b. Regression
c. Classification
***d. Random Forest***

**2. To find the minimum or the maximum of a function, we set the gradient to zero because:**
***a. The value of the gradient at extrema of a function is always zero***
b. Depends on the type of problem
c. Both A and B
d. None of the above

**3. The most widely used metrics and tools to assess a classification model are:**
a. Confusion matrix
b. Cost-sensitive accuracy
c. Area under the ROC curve
***d. All of the above***

**4. Which of the following is a good test dataset characteristic?**
a. Large enough to yield meaningful results
b. Is representative of the dataset as a whole
***c. Both A and B***
d. None of the above

**5. Which of the following is a disadvantage of decision trees?**
a. Factor analysis
b. Decision trees are robust to outliers
***c. Decision trees are prone to be overfit***
d. None of the above

**6. How do you handle missing or corrupted data in a dataset?**
a. Drop missing rows or columns
b. Replace missing values with mean/median/mode
c. Assign a unique category to missing values
***d. All of the above***

**7. What is the purpose of performing cross-validation?**
a. To assess the predictive performance of the models
b. To judge how the trained model performs outside the sample on test data
***c. Both A and B***

**8. Why is second order differencing in time series needed?**
a. To remove stationarity
b. To find the maxima or minima at the local point
***c. Both A and B***
d. None of the above

**9. When performing regression or classification, which of the following is the correct way to preprocess the data?**
***a. Normalize the data → PCA → training***
b. PCA → normalize PCA output → training
c. Normalize the data → PCA → normalize PCA output → training
d. None of the above

**10. Which of the folllowing is an example of feature extraction?**
a. Constructing bag of words vector from an email
b. Applying PCA projects to a large high-dimensional data
c. Removing stopwords in a sentence
***d. All of the above***

**11. What is pca.components in Sklearn?**
***a. Set of all eigen vectors for the projection space***
b. Matrix of principal components
c. Result of the multiplication matrix
d. None of the above options

**12. Which of the following is true about Naive Bayes ?**
a. Assumes that all the features in a dataset are equally important
b. Assumes that all the features in a dataset are independent
***c. Both A and B***
d. None of the above options

**13. Which of the following statements about regularization is not correct?**
a. Using too large a value of lambda can cause your hypothesis to underfit the data.
b. Using too large a value of lambda can cause your hypothesis to overfit the data.
c. Using a very large value of lambda cannot hurt the performance of your hypothesis.
***d. None of the above***

**14. How can you prevent a clustering algorithm from getting stuck in bad local optima?**
a. Set the same seed value for each run
***b. Use multiple random initializations***
c. Both A and B
d. None of the above

**15. Which of the following techniques can be used for normalization in text mining?**
a. Stemming
b. Lemmatization
c. Stop Word Removal
***d. Both A and B***

**16. In which of the following cases will K-means clustering fail to give good results?**
*1) Data points with outliers
2) Data points with different densities
3) Data points with non convex shapes*a. 1 and 2
b. 2 and 3
***c. 1, 2, and 3***
d. 1 and 3

**17. Which of the following is a reasonable way to select the number of principal components "k"?**
***a. Choose k to be the smallest value so that at least 99% of the varinace is retained.***
b. Choose k to be 99% of m (k = 0.99\*m, rounded to the nearest integer).
c. Choose k to be the largest value so that 99% of the variance is retained.
d. Use the elbow method

**18. You run gradient descent for 15 iterations with a=0.3 and compute J(theta) after each iteration. You find that the value of J(Theta) decreases quickly and then levels off. Based on this, which of the following conclusions seems most plausible?**
a. Rather than using the current value of a, use a larger value of a (say a=1.0)
b. Rather than using the current value of a, use a smaller value of a (say a=0.1)
***c. a=0.3 is an effective choice of learning rate***
d. None of the above

**19. What is a sentence parser typically used for?**
a. It is used to parse sentences to check if they are utf-8 compliant.
***b. It is used to parse sentences to derive their most likely syntax tree structures.***
c. It is used to parse sentences to assign POS tags to all tokens.
d. It is used to check if sentences can be parsed into meaningful tokens.

**20. Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction ho(x) = 0.2. This means**
a. Our estimate for P(y=1 | x)
***b. Our estimate for P(y=0 | x)***
c. Our estimate for P(y=1 | x)
d. Our estimate for P(y=0 | x)

**21. Which of the following clustering type has characteristic shown in the below figure?**


a) Partitional
***b) Hierarchical***
c) Naive bayes
d) None of the mentioned
***Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.***

**22. Hierarchical clustering is an agglomerative approach & K-means clustering follows partitioning approach.

*23. Hierarchical clustering is deterministic & K-means is not deterministic.***

**24. The process of forming general concept definitions from examples of concepts to be learned.**
A. Deduction
B. abduction
***C. induction***
D. conjunction

**25. Computers are best at learning**
***A. facts.***
B. concepts.
C. procedures.
D. principles.

**26. Data used to build a data mining model.**
A. validation data
***B. training data***
C. test data
D. hidden data

**27. Supervised learning and unsupervised clustering both require at least one**
***A. hidden attribute.***
B. output attribute.
C. input attribute.
D. categorical attribute.

**28. Supervised learning differs from unsupervised clustering in that supervised learning requires**
A. at least one input attribute.
***B. input attributes to be categorical.***
C. at least one output attribute.
D. ouput attriubutes to be categorical.

**29. A regression model in which more than one independent variable is used to predict the dependent variable is called**
A. a simple linear regression model
B. a multiple regression models
***C. an independent model***
D. none of the above

**30. A term used to describe the case when the independent variables in a multiple regression model are correlated is**
A. regression
B. correlation
***C. multicollinearity***
D. none of the above

**31. A multiple regression model has the form: y = 2 + 3x1 + 4x2. As x1 increases by 1 unit (holding x2 constant), y will**
A. increase by 3 units
B. decrease by 3 units
***C. increase by 4 units***
D. decrease by 4 units

**32. A multiple regression model has**
A. only one independent variable
***B. more than one dependent variable***
C. more than one independent variable
D. none of the above

**33. A measure of goodness of fit for the estimated regression equation is the**
A. multiple coefficient of determination
B. mean square due to error
***C. mean square due to regression***
D. none of the above

**34. The adjusted multiple coefficient of determination accounts for**
A. the number of dependent variables in the model
B. the number of independent variables in the model
C. unusually large predictors
***D. none of the above***

**35. The multiple coefficient of determination is computed by**
A. dividing SSR by SST
B. dividing SST by SSR
***C. dividing SST by SSE***
D. none of the above

**36. For a multiple regression model, SST = 200 and SSE = 50. The multiple coefficient of determination is**
A. 0.25
***B. 4.00***
C. 0.75
D. none of the above

**37. A nearest neighbor approach is best used**
A. with large-sized datasets.
***B. when irrelevant attributes have been removed from the data.***
C. when a generalized model of the data is desireable.
D. when an explanation of what has been found is of primary importance.

**38. Determine which is the best approach for each problem.**
*A. supervised learning
B. unsupervised clustering
C. data query*

**38.1. What is the average weekly salary of all female employees under forty years of age? (C)
38.2. Develop a profile for credit card customers likely to carry an average monthly balance of more than $1000.00. (A)
38.3. Determine the characteristics of a successful used car salesperson. (A)
38.4. What attribute similarities group customers holding one or several insurance policies? (A)
38.5. Do meaningful attribute relationships exist in a database containing information about credit card customers? (B)
38.6. Do single men play more golf than married men? (C)
38.7. Determine whether a credit card transaction is valid or fraudulent (A)**

**39. Another name for an output attribute.**
A. predictive variable
***B. independent variable***
C. estimated variable
D. dependent variable

**40. Classification problems are distinguished from estimation problems in that**
A. classification problems require the output attribute to be numeric.
B. classification problems require the output attribute to be categorical.
***C. classification problems do not allow an output attribute.***
D. classification problems are designed to predict future outcome.

**41. Which statement is true about prediction problems?**
A. The output attribute must be categorical.
B. The output attribute must be numeric.
C. The resultant model is designed to determine future outcomes.
***D. The resultant model is designed to classify current behavior.***

**42. Which statement about outliers is true?**
A. Outliers should be identified and removed from a dataset.
B. Outliers should be part of the training dataset but should not be present in the test data.
C. Outliers should be part of the test dataset but should not be present in the training data.
***D. The nature of the problem determines how outliers are used.***
E. More than one of a,b,c or d is true.

**43. Which statement is true about neural network and linear regression models?**
***A. Both models require input attributes to be numeric.***
B. Both models require numeric attributes to range between 0 and 1.
C. The output of both models is a categorical attribute value.
D. Both techniques build models whose output is determined by a linear sum of weighted input attribute values.
E. More than one of a,b,c or d is true.

**44. Which of the following is a common use of unsupervised clustering?**
***A. detect outliers***
B. determine a best set of input attributes for supervised learning
C. evaluate the likely performance of a supervised learner model
D. determine if meaningful relationships can be found in a dataset
E. All of a,b,c, and d are common uses of unsupervised clustering.

**45. The average positive difference between computed and desired outcome values, called as**
A. root mean squared error
B. mean squared error
C. mean absolute error
***D. mean positive error***

**46. Selecting data so as to assure that each class is properly represented in both the training and test set.**
A. cross validation
***B. stratification***
C. verification
D. bootstrapping

**47. The standard error is defined as the square root of this computation.**
***A. The sample variance divided by the total number of sample instances.***
B. The population variance divided by the total number of sample instances.
C. The sample variance divided by the sample mean.
D. The population variance divided by the sample mean.

**48. Data used to optimize the parameter settings of a supervised learner model.**
A. training
B. test
C. verification
***D. validation***

**49. Bootstrapping allows us to**
***A. choose the same training instance several times.***
B. choose the same test set instance several times.
C. build models with alternative subsets of the training data several times.
D. test a model with alternative subsets of the test data several times.

**50. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?**
A. There is no relationship between salary and years worked.
***B. Individuals that have worked for the company the longest have higher salaries.***
C. Individuals that have worked for the company the longest have lower salaries.
D. The majority of employees have been with the company a long time.
E. The majority of employees have been with the company a short period of time.

**51. The correlation coefficient for two real-valued attributes is –0.85. What does this value tell you?**
A. The attributes are not linearly related.
B. As the value of one attribute increases the value of the second attribute also increases.
***C. As the value of one attribute decreases the value of the second attribute increases.***
D. The attributes show a curvilinear relationship.

**52. The average squared difference between classifier predicted output and actual output.**
***A. mean squared error***
B. root mean squared error
C. mean absolute error
D. mean relative error

**53. Simple regression assumes a \_\_\_\_\_\_\_\_\_\_ relationship between the input attribute and output attribute.**
***A. linear***
B. quadratic
C. reciprocal
D. inverse

**54. Regression trees are often used to model \_\_\_\_\_\_\_ data.**
A. linear
***B. nonlinear***
C. categorical
D. symmetrical

**55. The leaf nodes of a model tree are**
A. averages of numeric output attribute values.
B. nonlinear regression equations.
***C. linear regression equations.***
D. sums of numeric output attribute values.

**56. Logistic regression is a \_\_\_\_\_\_\_\_ regression technique that is used to model data having a \_\_\_\_\_outcome.**
A. linear, numeric
B. linear, binary
C. nonlinear, numeric
***D. nonlinear, binary***

**57. This technique associates a conditional probability value with each data instance.**
A. linear regression
***B. logistic regression***
C. simple regression
D. multiple linear regression

**58. This supervised learning technique can process both numeric and categorical input attributes.**
***A. linear regression***
B. Bayes classifier
C. logistic regression
D. backpropagation learning

**59. With Bayes classifier, missing data items are**
A. treated as equal compares.
***B. treated as unequal compares.***
C. replaced with a default value.
D. ignored.

**60. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.**
A. agglomerative clustering
B. expectation maximization
C. conceptual clustering
***D. K-Means clustering***

**61. This clustering algorithm initially assumes that each data instance represents a single cluster.**
A. agglomerative clustering
B. conceptual clustering
***C. K-Means clustering***
D. expectation maximization

**62. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.**
A. agglomerative clustering
B. conceptual clustering
***C. K-Means clustering***
D. expectation maximization

**63. Machine learning techniques differ from statistical techniques in that machine learning methods**
A. typically assume an underlying distribution for the data.
***B. are better able to deal with missing and noisy data.***
C. are not able to explain their behavior.
D. have trouble with large-sized datasets.

**2 marks questions :**

**64. How do you handle missing or corrupted data in a dataset?**
A ) drop missing rows or columns
B) replace missing values with mean/median/mode
C) assign a unique category to missing values
***D) all of the above***

**65. What is the purpose of performing cross validation?**
A) To asses the predictive performance of the models
B) To judge how the trained model performs outside the sample on test data.
***C) Both A & B***

**66. Why is second order differencing in time series needed?**
A) To remove stationarity
B) To find the maxima or minima at the local point
***C) Both A & B***

**67. When performing regression or classification which of the following is the correct way to pre-process the data?**
***Normalize the data > PCA (Principal Component Analysis) > Training***

**68. Which of the following Is an example of feature extraction?**
A) Constructing bag of words vector from an email
B) Applying PCA projects to a large high dimensional data
C) Removing stopwards in a sentence
***D) All of the above***

**69. Which of the following is true about Naïve bay’s algorithm?**
A) Assume that all the features in a data set are equally important
B) Assume that all the features in a data set are independent
***C) Both A & B***

**70. Which of the following statements about regularisation is not correct?**
A) Using too a large a value of lamda can cause your hypothesis to underfit the data
B) Using too a large a value of lamda can cause your hypothesis to overfit the data
C) Using a very large value of lamda cannot hurt the performance of your hypothesis
***D) None of the above***

**71. How can you prevent a clustering algorithm from getting stuck in bad local optima?**
A) Set the same seed value for each run
B) Use multiple random initializations
***C) Both A & B***

**72. Which of the following techniques can be used for normalization in text mining?**
A) Stemming
B) Lemmatization
C) Stopward removal
***D) Both A & B***

**73. In which of the following cases will K means clustering fail to give good results?**
1. Data points with outliers
2. Data points with different densities
3. Data points with non convex shapes
***For all the three cases***

**74. What is a sentence parser typically used for?
*Answer: It is used to parse sentences to derive their most likely syntax tree structures.***

**75. Suppose you have trained a logistic regression classifier and it outputs a new example ‘X’ with a prediction HO(X) = 0.2. This means what?**
***Answer: Our estimate for P(Y) = 0 for X***

**76. What is pca.components\_ in SKlearn?
*Answer: Set of all Eigen vectors for the projection space.***

**77. Which of the following is an example of a deterministic algorithm?
*Answer: PCA*

78. A Pearson correlation between to variables is 0 but their values can still be related to each other?
*Answer: True***

**79. Imagine you are solving a classification problem with highly imbalanced class, the majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on the test data. Which of the following is true in such a case?**
1. Accuracy matrix is not is good idea for imbalanced class problems
2. Accuracy matrix is a good idea for imbalanced class problems
3. Precision and recall matrix are good for imbalanced class problems
4. Precision and recall matrix are not good for imbalanced class problems
***Option 1 & 3 are correct***

**80. Which of the following option is true for overall execution time for 5 fold cross validation with 10 different values of max\_depth?**
***Answer: More than 600 secs***

**81. What would you do in PCA to get the same projection as SVM?**
***Answer: Transform data to zero mean.***

**82. Which of the following value of K will have least leave-one-out cross validation accuracy?**
***Answer: 1-NN***

**83. Which of the following options can be used to get global minima K-means algorithm?**
A) Try to run algorithm for different centroid initialization
B) Adjust number of iterations
C) Find out the optimal no. of clusters
***D) All of the above***

**84. Imagine, you have a 28 \* 28 image and you run a 3 \* 3 convolution neural network on it with the input depth of 3 and output depth of 8.
Note: Stride is 1 and you are using same padding.**
A) 28 width, 28 height and 8 depth
***B) 13 width, 13 height and 8 depth***
C) 28 width, 13 height and 8 depth
D) 13 width, 28 height and 8 depth

**85. A feature F1 can take certain values: A, B, C, D, E & F and represents grade of from a college.
Which of the following statement is true for the above case?**
1. Feature F1 is an example of nominal variable
***2. Feature F1 is an example of ordinal variable***
3. It doesn’t belong to any of the above category
4. Both of these

**86. Assume that there is a blackbox algorithm which takes training data with multiple observations T1, T2, T3,………., Tn and a new observation Q1. The blackbox the nearest neighbour of Q1 say Ti and its corresponding class level Ci. Assume that this blackbox algorithm is same as 1- NN.
It is possible to construct a K-NN classification algorithm based on this blackbox alone where number of training observations is very large compared to K?

*Answer: True***

**87. Assume that there is a blackbox algorithm which takes training data with multiple observations T1, T2, T3,………., Tn and a new observation Q1. The blackbox the nearest neighbour of Q1 say Ti and its corresponding class level Ci. Assume that this blackbox algorithm is same as 1- NN.
Instead of using 1-NN blackbox we want to use the J-NN algorithm for blackbox, where J>1?
Which of the following option is correct for finding K-NN using J-NN?
*A) J must be a proper factor of K.***
B) J must be greater than K.
C) Not possible

**3 Marks Questions :**

**88. Which of the following statement is true ?**
***A. In Gradient Boosting (GD) and Stocharstic Gradient Boosting (SGD), we update set of parameters in an iterative manner.***
B. In Stocharstic Gradient Boosting (SGD), we have to run for all the sample for a single update of parameter in each iteration.
C. In Gradient Boosting (GD), we either use entire data or a subject of training data to update a parameter in each iteration.

**89. Which of the following hyper parameter of Random Forest increase, causes overfit the data ?**
***Answer : depth of tree***

**90. Imagine you are working with analytics vidya and you want to develop a machine learning algorithm which predict no. of views on the article. Your analysis is based on teachers name, author name, no. of articles written by same author on the analytics vidya platform in past etc. Which of the following evalution matrix would you choose in that case ?**
***Answer : Min square error***

**91. Lets say that you are using action funX in hidden layer of neural network at a particular neuron for given input you get the output -0.001. Which of the following activation function should X represent ?**
***Answer : fun 8***

**92. Which of the following are one of the important step to preprocess the text in NLP based project.**
A. stemming
B. stopward removal
C. object standardization
***D. All of these***

**93. Adding a anon important feature to a linear regression method may result in**
***Answer : increase in R^2 (square of R)***

**94. In KNN model, it is very likely to overfit due to cause of dimensionality. Which of the following option would you consider to handle this problem.**
***Answer : dimensionality reduction & feature selection***

**95. Which of the following is true about the gradient boosting tree ?**
A. in each stage introduce a regression tree to compensate the shortcoming of the existing model.
B. we can use gradient distance (GD) method to minimze the loss finction
***C. Both of these***

**96. To apply bagging to regression trees, which of the following are true in that case ?**
A. we build the n regression with n bootstrap sample
B. we take the average of n regression tree
C. each tree has high varience with low biased.
***D. All of the above***

**97. When you find noise in data, which of the following option will you considered in KNN ?**
***Answer : I will increase the value of k.***

**98. Suppose you want to predict the class of the data point, x=1 and y=1 using eucleadian distance in 3NN in which class these data points belongs to ?**
***Answer : positive (+) class***

**99. Which of the following will be eucleadian distance between the two data points A(1,3) and (2,3)**
***Answer : 1***

**100. Suppose you are working on a binary classification problems with three input features and you choose to apply a bagging algorithm X. On this data, you choose max\_features = 2 and the n\_estimators = 3, assume that each estimation has 70% accuracy. Note that algorithm X is aggregating the result of individual estimates based on maximum voting. What will be the maximum accuracy you can get ?**
***Answer : 100%***

**101. In random forest or gradient boosting algorithm, features can be of any type, for example it can be a continuous features or categorical features. Which of the following option is true when you consider this type of feature.**
***Answer : Both the algorithm can handle real valued attributes by discretizing them.***

**102. Which of the following is true about training and testing error in the case described below. Suppose you want to apply Adaboost algorithm on data D which has 'T' observation. You have set half of the data for training and half for testing initially. Now you want to increase the no. of data points for training. [ T1, T2, ------- Tn where T1 > T2 < T3 < ---- < Tn ]**
***Answer : the difference between training error and testing error decreases as the no. of observation increases.***

**103. Suppose you are given 3 variables x, y, z, the pearson corelation coefficient for (x,y), (y,z) & (x,z), c1, c2 and c3 respectively. Now you have added 2 in all the values of X, and substract 2 from all the values of Y and Z remains the same. The new coefficient (x,y), (y,z) & (x,z) are given by D1, D2 and D3 respectively. How do the values of D1, D2, D3 relates to c1, c2, c3 ?**
***Answer : D1=c1, D2=c2, D3=c3.***

**104. Which of the following techniques can be used for the purpose of keyword normalization, the process of converting a keyword into its base form?**
*1. Lemmatization
2. Levenshtein
3. Stemming
4. Soundexbr*

A) 1 and 2
B) 2 and 4
***C) 1 and 3***
D) 1, 2 and 3
E) 2, 3 and 4
F) 1, 2, 3 and 4

**105. Which of the following models can perform tweet classification with regards to context mentioned above?**
A) Naive Bayes
B) SVM
***C) None of the above***

**Explanation : you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.**

**106. What is the major difference between CRF (Conditional Random Field) and HMM (Hidden Markov Model)?**
A) CRF is Generative whereas HMM is Discriminative model
***B) CRF is Discriminative whereas HMM is Generative model***
C) Both CRF and HMM are Generative model
D) Both CRF and HMM are Discriminative model