**Final Exam:** **2053 Big Data Algorithms & Statistics 04.19.2022 - 2 hours**

01. What will happen to Tokens before giving to any Neural Network in NLP (1 marks)

Before the token are passed to the Neural networks they are converted into the numerical representation. This numerical representation is called vectorization. Vectorization is process of converting the words token into numerical form. Here unique words in a document specifies the size of the vector. For each word or token in each document the frequency of its occurrence is noted. Most used vectorization methods are

* Bag of words
* Tf-idf

02. What is overfitting? (1 marks)

A modelling error which tends to ignore all the noises in the data and tightly fits the given training sets in a way that it almost memorises the whole data sets is overfitting. If the models perform well in training data and very poor in test data then there are chances of overfitting. In overfitting the models collect the noise and links it with the features and label. Overfitting is mostly seen in non-linear machine learning approaches.

03. What is the difference between Entropy and Information Gain? (2 marks)

Information gain measures how much information a feature gives us about the class whereas entropy is a measure of uncertainty or disorder in the sample.

Entropy control how a decision tree decides to split the data and effects how a decision tree draws a boundary.

Information gain is significant when you are allowed to choose the variable that provide most information about a class among the given random variables. Whereas entropy tells how much information is encoded in a decision.

04.Why do we require Pruning in Decision Trees? Explain. (2 marks)

Pruning is a part of optimization where it reduces the size of the decision tree by cutting off the branches that do not contribute to the decision of classifying instances.

Mostly used pruning techniques are

* overfitting - pre-pruning
* early stopping and post-pruning

05. What is the difference between Data Mining and Machine Learning? (2 marks)

* Data mining is the process in which rules are derived from large quantities of data where as machine learning is the process where computer are trained in order to provide the result given the parameters to it.
* Data mining is not capable in self learning they provide certain outcome are result from the data gathered where self learning is possible in machine learning, and it derives insights and hidden pattern from the data.
* Generally, Data mining can’t work without human intervention but machine learning can work without the human intervention given that the machine learning pipeline is already built.

06.What are the keyword Normalization techniques in NLP (2 marks)

Lemmatization and stemming.

07. What are Support Vectors in SVM? (1 marks)

Points that are closer to the hyperplane and effects the position and orientation of hyperplane are support vectors. These points help in building Support vector model.

08.What are residuals in simple liner regression (1 marks)

Residual in simple linear regression are the difference between the value of the observed point from the regression line. It is the vertical distance between the observed point and regression line. If the observed point are above the regression line then it is positive residual, if below the line it is negative residual other wise the point lies in the regression line.

09.How many layers are there in the neural network? (2 marks)

To simply put there are three layers in a neural network (Input layer, Hidden Layer, Output layers)

* Input layer: As the name suggest input or data are consumed through this layer.
* Hidden layer: This layer consists of the neurons and the weights are adjusted in each iteration to improve the results.
* Output layer: Via this layer output is provided to the world and in some cases these outputs are back propagated to adjust the weights of the neuron.

10. What is the use of the Activation function? (1 marks)

Activation functions are used to introduce nonlinearity to the neural network. It is used to identify whether the output of the neural network is binary (yes or no, 0 or 1). t converts the values from 0 to 1, -1 to 1, and so o

11.The Sum of squared total (SST) and Sum of Squared Error (SSE) is (1 marks)

a) Equal to 1

b) Equal to zero

c) SSE > SST

d) SST > SSE

12. What are the pre-processing techniques in NLP (1 marks)

* Converting to Lower case.
* Tokenization.
* Removing Punctuation and special characters.
* Stop Word Removal.
* Stemming.
* Lemmatization.

13. List 3 cross validation techniques (2 marks)

* Hold out method
* Leave out One
* K -fold cross validation

14. What is overfitting and underfitting (2 marks)

Over fitting is a modelling error when the model is tightly fitted to the given data sets its like it almost memorise the given data set and perform poor in other datasets.

Underfitting is a modelling error where the model is not able to capture all the feature of the data sets thus performs poor in any given datasets either it is training or testing. It is not able to generalise any datasets.

15. List any 3 activation functions. (3 marks)

* Linear or Identity Activation Function.
* Non-linear Activation Function.
* Sigmoid or Logistic Activation Function.

16. What is a regression model (1 marks)

Y=mx+b

Regression model represents the relation between dependent and independent variable. It gives the equation of the line.

17.What is Backpropagation? (1marks)

The feedback loop that helps to improve the weight of neuron is backpropagation. Artificial neural network will use back propagation as a learning algorithm to compute a gradient descent with respect to weights.

18. List down the different type of nodes in Decision Trees. (2 marks)

* Chances Node: Chances Node provides probability of certain outcomes. They are generally represented with circle in decision tree.
* Decision Nodes: Decision nodes provides the decision to be made and are represented with square.
* End Nodes: It shows the final result of at decision path or branch.

19.List down at least three advantages of Decision Trees. (2 marks)

* As trees can be visualized it is simple to understand and visualize
* It can handle both categorical and numerical data.
* It can handle multi output problems.
* As compared to other techniques it needs less data preparation

20. Explain the difference between K-Means and KNN algorithms? (2 marks)

* The major difference is K-Means is an unsupervised learning clustering algorithms where as KNN is supervised learning classification algorithm.
* K Means is used for classification where as KNN is used for classification and regressions.
* K in K-Means represents number of classes where as K in KNN represents number of nearest neighbour.
* KNN performs better if all the data are of same scale but this is not true in case of K means algorithm.

21. What is the process of identifying people and organization from a given sentence or paragraph is called in NLP? (1 marks)

The process of identifying people and organization from a given sentence or paragraph in NLP is called as Dependency parsing or Part of speech Tagging(POS)

22. What is the main difference between supervised and unsupervised machine learning? (2 marks)

* Dataset used is one of the major differences in supervised and unsupervised learning in supervised learning the data sets are labeled so the machine have examples for which they can learn but it is not the case for unsupervised learning.
* Supervised learning takes feedback from the output and improves itself where as there are no feed back in unsupervised learning.
* Supervised learning predicts output but unsupervised learning finds the hidden pattern in data
* Supervised learning can be categorized to classification and regression problems, unsupervised learning can be categorized to clustering and association problems.

23. Write the equation for (SST) the Sum of squared total

The equation for (SST) the Sum of squared total Total SS = Σ(Yi – mean of Y)2.

24. What is the process of converting a sentence or paragraph into tokens in NLP (1 marks)

Tokenization

25. What is the difference between covariance and correlation? (1 marks)

Covariance provides direction of linear relationship where as correlation provides both directions and strength of the relationship.

26.What is the meaning of term weight initialization in neural networks? (2 marks)

Weight initialization is process of setting the initial weights to the neural networks. They are some small random values assigned for optimization. With these assigned weights the neural network starts learning or training and as they train, they adjust the weight of the neuron that optimizes the result.

27. What are the different types of Learning in Machine learning (1 marks)

* Supervised learning
* Unsupervised learning
* Reinforcement learning

28. Compare Hierarchical Clustering and k-Means Clustering (2 marks)

* Kmeans clustering uses predefined set of clusters and assigns records on based of measure of some distance metrics whereas Hierarchical clustering is done with dendrogram representation.
* K means needs advanced knowledge about the cluster to divide it but in hierarchical clustering cluster can be created at any point in dendrogram.
* K means clustering are less computationally intensive as compared to hierarchical clustering.
* If data set is large enough k means is preferred over hierarchical clutsering.

29.Explain Data Normalization. (1 marks)

The process of transforming data so that the data with higher value does not dominate the result is called Data normalization. Some of them are min max scaling, Z scaling.

30.What is the technique that can be used for keyword normalization in NLP, the process of converting a keyword into its base form? (1 marks)

Lemmatization and stemming can be used for Keyword Normalization,

Stemming is used to convert a keyword into its base form

31. List out at least three disadvantages of Decision Trees. (3 marks)

* Decision trees can create complex trees that do not generalise the data well
* They are unstable, a small variation can lead to drastic change in result
* They might create biases if some classes dominate in dataset.
* They donot express themselves easily so there are concepts that are hard to learn

32. What is Precision and Recall? (2 marks)

Generally, Precision can be seen as measure of quality and recall can be seen as measure of quantity. However, mathematically,

Precision is the ratio of number of positive and number of true positive plus number of false positive.

P=Tp/(Tp+Fp)

Recall is the ratio of number of true positive and number of true positive plus false negative.

R=Tp/(Tp+Fn)