Multiple Choice Questions
On

DEEP LEARNING

University Academy
Teaching|Training|Informative

By

Mr. Sandeep Vishwakarma
Assistant Professor
Dr. A.P.J. Abdul Kalam
Technical University,Lucknow

250+ MCQ

University Academy
Teaching | Training | Informative
YouTube: www.youtube.com/c/UniversityAcademy

1. Computers are best at learning
- **(A)** facts.
- **(B)** concepts.
- **(C)** procedures.
- **(D)** principles.

Correct Answer: A

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

Correct Answer: A

3. 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.

Correct Answer: B

4. Another name for an output attribute.
- **(A)** predictive variable
- **(B)** independent variable
- **(C)** estimated variable
- **(D)** dependent variable

Correct Answer: B

5. 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.

Correct Answer: C

6. 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.

Correct Answer: D

7. 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.

Correct Answer: D

8. Assume that we have a dataset containing information about 200 individuals. One hundred of these individuals have purchased life insurance. A supervised data mining session has discovered the following rule:

IF age < 30 & credit card insurance =

yes THEN life insurance = yes

Rule Accuracy: 70%

Rule Coverage: 63%

How many individuals in the class *life insurance= no* have credit card insurance and are less than 30 years old?

(A) 63

(B) 70

(C) 30

(D) 18

Correct Answer: D

9. 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.

Correct Answer: A

10. Unlike traditional production rules, association rules

(A) allow the same variable to be an input attribute in one rule and

an output attribute in another rule.

(B) allow more than one input attribute in a single rule.

(C) require input attributes to take on numeric values.

(D) require each rule to have exactly one categorical output attribute.

Correct Answer: A

11. 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.

Correct Answer: E

12. The average positive difference between computed and desired outcome values.

(A) root mean squared error

(B) mean squared error

(C) mean absolute error

(D) mean positive error

Correct Answer: C

13. Given desired class C and population P, lift is defined as

(A) the probability of class C given population P divided by the probability of C given a sample taken from the population.

(B) the probability of population P given a sample taken from P.

(C) the probability of class C given a sample taken from population P.

(D) the probability of class C given a sample taken from population P divided by the probability of *C* within the entire population *P*.

Correct Answer: D

14. Which of the following is (are) application(s) of deep learning?
(A) Video captioning
(B) Visual question answering

(C) Video summarization

(D) All of the above
Correct Answer: D

15. Deep learning works well despite of _____problem(s).

(A) High capacity (Susceptible to overfitting)
(B) Numerical instability (vanishing/exploding gradient)
(C) Sharp minima
(D) All of the above
Correct Answer: A

16. An average human brain has around_____neurons(aprox)
(A) 10^2
(B) 10^{11}
(C) 10^{10}
(D) None of the above
Correct Answer: B

17. Which of the following statements is true for the Perceptron Learning Algorithm?

Statement I: Threshold needs to be hand coded

Statement II: Threshold can be learnt

(A) I & II
(B) I
(C) II
(D) None of the statements
Correct Answer: C

18. Sigmoid is
(A) Smooth
(B) Continuous
(C) Differentiable
(D) All of these
Correct Answer: D

19. Which kind of optimisation is required in 2-class linear SVM training procedure?
(A) Linear

(B) Qudratic
(C) Cubic
(D) Depends on dimension of the data

Correct Answer: B

20. The number of Lagrange multiplier required for estimating SVM training parameters depends on

(A) The dimension of the data
(B) The number of classes
(C) The total number of support vectors from training samples
(D) All of the above

Correct Answer: C

21. The distance of a feature vector [2, 3, -2] from separating plane x1+2x2+2x3+5=0 is given by

(A) 5
(B) 3/13
(C) 3
(D) 13

Correct Answer: C

22. For the separable data with D-dimensional feature vectors, K classes and a total of N examples, how many unknown variables are to be estimated in SVM training to get a separating plane 'which maximizes margin? How many constraints?

(A) D variables, N constraints
(B) (D+1) variables, N constraints
(C) D variables, K constraints
(D) (D+1) Variables, K constraints

Correct Answer: B

23.The values of Lagrange multipliers corresponding to the support vectors can be:
(A) Less than zero
(B) Greater than zero
(C) Any real number
(D) Any real non-zero number

Correct Answer: B

24. A data point with 5 dimensions [27,40, -15,30,38] obtain a score [18, 20, -5, -15,19].
Find the Hinge loss incurred by second class (class 2) with a margin of 5 is given by:
(A) 37
(B) 7
(C) 3
(D) 120

Correct Answer: B

25. One form of sigmoid function is logistic function

(A) True
(B) False

Correct Answer: A

26. In sigmoid (logistic) function when WTX=0, then sigmoid value is equal to
    (A) 0
    (B) 1
    (C) ∞
    (D) 0.5
    Correct Answer: D

27. In logistic function, when we increase the value of 'w', the function becomes
    (A) Step function
    (B) Sigmoid function
    (C) quadratic function
    (D) None of these
Correct Answer: B

28. Which parameter(s) need to be learnt in minimizing objective function in supervised learning?

    (A) Only Weight

    (B) Only Bias

    (C) Both Weight and Bias

    (D) Learning rate

    (E) None of these
Correct Answer: C

29 . Which of the following statement is true?
    (A) An event with high probability has high information content
    (B) An event with low probability has high information content
    (C) An event with low probability has low information content
    (D) None of the above
Correct Answer: B

30. Why convolution neural network is taking off quickly in recent times
    (A) Access to large amount of digitized data
    (B) Integration of feature extraction within the training process
    (C) Availability of more computational power
    (D) All of the above
Correct Answer: D

31. Momentum based gradient descent algorithm and Nesterov accelerated gradient descent are faster than Stochastic gradient descent algorithm"
    (A) True
    (B) False
Correct Answer: B

32.Consider the following statement, "It takes less time to navigate the regions having a gentle slope"
The above statement is true in case of
    I.   Gradient descent algorithm
    II.   Momentum based gradient descent algorithm
    (A) I
    (B) II
    (C) II & I
Correct Answer: B

33. Identify the technique that is used to achieve relatively better learning rate by updating w using bunch of different values of η.
   (A) Bias Correction
   (B) Line Search
   (C) Stochastic
   (D) All the above
Correct Answer: B

34. There is no guarantee that the loss decreases at each step in a stochastic Gradient Descent"
   (A) True
   (B) False
Correct Answer: A

35. Identify the advantages of Nesterov accelerated gradient descent.
   I. Corrects its course quicker than Momentum-based gradient descent
   II. Oscillations are smaller
   III. Chances of escaping minima valley are also smaller
   (A) I
   (B) only II only
   (C) IIand III
   (D) I, II and III
Correct Answer: D

36. Pick out the methods for annealing learning rate that has only number of epochs as the hyperparameter.
   (A) Step decay
   (B) Exponential Decay
   (C) 1/t Decay
Correct Answer: A

37.Adagrad got stuck when it was close to convergence. How does RMSProp overcome this problem?
   (A) More Aggressive on decay
   (B) Less Aggressive on decay
   (C) No decay
Correct Answer: B

38.Which of the following gradient descent algorithm suffers from more oscillations?

   (A) Momentum based gradient descent
   (B) Nesterov accelerated gradient descent
   (C) Vanilla gradient descent
   (D) None of the above
Correct Answer: A

39. In a neural network, knowing the weight and bias of each neuron is the most important step. If you can somehow get the correct value of weight and bias for each neuron, you can approximate any function. What would be the best way to approach this?

   (A) Assign random values and pray to God they are correct
   (B) Search every possible combination of weights and biases till you get the best value
   (C) Iteratively check that after assigning a value how far you are from the best values, and slightly change the assigned values values to make them better
   (D) None of these
Correct Answer: C

40..What are the steps for using a gradient descent algorithm?

1. Calculate error between the actual value and the predicted value
2. Reiterate until you find the best weights of network
3. Pass an input through the network and get values from output layer
4. Initialize random weight and bias
5. Go to each neurons which contributes to the error and change its respective values to reduce the error

(A) 1, 2, 3, 4, 5
(B) 5, 4, 3, 2, 1
(C) 3, 2, 1, 5, 4
(D) 4, 3, 1, 5, 2
Correct Answer: D

41.."Convolutional Neural Networks can perform various types of transformation (rotations or scaling) in an input". Is the statement correct True or False?

(A) True
(B) False
Correct Answer: B

42 .Which of the following techniques perform similar operations as dropout in a neural network?
(A) Bagging
(B) Boosting
(C) Stacking
(D) None of these
Correct Answer: A

43. Which of the following gives non-linearity to a neural network?
(A) Stochastic Gradient Descent
(B) Rectified Linear Unit
(C) Convolution function
(D) None of the above
Correct Answer: B

44.Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions) ?
(A) As number of hidden layers increase, model capacity increases
(B) As dropout ratio increases, model capacity increases
(C) As learning rate increases, model capacity increases
(D) None of these
Correct Answer: A

45. If you increase the number of hidden layers in a Multi Layer Perceptron, the classification error of test data always decreases. True or False?
(A) True
(B) False
Correct Answer: B

46. Can a neural network model the function $(y=1/x)$?
(A) Yes
(B) No
Correct Answer: A

47. In which neural net architecture, does weight sharing occur?

(A) Convolutional neural Network
(B) Recurrent Neural Network
(C) Fully Connected Neural Network
(D) Both A and B
Correct Answer: D

48. Batch Normalization is helpful because
(A) It normalizes (changes) all the input before sending it to the next layer
(B) It returns back the normalized mean and standard deviation of weights
(C) It is a very efficient backpropagation technique
(D) None of these
Correct Answer: A

49. Instead of trying to achieve absolute zero error, we set a metric called bayes error which is the error we hope to achieve. What could be the reason for using bayes error?
(A) Input variables may not contain complete information about the output variable
(B) System (that creates input-output mapping) may be stochastic
(C) Limited training data
(D) All the above
Correct Answer: D

50.The number of neurons in the output layer should match the number of classes (Where the number of classes is greater than 2) in a supervised learning task. True or False?
(A) True
(B) False
Correct Answer: B

51.In a neural network, which of the following techniques is used to deal with overfitting?
(A) Dropout

(B) Regularization
(C) Batch Normalization
(D) All of these
Correct Answer: D

52.$Y = ax^2 + bx + c$ (polynomial equation of degree 2)

Can this equation be represented by a neural network of single hidden layer with linear threshold?

(A) Yes
(B) No
Correct Answer: B

53. What is a dead unit in a neural network?

(A) A unit which doesn't update during training by any of its neighbour
(B) A unit which does not respond completely to any of the training patterns
(C) The unit which produces the biggest sum-squared error
(D) None of these
Correct Answer: A

54 .Which of the following statement is the best description of early stopping?
(A) Train the network until a local minimum in the error function is reached
(B) Simulate the network on a test dataset after every epoch of training. Stop training when

the generalization error starts to increase
- **(C)** Add a momentum term to the weight update in the Generalized Delta Rule, so that raining converges more quickly
- **(D)** A faster version of backpropagation, such as the `Quickprop' algorithm

Correct Answer: B

55. What if we use a learning rate that's too large?
- **(A)** Network will converge
- **(B)** Network will not converge
- **(C)** Can't Say

Correct Answer: B

56. Which gradient technique is more advantageous when the data is too big to handle in RAM simultaneously?
- **(A)** Full Batch Gradient Descent
- **(B)** Stochastic Gradient Descent

Correct Answer: B

57. What are the factors to select the depth of neural network?

1. Type of neural network (eg. MLP, CNN etc)
2. Input data
3. Computation power, i.e. Hardware capabilities and software capabilities
4. Learning Rate
5. The output function to map

- **(A)** 1, 2, 4, 5
- **(B)** 2, 3, 4, 5
- **(C)** 1, 3, 4, 5
- **(D)** All of these

Correct Answer: D

58. Consider the scenario. The problem you are trying to solve has a small amount of data. Fortunately, you have a pre-trained neural network that was trained on a similar problem. Which of the following methodologies would you choose to make use of this pre-trained network?
- **(A)** Re-train the model for the new dataset
- **(B)** Assess on every layer how the model performs and only select a few of them
- **(C)** Fine tune the last couple of layers only
- **(D)** Freeze all the layers except the last, re-train the last layer

Correct Answer: D

59. Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional network.
- **(A)** True
- **(B)** False

Correct Answer: B

60. Which kind of activation function is typical for a convolution layer in an RNN
- **(A)** Gaussian
- **(B)** Sigmoid
- **(C)** Hyperbolic Tangent
- **(D)** ReLU

Correct Answer: D

61. Which of the following method use arithmetic operation in evaluation of word representation?
- **(A)** Semantic relatedness

(B) Synonym detection
(C) Semantic analogy
(D) None of the above

Correct Answer: C

62. Which of the following model directly learn word representations?
   (A) Prediction based model
   (B) Count based model
   (C) Both prediction and count based model
   (D) None of these

Correct Answer: A

63. Which of the following is correct in word representation model?
   I. In continuous bag of word model, the softmax function is computationally expensive.
   II. In continuous bag of word model, the softmax function is computationally inexpensive.
   III. In Skip-gram model, the softmax function is computationally inexpensive.
   IV. In Skip-gram model, the softmax function is computationally expensive

   (A) I only
   (B) II & III
   (C) III only
   (D) I & IV

Correct Answer: D

64. Which of the following solution constructs a binary tree in learning word representation using prediction- based models?
   (A) Use negative sampling
   (B) Use contrastive estimation
   (C) Use hierarchical softmax
   (D) None of these

Correct Answer: C

65. Which of the following method(s) use dot product in the evaluation of word representation?
   I. Semantic relatedness
   II. Synonym detection
   III. Semantic analogy

   (A) I & II
   (B) III
   (C) I, II, & III
   (D) None of the above

Correct Answer: A

66. If you need to design a model for textual entailment from the text, then which of the following steps will you choose?
   I. CNN is used to encode the text
   II. CNN is used to decode the text
   III. RNN is used to decode the text from the encoding
   IV. RNN is used to encode the text.
   V. RNN is used to encode the text and decode the text.

   (A) IV, II
   (B) I, III
   (C) V
   (D) None of these

Correct Answer: C

67. The problem of generating the sentence given an image can be possibly solved with the encoder-decoder architecture.

(A) Yes

(B) No

Correct Answer: A

68. For document classification and summarization, it is important to look at the important sentences and important words. What kind of "attention" mechanism is required for encoding?

(A) Hierarchical

(B) Ungraded

(C) Sequential

(D) Unordered

Correct Answer: A

69. 48 filters of size 21 x 21 is applied to an image of size 327 x 327, with zero padding and stride of 3. The image is an RGB image. The depth of the filter is same as the depth of image. What will be the volume of the final image?

(A) 103 x 103 x 3

(B) 103 x 103 x 48

(C) 327 x 327 x 3

(D) 327 x 327 x 48

Correct Answer: B

70. Consider the following:

W= [0.2, 0.7, 0.05, 0.75, 0.86, 0.21]

X= [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

What output St is obtained by sliding the filter Wt over the input Xt ?

(A) [1.306, 1.421, 1.567, 2.002]

(B) [1.031, 1.308, 1.585, 1.862]

(C) [2.345, 2.121, 3.547, 2.409]

(D) [2.127, 2.229, 3.212, 3.421]

Correct Answer: B

71. While reconstructing images from embeddings, if we consider deeper layers of the network, how is the image affected?

(A) The reconstructed image is blur

(B) A more realistic copy of the original image is reconstructed

(C) The reconstructed image is sharpened

(D) The reconstructed image becomes abstract

Correct Answer: D

72. Which of the following is true for content and style of image w.r.t. Deep Art?

(A) Content should match and Style shouldn't match

(B) Content shouldn't match and Style should match

(C) Both Content and Style should match

(D) Both Content and Style shouldn't match

Correct Answer: C

73. It is very easy to fool a CNN. Which of the following statements doesn't support the above argument?

(A) The CNN draws very stringent/strict boundaries for classification.

(B) The decision boundaries are loosely drawn.

(C) The input dimension for a certain class is very very large.

(D) None

Correct Answer: A

74. Which of the following includes major tasks of NLP?

(A) Discourse Analysis
(B) Automatic Summarization
(C) Machine Translation
(D) All of the above
Correct Answer: D

76. Which of the following is used to mapping sentence plan into sentence structure?
(A) Text planning
(B) Sentence planning
(C) Text Realization
(D) None of the Above
Correct Answer: C

76. Given a sound clip of a person or people speaking, determine the textual representation of the speech.

(A) Text-to-speech
(B) Speech-to-text
(C) All of the mentioned
(D) None of the mentioned
Correct Answer: B

77._____ is the step in which an input sentence is converted into a hierarchical structure that corresponds to the units of meaning in the sentence.

(A) Syntactic Processing
(B) Semantic Processing
(C) Graph Processing
(D) All of the mentioned
Correct Answer: A

78. Which of the following statement about Skip-gram model are correct?

(A) It predicts the center word from the surrounding context words
(B) The final word vector for a word is the average or sum of the input vector v and output vector u corresponding to that word
(C) When it comes to a small corpus, it has better performance than GloVe
(D) It makes use of global co-occurrence statistics

Correct Answer: B

79. Which of the following statements is true of language models?

(A) Neural window-based models share weights across the window
(B) Neural window-based language models suffer from the sparsity problem, but n-gram language models do not
(C) The number of parameters in an RNN language model grows with the number of time step
(D) Neural window-based models can be parallelized, but RNN language models cannot
Correct Answer: D

80 Suppose that you are training a neural network for classification, but you notice that the training loss is much lower than the validation loss. Which of the following can be used to address the issue (select all that apply)?

(A) Use a network with fewer layers

(B) Decrease dropout probability

(C) Increase L2 regularization weight

(D) Increase the size of each hidden layer

Correct Answer: A

81. Which of the following scientists created the first Bioinformatics database?

(A) Dayhoff

(B) Pearson

(C) Richard Durbin

(D) Michael.J.Dunn

Correct Answer: A

82. The human genome contains approximately_____.

(A) 6 billion base pairs

(B) 5 billion base pairs

(C) 3 billion base pairs

(D) 4 billion base pairs

Correct Answer: C

83. The term Bioinformatics was coined by_____.

(A) J.D Watson

(B) Pauline Hogeweg

(C) Margaret Dayhoff

(D) Frederic Sanger

Correct Answer: B

84. The laboratory work using computers and associated with web-based analysis generally online is referred to as_____.

(A) In silico

(B) Dry lab

(C) Wet lab

(D) All of the above

Correct Answer: A

85. The laboratory work using computers and computer-generated models generally offline is referred to as_____.

(A) Insilico

(B) Wet lab

(C) Dry lab

(D) All of the above

Correct Answer: C

86. A type or class of biometric system, such as face recognition, fingerprint recognition,

iris recognition, etc.

(A) Nodal Points

(B) Modality

(C) Model

(D) Identity

Correct Answer: B

87. What enables people to recognize people, animals and inanimate objects reliably?

(A) Speech

(B) Vision

(C) Hear

(D) Perception

Correct Answer: B

88. Which provides a framework for studying object recognition?

(A) Learning
(B) Unsupervised learning
(C) Supervised learning
(D) None of the mentioned

Correct Answer: C

89. How the distance between two shapes can be defined?
(A) Weighted sum of the shape
(B) Size of the shape
(C) Shape context
(D) None of the mentioned

Correct Answer: A

90. Which of the following is a bottleneck for deep learning algorithm?
(A) Data related to the problem
(B) CPU to GPU communication
(C) GPU memory
(D) All of the above

Correct Answer: D

91. Which of the following would be the best for a non-continuous objective during optimization in deep neural net?
(A) L-BFGS
(B) SGD
(C) AdaGrad
(D) Subgradient Method

Correct Answer: D

92. A recurrent neural network can be unfolded into a full-connected neural network with infinite length
(A) True
(B) False

Correct Answer: A

93. Backpropagation works by first calculating the gradient of ___ and then propagating it backwards.
(A) Sum of squared error with respect to inputs
(B) Sum of squared error with respect to weights
(C) Sum of squared error with respect to outputs
(D) None of the above

Correct Answer: A

94. Inductive learning is based on the knowledge that if something happens a lot it is likely to be generally _____
   (A) True
   (B) False

Correct Answer: B

95. Outputs of learning are determined by the _____
   (A) Application
   (B) Validation
   (C) Training
   (D) None of the given

Correct Answer: A

96. The tractable problems are further divided into structured and _____ problems
   (A) Non-structured
   (B) Complex
   (C) Simple
   (D) None of the given

Correct Answer: B

97. Given above is a description of a neural network. When does a neural network model become a deep learning model?
   (A) When you add more hidden layers and increase depth of neural network
   (B) When there is higher dimensionality of data
   (C) When the problem is an image recognition problem
   (D) None of these

Correct Answer: A

98. Suppose a convolutional neural network is trained on ImageNet dataset (Object recognition dataset). This trained model is then given a completely white image as an input.The output probabilities for this input would be equal for all classes. True or False?
   (A) True
   (B) False

Correct Answer: B

99. Which of the following gives non-linearity to a neural network?
   (A) Stochastic Gradient Descent
   (B) Rectified Linear Unit
   (C) Convolution function
   (D) None of the above

Correct Answer: B

100, In which neural net architecture, does weight sharing occur?
   (A) Convolutional neural Network
   (B) Recurrent Neural Network
   (C) Fully Connected Neural Network
   (D) Both A and B

Correct Answer: D

101. 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

Correct Answer: A

102. In determination of weights by learning, for noisy input vectors what kind of learning should be employed?.

   (A) hebb learning law

   (B) widrow learning law

   (C) hoff learning law

   (D) no learning law

Correct Answer: D

103.Number of output cases depends on what factor?

   (A) number of inputs

   (B) number of distinct classes

   (C) total number of classes

   (D) none of the mentioned

Correct Answer: B

104. Which architecture of neural network would be better suited to solve the problem?

   (A) End-to-End fully connected neural   network

   (B) Convolutional neural network followed by recurrent units

   (C) Neural Turing Machine

   (D) All of above. Correct

Correct Answer: B

105. Which of the following is a representation learning algorithm?

   (A) Neural network

   (B) Random Forest

   (C) k-Nearest neighbor

   (D) All of the above

Correct Answer: A

106. Perplexity is a commonly used evaluation technique when applying deep learning for NLP tasks. Which of the following statement is correct?

   (A) Higher the perplexity the better

   (B) Lower the perplexity the better

Correct Answer: B

107. What is generally the sequence followed when building a neural network architecture for semantic segmentation for image?

   (A) Convolutional network on input and   deconvolutional network on output

   (B) Deconvolutional network on input and   convolutional network on output

Correct Answer: A

108. Deep learning can be applied to which of the following NLP tasks?

   (A) Machine translation

   (B) Sentiment analysis

   (C) Question Answering system

   (D) All of the above

Correct Answer: D

109. What is the technical difference between vanilla backpropagation algorithm and backpropagation through time (BPTT) algorithm?

  Unlike backprop, in BPTT we sum   up gradients for corresponding weight for each time step

  Unlike backprop, in BPTT we subtract gradients for corresponding weight for each time step

Correct Answer: A

Practice Question
(Option In bold font is Answer)

110. Which of the following is FALSE about Deep Learning and Machine Learning algorithms?
    a. Deep Learning algorithms work efficiently on high amount of data
    b. **Feature Extraction needs to be done manually in both ML and DL algorithms**
    c. Deep Learning algorithms are best suited for unstructured data
    d. Deep Learning algorithms require high computational power

111. Which of the following is FALSE about Deep Learning and Machine Learning algorithms?
    a. Data augmentation can be done easily in Deep Learning as compared to Machine Learning
    b. Deep Learning algorithms efficiently solve computer vision problems
    c. **Deep Learning algorithms are more interpretable as compared to Machine Learning algorithms**
    d. None of the above

112. Which of the following is TRUE about Perceptrons?
    a. It is an artificial neuron and is the fundamental unit of a neural network
    b. It is also called single layer neural network or search single layer binary linear classifier
    c. A neural network is a composition of perceptrons connected in different ways
    d. All of the above

113. Which of the following is FALSE about Perceptrons?
    a. **Perceptron can learn both linearly and non-linearly separable functions**
    b. It cannot implement XOR gate as it cannot be classified by a linear separator
    c. It can easily implement OR and AND gates as these are linearly separable
    d. None of the above

114. Which of the following is the structure of the input to an artificial neuron?
    a. Weighted sum of inputs + bias
    b. Sum of inputs + bias
    c. **Weighted sum of bias + inputs**
    d. Sum of bias + inputs

115. Which of the following is FALSE about Neural Networks?
    a. During backward propagation, we update the weights using gradient descent algorithm
    b. We can use different activation functions in different layers
    c. **We can use different gradient descent algorithms in different epochs**
    d. None of the above

116. Which of the following is FALSE about Hidden Layers in Neural Networks?
    a. Abstract representation of the training data is stored in the hidden layers
    b. Feature extraction happens at the hidden layers
    c. **Increasing the number of hidden layers always leads to higher accuracy**
    d. Increasing the number of hidden layers above a certain point may lead to overfitting

117. Which of the following is FALSE about Deep Neural Networks?
    a. These are computationally more complex as compared to shallow networks

b. These have less generalization capabilities as compared to shallow networks
   c. These may suffer from overfitting problem
   d. These may suffer from vanishing gradients problem

118.   Which of the following is TRUE about Shallow Neural Networks?
   a. These are computationally more complex as compared to deep neural networks
   b. These have better generalization capabilities as compared to deep neural networks
   c. These are better in learning and extracting features as compared to deep neural networks
   d. None of the above

119.   Which of the following is a type of neural network?
   a. CNN (Convolutional Neural Network)
   b. Capsule Neural Networks
   c. Autoencoders
   d. All of the above

120.   Which of the following is a type of neural network?
   a. RNN (Recurrent Neural Network)
   b. LSTM (Long Short-Term Memory Networks)
   c. Restricted Boltzmann Machines (RBM)
   d. All of the above

121.   Which of the following is FALSE about Radial Basis Function Neural Network?
   a. It resembles to RNNs which have feedback loops
   b. It uses radial basis function as activation function
   c. While outputting, it considers the distance of a point with respect to the center
   d. None of the above

122.   Which of the following components make a neural network non-linear in nature?
   a. Hidden Layers
   b. Activation Functions
   c. Weights and Bias
   d. Regularization and Dropout

123.   What is the purpose of an activation function?
   a. To decide whether a neuron will fire or not
   b. To increase the depth of a neural network
   c. To create connectivity among hidden layers
   d. To normalize the inputs

124.   Which of the following is FALSE about activation functions?
   a. Activation functions help in achieving non-linearity in deep neural networks
   b. Activation functions help in reducing overfitting problem
   c. These are also called squashing functions as these squash the output under a certain range
   d. Commonly used activation functions are step, sigmoid, tanh, ReLU and softmax

125.   Which of the following is FALSE about step activation function?
   a. It either outputs 0 or 1
   b. It is linear in nature
   c. It is also called Threshold activation function
   d. None of the above

126.   Which of the following is FALSE about step activation function?
   a. It is a differentiable activation function
   b. Derivative of a step function is always zero

    c. It is non-linear in nature

    d. None of the above

127.   Which of the following is FALSE about sigmoid and tanh activation function?

    a. Both are non-linear activation functions

    b. Output of sigmoid ranges from 1 to 1 while output of tanh ranges from 0 to 1

    c. Output of both sigmoid and tanh is smooth, continuous and differentiable

    d. None of the above

128.   Which of the following is FALSE about sigmoid and tanh activation function?

    a. These cannot blow up the activations unlike ReLU

    b. Both functions output a probability value instead of discrete values like 0 and 1

    c. Sigmoid is zero centered as its values range from 1 to 1

    d. None of the above

129.   Which of the following is FALSE about sigmoid and tanh activation function?

    a. These do not suffer from vanishing and exploding gradient problems unlike ReLU

    b. These involve computing the exponential (which may be expensive) unlike ReLU

    c. These are non-linear in nature like ReLU

    d. None of the above

130.   Output of step (threshold) activation function ranges from:

    a. Either 0 or 1

    b. 0 to 1

    c. -1 to 1

    d. Either -1 or 1

131.   Output of sigmoid activation function ranges from:

    a. 0 to 1

    b. .-1 to 1

    c. -1 to o

    d. 0 to 9

132.   Output of tanh activation function ranges from:

    a. 0 to 1

    b. -1 to 1

    c. -1 to o

    d. 0 to 9

133.   Output of which of the following activation functions is zero centered?

    a. Hyperbolic Tangent

    b. Sigmoid

    c. Softmax

    d. ReLU

134.   ReLU activation function outputs zero when:

    a. Input is zero

    b. Input is less than or equal to zero

    c. Input is greater than or equal to zero

    d. Input is zero or one

135.   ReLU outputs same value as input when:

    a. Input is greater than zero

    b. Input is greater than or equal to zero

    c. Input lies between zero and one

    d. Input is zero or one

136. Which of the following is FALSE about ReLU activation function?
    a. It does not require exponent calculation unlike sigmoid and tanh
    b. It outputs 0, if the input is less than or equal to 0, otherwise output will be same as the input
    c. ReLU is preferred over sigmoid and tanh for output layer in a neural network
    d. None of the above
137. Which of the following is FALSE about Dying ReLU?
    a. In this situation, neurons become inactive and only output 0 for any input
    b. Dying ReLU neurons do not play any role in discriminating the input and become useless
    c. Dying ReLU problem occurs when learning rate is too low
    d. None of the above
138. Dying ReLU issue occurs when
    a. Learning rate is too high
    b. There is a large negative bias
    c. Any of the above occurs
    d. None of the above
139. Which of the following is NOT a solution for Dying ReLU?
    a. Batch normalization
    b. Dropout
    c. Leaking ReLU
    d. Low learning rate
140. Which of the following is a solution for Dying ReLU?
    a. Leaking ReLU
    b. Max ReLU
    c. Parametric ReLU (PRELU)
    d. All of the above
141. Which of the following is TRUE about Leaking ReLU?
    a. Leaky ReLU is the most common method to alleviate a dying ReLU
    b. Leaking ReLU adds a slight slope in the negative range to prevent dying ReLU issue
    c. Parametric ReLU (PRELU) is a type of leaky ReLU that uses parameterized slope value
    d. All of the above
142. Which of the following is FALSE about Softmax function?
    a. Softmax function is a generalized version of a sigmoid function
    b. Softmax function is usually used in the hidden layers of neural networks
    c. Softmax function is usually used for multi-class classification problem
    d. None of the above
143. Which of the following is TRUE about Softmax and Sigmoid function?
    a. Sum of probabilities of all softmax units are supposed to be 1
    b. Sum of probabilities of all sigmoid units are supposed to be 1
    c. Sum of probabilities of both sigmoid and softmax units are supposed to be 1
    d. All of the above
144. Which of the following is TRUE about Softmax and Sigmoid function?
    a. Softmax is usually used for hidden layers and sigmoid for outer layers
    b. Sigmoid is usually used for hidden layers and softmax for outer layers
    c. Softmax function is usually used for binary classification problem
    d. All of the above

145. What is the purpose of the Gradient Descent algorithm?
   a. To normalize the inputs
   b. To minimize the weights and bias
   c. To minimize the loss function
   d. To prevent model from overfitting
146. Gradient Descent computes derivative of loss function w.r.t:
   a. input
   b. activation value
   c. weight
   d. bias
147. Which of the following is TRUE about Batch Gradient Descent?
   a. Performs model updates at the end of each training epoch
   b. Requires a lot of memory while gradient computation
   c. Training speed usually becomes very slow for large datasets
   d. All of the above
148. Which of the following is FALSE about Stochastic Gradient Descent?
   a. Performs model updates for each single observation
   b. Frequent model updates can also result in a noisy gradient signal
   c. Learning is slower than batch gradient descent
   d. None of the above
149. Which of the following is FALSE about Mini Batch Gradient Descent?
   a. Performs model updates per batch of training data
   b. Batch size is a hyper-parameter which needs to be fine-tuned
   c. It maintains a balance between BGD and SGD
   d. None of the above
150. Which of the following is TRUE about Momentum?
   a. It helps in accelerating SGD in a relevant direction
   b. It helps to prevent unwanted oscillations by adding up the speed
   c. It helps to know the direction of the next step with the knowledge of the previous steps
   d. All of the above
151. Which of the following is TRUE about Momentum?
   a. It helps in accelerating SGD in a relevant direction
   b. It helps SGD in avoiding local minima
   c. It helps in faster convergence
   d. All of the above
152. Which of the following SGD variants is momentum based?
   a. Nesterov
   b. Adagrad
   c. RMSprop
   d. All of the above
153. Which of the following SGD variants is NOT momentum based?
   a. Nesterov
   b. AdaDelta
   c. Adam
   d. All of the above are momentum based
154. Which of the following SGD variants is NOT based on adaptive learning?
   a. Nesterov

      b. Adagrad
      c. AdaDelta
      d. RMSprop

155. Which of the following SGD variants is based on adaptive learning?
      a. Adam
      b. Adagrad
      c. AdaDelta
      d. All of the above

156. Which of the following SGD variants is based on both momentum and adaptive learning?
      a. RMSprop
      b. Adagrad
      c. Adam
      d. Nesterov

157. Which of the following SGD variants is based on both momentum and adaptive learning?
      a. RMSprop
      b. Adagrad
      c. AdaDelta
      d. None of the above

158. Which of the following is FALSE about Adagrad, AdaDelta, RMSprop and Adam?
      a. AdaDelta and RMSprop are an extension of Adagrad
      b. RMSprop considers both momentum and adaptive learning
      c. Adagrad mainly focuses on adaptive learning rate instead of momentum
      d. None of the above

159. Which of the following is TRUE about Nesterov, RMSprop and Adam?
      a. A Nesterov and RMSprop are an extension of Adam
      b. B. Nesterov considers both momentum and adaptive learning
      c. C. RMSprop mainly focuses on adaptive learning rate instead of momentum
      a. All of the above

160. CNN is best suited for:
      a. Image Classification
      b. Natural Language Processing
      c. Image Captioning
      d. All of the above

161. Which of the following layers is NOT a part of CNN?
      a. Convolutional Layer
      b. Pooling Layer
      c. Code Layer
      d. Fully connected Layer

162. Which of the following terms is NOT associated with CNN?
      a. Filters (Kernels)
      b. Forget Gates
      c. Zero and Valid Padding
      d. Strides

163. Which of the following is FALSE about CNN?
      a. We must flatten the output before feeding it to a fully connected layer
      b. There can be only one fully connected layer in CNN

    c. We can use as many convolutional and pooling layers in CNN

    d. None of the above

164. Which of the following is TRUE about Convolutional Layer in CNN?

    a. It is used to extract features from an input image

    b. It uses a matrix filter and performs convolution operation to detect patterns in the image

    c. We may lose some information about edges and corners of an image during convolution

    d. All of the above

165. Which of the following is FALSE about Kernels in CNN?

    a. Kernels can be used in convolutional as well as in fully connected layers

    b. Kernels act as pattern detectors

    c. Kernels lead to dimensionality reduction

    d. We may lose some information about edges and corners of an image using Kernels

166. Which of the following is FALSE about Kernels in CNN?

    a. Kernels can be used in convolutional as well as in pooling layers

    b. Kernels keep sliding over an image to extract different components or patterns of an image

    c. Kernels extract simple features in initial layers and complex features in deeper layers

    d. None of the above

167. Which of the following is FALSE about Padding in CNN?

    a. Padding is used to prevent the loss of information about edges and corners during convolution

    b. There are two types of padding: Zero Padding and Valid Padding (no padding)

    c. There is no reduction in dimension when we use valid padding

    d. In zero padding, we pad the image with zeros so that we do not lose any edge information

168. Which of the following is FALSE about Padding in CNN?

    a. We should use valid padding if we know that information at edges is not that much useful

    b. We compromise to lose some edge information of the image in zero padding

    c. There is no reduction in dimension when we use zero padding

    d. In valid padding, we drop the part of the image where the filter does not fit

169. Which of the following is FALSE about zero padding?

    a. It is used to preserve the spatial size of the input volume

    b. It is used to preserve edge information of the image

    c. It is used to preserve resolution of the image

    d. None of the above

170. Which of the following is TRUE about Padding in CNN?

    a. Padding is used in convolution layer as well as in pooling layer

    b. Padding is used in convolution layer as well as in fully connected layer

    c. Padding is used in fully connected layer as well as in pooling layer

    d. Padding is used only in convolution layer

171. Filter of size 3X3 is rotated over input matrix of size 4X4 (stride=1). What will be the size of output matrix after applying zero padding?

    a. 4X4

    b. 3X3

    c. 2X2

       d. 1X1
172.   Filter of size 3X3 is rotated over input matrix of size 4X4 (stride=1). What will be the size of output matrix after applying valid padding?
    a. 4X4
    b. 3X3
    c. 2X2
    d. 1X1
173.   Which of the following is FALSE about Pooling Layer in CNN?
    a. Pooling layer must be added after each convolutional layer
    b. Output of convolutional layer acts as an input to the pooling layer
    c. It does down-sampling of an image which reduces dimensions by retaining vital information
    d. It does feature extraction and detects components of the image like edges, corners etc.
174.   Which of the following is TRUE about Pooling Layer in CNN?
    a. We can use Max, Min, Average or Sum pooling in CNN
    b. It helps in retaining the most useful information and throwing away useless information
    c. It reduces resolution and dimension and hence reduces computational complexity
    d. All of the above
175.   Which of the following is FALSE about Pooling Layer in CNN?
    a. It helps in reducing overfitting
    b. It reduces computational complexity
    c. It increases image resolution
    d. None of the above
176.   Which of the following is NOT a hyper-parameter in CNN?
    a. Code size for compression
    b. Number and size of kernels in a convolutional layer
    c. Padding in a convolutional layer (zero or valid padding)
    d. Number of convolutional layers
177.   Which of the following model is best suited for sequential data?
    a. Convolutional Neural Networks (ConvNots)
    b. Capsule Neural Networks (CapsNots)
    c. RNN (Recurrent Neural Network)
    d. Autoencoders
178.   Which of the following model contains internal memory?
    a. Convolutional Neural Networks (ConvNets)
    b. Capsule Neural Networks (CapsNets)
    c. RNN (Recurrent Neural Network)
    d. Autoencoders
179.   Which of the following is FALSE about ISTM?
    a. LSTM is an extension for RNA which extends its memory
    b. LSTM enables RNN to learn long-term dependencies
    c. LSTM solves the exploding gradients issue in RNN
    d. None of the above
180.   Which of the following is TRUE about LSTM?
    a. It uses forget gates, input gates and output gates
    b. These gates are analog in nature
    c. It uses feedback loop to remember the data

    d. All of the above

181. Which of the following is FALSE about gated cells in LSTM?
    a. LSTMs store information in these gated cells
    b. These gates are digital in nature
    c. Gates of these cells open and close based on some decisions
    d. Data can be stored, deleted and read from these gated cells like computer storage

182. Which of the following is FALSE about Forget Gate in LSTM?
    a. It decides what information it needs to forget or throw away
    b. It outputs a number between 0 and 1
    c. 0 represents completely keep this info while 1 represents completely forget this info
    d. None of the above

183. Which of the following types of RNN is also called Vanilla Neural Network?
    a. One to One
    b. One to Many
    c. Many to One
    d. Many to Many

184. Which of the following types of RNN is best suited for image captioning?
    a. One to One
    b. One to Many
    c. Many to One
    d. Many to Many

185. Which of the following types of RNN is best suited for sentiment analysis?
    a. One to One
    b. One to Many
    c. Many to One
    d. Many to Many

186. Which of the following types of RNN is best suited for machine translation?
    a. One to One
    b. One to Many
    c. Many to One
    d. Many to Many

187. Which of the following is NOT an application of RNN?
    a. Time series prediction
    b. Anomaly detection
    c. Weather prediction
    d. Stock market prediction

188. Which of the following is NOT an application of RNN?
    a. Algorithmic trading
    b. Understanding DNA sequence
    c. Image compression
    d. Image captioning

189. Which of the following is an application of RNN?
    a. Text mining
    b. Sentiment analysis
    c. Text and Speech analysis
    d. All of the above

190. Which of the following is an application of RNN?

a. Natural Language Processing
b. Audio and Video analysis
c. Stock market prediction
d. All of the above

191. Which of the following is NOT a generalization technique in neural networks?
   a. Xavier Weight Initialization
   b. Dropout
   c. Li and L2 Regularization
   d. Data Augmentation

192. Which of the following is FALSE about Dropout?
   a. Dropout is a regularization technique
   b. Dropout reduces overfitting
   c. Dropout solves vanishing gradient problem
   d. None of the above

193. Which of the following is TRUE about Dropout?
   a. Dropout randomly switches off some neurons in the network
   b. Dropout solves overfitting and underfitting problem
   c. Dropout solves vanishing and exploding gradient problem
   d. All of the above

194. Which of the following is FALSE about Dropout?
   a. Dropout is a learnable parameter in the network
   b. Dropout increases the accuracy and performance of the model
   c. Dropout introduces sparsity in the network
   d. Dropout makes training process noisy

195. Which of the following is FALSE about Dropout?
   a. Dropout is implemented per layer in a network
   b. Dropout is a hyper-parameter
   c. Dropout can be used in input, hidden and output layers
   d. None of the above

196. Which of the following is TRUE about Dropout?
   a. Dropout can be implemented only in hidden layers
   b. Dropout can be implemented in hidden and input layer
   c. Dropout can be implemented in hidden and output layer
   d. Dropout can be implemented in input, hidden and output layers

197. Which of the following is TRUE about Dropout?
   a. Dropout should be implemented only during training phase, not in testing phase
   b. Dropout should be implemented during training phase as well as during testing phase
   c. Dropout can be compared to boosting technique in machine learning
   d. Dropout is computationally complex as compared to Li and La regularization methods

198. Which of the following is FALSE about Dropout?
   a. Dropout can be compared to bagging technique in machine learning
   b. Dropout is implemented only in hidden layers
   c. Dropout is implemented only during training phase of the model

199. Data Augmentation helps in:
   a. Reducing overfitting
   b. Increasing generalization capacity of the network
   c. Generating data from data

      d. All of the above

200. Which of the following is TRUE about Batch Normalization?
   a. It occurs per epoch
   b. In this process, we normalize the inputs to each hidden layer
   c. It may lead to overfitting
   d. All of the above

201. Which of the following is FALSE about Patch Normalization?
   a. It may lead to overfitting
   b. It solves dying ReLU problem
   c. It introduces regularization in the network
   d. It solves vanishing and exploding gradient issues

202. Which of the following is FALSE about Batch Normalization?
   a. It occurs per batch
   b. It increases overall training period
   c. It makes the gradients stable
   d. None of the above

203. Which of the following is NOT a hyperparameter in a neural network?
   a. Number of hidden layers
   b. Number of neurons in a layer
   c. Weights and bias
   d. Dropout

204. Which of the following is a hyperparameter in a neural network?
   a. Activation Function
   b. Learning Rate
   c. Momentum
   d. All of the above

205. Which of the following is a hyperparameter in a neural network?
   a. Number of epochs
   b. Batch size
   c. Loss Function
   d. All of the above

206. Which of the following is a hyperparameter tuning method in a neural network?
   a. Coordinate Descent
   b. Grid Search and Random Search
   c. Cross-validation
   d. All of the above

207. Which of the following is FALSE about Epochs in a neural network?
   a. Number of epochs is the number of times the whole training data is shown to the network
   b. Number of epochs is a learnable parameter
   c. Time to complete one epoch depends upon the batch size
   d. None of the above

208. Which of the following is FALSE about Batch Size in a neural network?
   a. It is the number of samples passed to the network at one time after which parameter update happens
   b. Batch size is a hyperparameter
   c. Batch size should usually be taken in power of 2
   d. None of the above

209. Using too many epochs while training a network may lead to:
    a. High training time
    b. Overfitting
    c. Unnecessary time wastage
    d. All of the above
210. Using large size batches while training a network may lead to:
    a. Lesser accuracy
    b. Overfitting
    c. High consumption of computational resources
    d. All of the above
211. Model accuracy is not improving after so many epochs. There could be an issue with:
    a. A Network architecture
    b. B. Training data
    c. C. Fine-tuning of hyperparameters
    a. Any of the above
212. Which of the following is FALSE about Autoencoders?
    a. It is an unsupervised deep learning algorithm
    b. It is like a data compression algorithm which performs dimensionality reduction
    c. More the number of code layers, more is the data compression
    d. In autoencoders, output is nearly same as that of the input
213. Which of the following is FALSE about Autoencoders?
    a. It can only compress and decompress the data similar to what it has been trained on
    b. All autoencoders are generative in nature
    c. Difference between the input and output is called reconstruction loss
    d. Output is not exactly same as the input, it is a close but degraded representation
214. Which of the following is FALSE about Autoencoders?
    a. It compresses the input into a latent-space representation and then reconstruct the output from it
    b. Its objective is to minimize this reconstruction loss so that the output is similar to the input
    c. Autoencoders are best suited for image coloring, image captioning and image recognition
    d. Autoencoder possess generalization capabilities
215. Which of the following is FALSE about Encoders in Autoencoders?
    a. It is the layer in which the model learns how to reduce the input dimensions
    b. It compresses the input into a latent space representation
    c. Number of nodes per layer increases with each subsequent layer in the encoder
    d. None of the above
216. Which of the following is FALSE about Decoders in Autoencoders?
    a. It is the layer in which the model learns how to reconstruct the data
    b. The decoded image is a lossy reconstruction of the original image
    c. Number of nodes per layer decreases with each subsequent layer in the decoder
    d. None of the above
217. Which of the following is TRUE about Encoders and Decoders in Autoencoders?
    a. Number of nodes per layer increases with each subsequent layer in the encoder
    b. Number of nodes per layer decreases with each subsequent layer in the encoder
    c. Number of nodes per layer decreases with each subsequent layer in encoder and decoder

      d. Number of nodes per layer increases with each subsequent layer in encoder and decoder

218. Which of the following is FALSE about Bottleneck / Code in Autoencoders?
      a. Number of code layers is n hyper parameter which needs to be fine-tuned
      b. It decides which aspects of the data are relevant and which aspects can be thrown away
      c. It is the layer that contains the compressed representation of the input data
      d. This compressed representation is the lowest possible dimensions of the input data

219. Which of the following is FALSE about code layer in autoencoder?
      a. code size results in more data compres
      b. Small code size results in more generalization
      c. Small code size results in more accuracy
      d. None of the above

220. Which of the following is FALSE about Reconstruction Loss in Autoencoders?
      a. This is the method that measures how well the decoder is performing
      b. This is the method that measures how close the output is to the original input
      c. For an anomaly, reconstruction loss is less as compared to the regular data
      d. None of the above

221. Which of the following is TRUE about Undercomplete autoencoder?
      a. A We limit the number of nodes in the hidden layers to increase its generalization capabilities
      b. B. We use Li and L2 regularization to increase its generalization capabilities
      c. C. We add random noise to the input to increase its generalization capabilities
      a. All of the above

222. Which of the following is TRUE about Sparse autoencoder?
      a. We limit the number of nodes in the hidden layers to increase its generalization capabilities
      b. We use Li and L2 regularization to increase its generalization capabilities
      c. We add random noise to input to increase its generalization capabilities
      d. All of the above

223. Which of the following is TRUE about Denoising autoencoder?
      a. We limit the number of nodes in the hidden layers to increase its generalization capabilities
      b. We use Li and L2 regularization to increase its generalization capabilities
      c. We add random noise to input to increase its generalization capabilities
      d. All of the above

224. Which of the following is NOT a way to increase generalization in autoencoder?
      a. Use larger code size
      b. Use Li and L2 regularization
      c. Add some random noise to the input
      d. Limit the number of nodes in the hidden layers

225. Which of the following is NOT a hyper-parameters in Autoencoders?
      a. Batch size and Code size
      b. Number of code layers
      c. Number of nodes per layer
      d. Number of layers in encoder and decoder

226. Which of the following is NOT a hyper-parameters in Autoencoders?
      a. Number and size of Kernels

     b. Regularization and Dropout

     c. Sparsity parameters in sparse autoencoders

     d. Activation functions such as sigmoid, tanh, softmax, and ReLU

227. Which of the following is FALSE about Code Size hyper-parameter in Autoencoders?

     a. It represents number of nodes in the code layer

     b. Larger size results in more compression

     c. Smaller code size results in more generalization

     d. None of the above

228. Which of the following is FALSE about PCA and Autoencoders?

     a. Both PCA and Autoencoders are dimensionality reduction techniques

     b. PCA works well with non-linear data while Autoencoders are best suited for linear data

     c. Output of both PCA and Autoencoders is lossy

     d. None of the above

229. Which of the following is NOT an application of Autoencoders?

     a. Image coloring

     b. Image captioning

     c. Anomalies and outliers detection

     d. Dimensionality reduction

230. Which of the following is NOT an application of Autoencoders?

     a. B. Predicting next word in a sentence

     b. C. Detecting anomalies in the signal

     a. Lowering the dimensions for better visualization

231. Which of the following is FALSE about Autoencoders and RBM?

     a. Both are unsupervised deep learning algorithms

     b. Both are generative deep learning models

     c. Both help in data compression

     d. None of the above

232. Which of the following is FALSE about Restricted Boltzmann Machines (RBM)?

     a. RBM is a discriminative deep learning algorithm

     b. RBM is probabilistic and unsupervised deep learning algorithm

     c. RBM is a shallow and two-layer (input and hidden) neural network

     d. There is no output layer in RBM unlike Autoencoders

233. Which of the following is FALSE about Restricted Boltzmann Machines (RBM)?

     a. No node is connected to each other across the layers

     b. No two nodes of the same layer are linked in RBM

     c. It finds joint probability distribution that maximizes the log-likelihood function

     d. All of the above

234. Which of the following is TRUE about NLP?

     a. We must take care of Syntax, Semantics and Pragmatics in NLP

     b. Preprocessing tasks include Tokenization, Stemming, Lemmatization and Vectorization

     c. NLP can be used for spam filtering, sentiment analysis and machine translation

     d. All of the above

235. Which of the following is NOT an application of NLP?

     a. Personal assistants (like Alexa, Siri, Google Assistant, Cortana etc.)

     b. Image recognition

    c. Auto-correct grammatical mistakes

    d. Named entity recognition

236.   Which of the following is an application of NLP?

    a. Google Translate

    b. Google Assistant

    c. Chatbots

    d. All of the above

237.   Which of the following is FALSE about Tokenization in NLP?

    a. It is used to remove stopwords from the text

    b. Each word in a text is called a token

    c. We can use regular expressions to find out tokens from the text

    d. None of the above

238.   Which of the following is a correct order of preprocessing of raw data in NLP?

    a. Remove Punctuation, Tokenization, Remove Stopwords, Stemming/Lernmatization, Vectorization

    b. Tokenization, Vectorization, Remove Punctuation, Remove Stopwords, Stemming/Lemmatization

    c. Stemming / Lemmatization, Remove Punctuation, Tokenization, Remove Stopwords, Vectorization

    d. Remove Stopwords, Remove Punctuation, Stemming / Lemmatization, Vectorization, Tokenization

239.   Which of the following is FALSE about preprocessing of raw data in NLP?

    a. We remove stopwords and do stemming to decrease the number of tokens

    b. We should remove all the punctuation marks and stopwords from the text

    c. Lemmatization is a process of removing punctuation and stopwords from the text

    d. Vectorization is used to encode tokens into numbers to create feature vectors

240.   Which of the following is FALSE about Stemming and Lemmatization in NLP?

    a. Lemmatization is more powerful and sophisticated as compared to stemming

    b. Lemmatization is fast and but more complex as compared to the stemming

    c. Both are used to reduce the inflected words to their word stem or root

    d. Both are used to reduce the number of tokens

241.   Which of the following is TRUE about Stemming and Lemmatization in NLP?

    a. Stemming considers context of a word in which it is used in a sentence while Lemmatization does not

    b. Stemming provides more accurate results as compared to Lemmatization

    c. Stemming is faster than Lemmatization

    d. All of the above

242.   Which of the following is TRUE about Vectorization in NLP?

    a. It is used to encode tokens into feature vectors which algorithms can understand

    b. Document term matrix is used to represent the words in the text in the form of matrix of numbers

    c. Count, N-gram and TF-IDF are the types of Vectorization

    d. All of the above

243.   Which of the following frameworks can be used to implement deep learning models?

    a. TensorFlow

    b. Keras

    c. MXNet

d. All of the above

244. Which of the following frameworks can be used to implement deep learning models?
   a. PyTorch
   b. Caffe
   c. Theano
   d. All of the above

245. Which of the following frameworks can be used to implement deep learning models?
   a. DLAJ
   b. CNTK
   c. Chainer
   d. All of the above

246. Which of the following frameworks can be used to implement deep learning models?
   a. Sonnet
   b. Gluon
   c. Swift
   d. All of the above

247. Which of the following frameworks can be used to implement deep learning models?
   a. ONNX
   b. Paddle Paddle
   c. MATLAB
   d. All of the above

249. Which of the following is FALSE about Weights and Bias?
   a. Biases are usually initialized to o (or close to o)
   b. Weights are usually initialized using Xavier technique
   c. Both weight and bias are hyperparameters
   d. None of the above

250. Which of the following is TRUE about Weight Initialization?
   a. If weights are too high, it may lead to vanishing gradient
   b. If weights are too low, it may lead to exploding gradient
   c. Model may never converge due to wrong weight initialization
   d. All of the above

251. Which of the following is FALSE about Weight Initialization?
   a. Training period may increase due to wrong weight initialization
   b. Vanishing and exploding gradient issues may arise due to wrong weight initialization
   c. Initially, we should set all weights to zero while training
   d. None of the above

252. Which of the following is FALSE about Xavier Weight Initialization with ReLU function?
   a. We generate random weights from a Gaussian distribution (mean = 0, std dev = 1)
   b. Multiply above random weights with the square root of (1/n) where n is number of input units
   c. Multiply above random weights with the square root of (2/n) where n is number of input units
   d. None of the above

253. Which of the following is FALSE about Xavier Weight Initialization with Sigmoid function?
   a. We generate random weights from a Gaussian distribution (mean = 0, std dev = 1)

b. Multiply above random weights with the square root of (1/n) where n is number of input units

c. Multiply above random weights with the square root of (2/n) where n is number of input units

d. None of the above