1. …the weights may be reduced to zero.

(A) L1 and L2 (B) L1 (C) L2 (D) None of the above

2. Bagging is an ensemble technique that:

(A) Combines predictions using a weighted average
(B) Trains multiple models on different subsets of the data
(C) Constructs an ensemble by iteratively updating weights
(D) Uses a committee of experts to make predictions

3. Which of the following is/are Limitations of deep learning?

(A) Data labeling (B) Obtain huge training datasets
(C) Both (A) and (B) (D) None of the previous

4. Which neural network has only one hidden layer between the input and output?

(A) Shallow neural network (B) Deep neural network
(C) Feed-forward neural networks (D) Recurrent neural networks

5. CNN is mostly used when there is an?

(A) structured data (B) unstructured data (C) both (A) and (B) (D) None of the previous

6. Which of the following is well suited for perceptual tasks?

(A) feed-forward neural networks
(B) recurrent neural networks
(C) convolutional neural networks
(D) Reinforcement learning

7. Which of the following is/are Common uses of RNNs?

(A) BusinessesHelp securities traders to generate analytic reports
(B) Detect fraudulent credit-card transaction
(C) Provide a caption for images
(D) All of the above

8. Boosting is an ensemble technique that:

(A) Combines predictions using a weighted average
(B) Trains multiple models on different subsets of the data
(C) Constructs an ensemble by iteratively updating weights
(D) Uses a committee of experts to make predictions

9. What steps can we take to prevent overfitting in a Neural Network?

(A) Data Augmentation (B) Weight Sharing
(C) Early Stopping (D) Dropout
(E) All of the previous

10. Which of the following is an example of an ensemble learning algorithm?

(A) Decision tree (B) SVM (C) Random Forest (D) KNN

11. AdaBoost is an example of:

(A) Bagging algorithm (B) Boosting algorithm
(C) Randomized algorithm (D) Reinforcement learning algorithm

12. Gradient Boosting is an ensemble technique that:

- (A) Combines predictions using a weighted average
- (B) Trains multiple models on different subsets of the data
- (C) Constructs an ensemble by iteratively updating weights
- (D) Uses a committee of experts to make predictions

13. XGBoost is a popular implementation of:

- (A) Bagging algorithm
- (B) Boosting algorithm
- (C) Random Forest Algorithm
- (D) K-Means clustering algorithms

14. Stacking is an ensemble technique that:

- (A) Combines predictions using a weighted average
- (B) Trains multiple models on different subsets of the data
- (C) Constructs an ensemble by iteratively updating weights
- (D) Trains a meta-model to make predictions based on outputs of base models

15. Which ensemble learning algorithm uses bootstrapping and feature sampling?

- (A) Random Forest
- (B) AdaBoost
- (C) Gradient Boosting
- (D) Stacking

16. The purpose of using ensemble learning is to:

- (A) Reduce overfitting and improve generalization
- (B) Increase training time and complexity
- (C) Decrease the number of models required
- (D) Eliminate the need for labeled data

17. Bagging algorithms are effective in:

- (A) Handling imbalanced datasets
- (B) sequential data prediction
- (C) Clustering high-dimensional data
- (D) Text classification tasks

18. Which ensemble learning algorithm assigns weights to base models based on their performance?

- (A) AdaBoost
- (B) Random Forest
- (C) Gradient Boosting
- (D) Stacking

19. Which ensemble learning algorithm uses a committee of experts to make predictions?

- (A) Bagging
- (B) Boosting
- (C) Random Forest
- (D) Stacking

20. Which ensemble learning algorithm is prone to overfitting if the base models are too complex?

- (A) Bagging
- (B) Boosting
- (C) Random Forest
- (D) Stacking

21. Which ensemble learning algorithm can handle both regression and classification tasks?

- (A) Bagging
- (B) AdaBoost
- (C) Gradient Boosting
- (D) Stacking
- (E) All of the previous

22. Ensemble learning algorithms are useful when:

- (A) The dataset is small and low-dimensional
- (B) The dataset is large and high-dimensional
- (C) The dataset is perfectly balanced
- (D) The dataset contains categorical variables

23. Ensemble learning algorithms can improve model performance by:

- (A) Reducing bias
- (B) Reducing variance
- (C) Increasing interpretability
- (D) Increasing training time

24. Which ensemble learning algorithm can handle both numerical and categorical data without requiring one-hot encoding?

- (A) Bagging
- (B) AdaBoost
- (C) Graident Boosting
- (D) Stacking

25. Which ensemble learning algorithm is less sensitive to outliers?

- (A) Bagging
- (B) Boosting
- (C) Random Forest
- (D) Stacking

26. The majority voting method in ensemble learning refers to:

- (A) Combining predictions by averaging their probabilities
- (B) Combining predictions by taking the mode of their classes
- (C) Combining predictions by multiplying their probabilities
- (D) Combining predictions by taking the median of their values

27. Which ensemble learning algorithm can handle missing values in the dataset?

- (A) Bagging
- (B) AdaBoost
- (C) Gradient Boosting
- (D) Stacking

28. Ensemble learning algorithms are useful for:

- (A) Improving model stability
- (B) Increasing model complexity
- (C) Reducing feature importance
- (D) Eliminating the need for cross-validation

29. Which ensemble learning algorithm can handle non-linear relationships in the data?

- (A) Bagging
- (B) AdaBoost
- (C) Graident Boosting
- (D) Stacking

30. Ensemble learning algorithms are effective in:

- (A) Reducing model interpretability
- (B) Increasing model training
- (C) Handling unbalanced datasets
- (D) Eliminating the need for hyperparameter tuning

31. Which ensemble learning algorithm can handle both numerical and categorical features effectively?

- (A) Bagging
- (B) AdaBoost
- (C) Gradient Boosting
- (D) Stacking

32. Which ensemble learning algorithm is less susceptible to overfitting compared to others?

- (A) Bagging
- (B) Boosting
- (C) Random Forest
- (D) Stacking

33. Which ensemble learning algorithm uses a weighted sum of predictions from base models?

- (A) Bagging
- (B) AdaBoost
- (C) Gradient boosting
- (D) Stacking

34. Which ensemble learning algorithm can be used to identify important features in a dataset?

- (A) Bagging
- (B) AdaBoost
- (C) Gradient Boosting
- (D) Stacking

35. The ReLu activation has no effect on back-propagation and the vanishing gradient.

- (A) True
- (B) False
- (C) can be true and false
- (D) can't say

36. Why is the vanishing gradient a problem?

- (A) Training is quick if the gradient is large and slow if it's small
- (B) with back propagation, the gradient becomes smaller as it works back through the net
- (C) The gradient is calculated multiplying two numbers between 0 and 1
- (D) All of the previous

37. Which of the following functions can be used as an activation function in the output layer if we wish to predict the probabilities of n classes $(p_1, p_2, ..., p_k)$ such that sum of p over all n equals to 1?

- (A) Softmax
- (B) ReLu
- (C) Sigmoid
- (D) $\tanh$

38. Which of the following would have a constant input in each epoch of training a Deep Learning model?

- (A) Weight between input and hidden layer
- (B) Weight between hidden and output layer
- (C) Biases of all hidden layer neurons
- (D) Activation Function of output layer
- (E) none of the previous

39. Assume a simple MLP model with 3 neurons and inputs = 1,2,3. The weights to the input neurons are 4,5 and 6 respectively. Assume the activation function is a linear constant value of 3. What will be the output ?

(A) 32 (B) 64 (C) 96 (D) 128

40. The input image has been converted into a matrix of size 28 X 28 and a kernel/filter of size 7 X 7 with a stride of 1. What will be the size of the convoluted matrix?

(A) 20×20 (B) 21×21 (C) 22×22 (D) 25×25

41. The number of nodes in the input layer is 10 and the hidden layer is 5. The maximum number of connections from the input layer to the hidden layer are …

(A) 50 (B) less than 50 (C) more than 50 (D) it's an arbitrary value.

42. Which of the following statements is true when you use 1×1 convolutions in a CNN?

(A) It can help in dimensionality reduction
(B) It can be used for feature pooling
(C) It suffers less overfitting due to small kernel size
(D) all of the previous

43. Deep learning algorithms are …more accurate than machine learning algorithm in image classification.

(A) 33 % (B) 37% (C) 40% (D) 41%

44. Which of the following are universal approximators?

(A) Kernel SVM (B) Neural Networks (C) Boosted Decision trees (D) All of the above

45. In which of the following applications can we use deep learning to solve the problem?

(A) Protein structure prediction (B) Prediction of chemical reactions
(C) Detection of exotic particles (D) all of the previous

46. Which of following activation function can't be used at output layer to classify an image ?

(A) Sigmoid (B) tanh (C) ReLU (D) None of the previous

47. Dropout can be applied at visible layer of Neural Network model?

(A) True (B) False

48. Which of the following neural network training challenge can be solved using batch normalization?

(A) overfitting (B) Restrict activation to become too high or low
(C) Training is too slow (D) Both (B) and (C)
(E) All of the previous

49. Changing Sigmoid activation to ReLu will help to get over the vanishing gradient issue?

(A) True (B) False

50. In CNN, having max pooling always decrease the parameters?

(A) True (B) False (C) can be true and false (D) can't say

51. Bagging is more sensitive to noise.

(A) True (B) False

52. What is **true** about the functions of a Multi Layer Perceptron?

(A) The first neural nets that were born out of the need to address teh inaccuracy of an early classifier, the perceptron.
(B) It predicts which group of given set of inputs falls into.
(C) It generates a score that determines the confidence level of the prediction
(D) all of the previous

53. Select reason(s) for using a Deep Neural Network.

- (A) Some patterns are very complex and can't be deciphered precisely by alternate means
- (B) Deep nets are great at recognizing patterns and using them as building blocks in deciphering inputs
- (C) We finally have the technology, GPUs, to accelerate the training process by several folds of magnitude.
- (D) All of the above

54. Sentiment analysis using Deep Learning is a many-to one prediction task

(A) True (B) False (C) Can be true and false (D) can't say

55. BackPropogation cannot be applied when using pooling layers

(A) True (B) False

56. What is the primary purpose of regularization in deep learning?

- (A) to increase computational efficiency
- (B) to reduce the number of layers in a neural network
- (C) to prevent overfitting
- (D) to speed up the training process

57. Which of the following regularization techniques adds a penalty term based on the absolute values of the weights?

(A) L1 regularization (B) L2 regularization (C) Dropout (D) Elastic Net

58. In neural networks, what does L2 regularization encourage?

(A) Sparse weight matrices (B) large weight values
(C) small weight values (D) No impact on weight values

59. How does dropout regularization work in a neural network?

- (A) It randomly drops input features during training
- (B) It randomly drops entire layers during training
- (C) It adds noise to the input data
- (D) It introduces a penalty term for large weights.

60. Which regularization technique combines both L1 and L2 penalties?

(A) Dropout (B) Ride regression
(C) Elastic Net (D) Batch Normalization

61. What is the purpose of early stopping as a form of regularization?

- (A) To stop the training process when the model is underfitting
- (B) To prevent the model from memorizing the training data
- (C) To speed up the convergence of the training process
- (D) To reduce the impact of outliers in the training data

62. Which of the following statements is true about the bias-variance tradeoff in the context of regularization?

- (A) Regularization always increases bias and decreases variance
- (B) Regularization always increases both bias and variance
- (C) Regularization can help balance bias and variance
- (D) Regularization has no impact on the bias-variance tradeoff

63. In the context of neural networks, what does weight decay refer to?

- (A) The gradual increase in weight values during training
- (B) The gradual decrease in weight values during training
- (C) The removal of unnecessary weights from the network
- (D) The introduction of noise to the weight values

64. Which of the following is a disadvantage of using a high regularization strength in a neural network?

(A) Increased risk of overfitting
(B) Faster convergence during training
(C) Enhanced generalization to new data
(D) Reduced capacity to capture complex patterns

65. What is weight decay?

(A) A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration.
(B) Gradual corruption of the weights in the neural network if it's training on noisy data.
(C) The process of gradually decreasing the learning rate during training
(D) A technique to avoid vanishing gradient by imposing a ceiling on the values of the weights.

66. If you have 10,000,000 examples, how would you split the train/dev/test set?

(A) 98% train. 1% dev. 1% test
(B) 33% train. 33% dev. 33% test
(C) 60% train. 20% dev. 20% test

67. The dev and test set should:

(A) Come from the same distribution
(B) Come from different distributions
(C) Be identical to each other(same (x, y) pairs)
(D) Have the same number of examples

68. If your Neural Network model seems to have high variance, what of the following would be promising things to try? (choose all that apply)

(A) Make the Neural network deeper
(B) Get more training data
(C) Get more test data
(D) Increase the number of units in each hidden layer
(E) Add regularization

69. You are working on an automated check-out kiosk for a supermarket, and are building a classifier for apples, bananas and oranges. Suppose your classifier obtains a training set error of 0.5% and a dev set error of 7%. Which of the following are promising things to try to improve your classifier? (Check all that apply)

(A) Increase the regularization parameter lambda
(B) decrease the regularization parameter lambda
(C) get more training data
(D) use a bigger neural network

70. What happens when you increase the regularization hyperparameter lambda?

(A) Weights are pushed twoard becoming smaller (closer to 0)
(B) weights are pushed toward becoming bigger (further from 0)
(C) doubling lambda should roughly result in doubling the weights
(D) Gradient descent taking bigger steps with each iteration (proportional to lambda)

71. With the inverted dropout, at test time:

(A) You don't apply dropout (do not randomly eliminate units), but keep `1/keep_prob` factor in the calculations used in training
(B) You don't apply dropout (do not randomly eliminate units) and do not keep the `1/keep_prob` factor in the calculations usd in the training
(C) You apply dropout (randomly eliminate units) but keep `1/keep_prob` factor in the calculations used in training
(D) You apply dropout (randomly eliminate units) and do not keep `1/keep_prob` factor in the calculations used in training

72. Which of these techniques are useful for reducing variance (reduce overfitting)? (check all that apply)

(A) Dropout
(B) Gradient Checking
(C) Data augmentation
(D) Vanishing gradient
(E) Xavier initialization
(F) L2 regularization
(G) Exploding gradient

73. Why do we normalize the inputs x?

(A) Normalization is another word for regularization–it helps to reduce variance
(B) It makes the cost function faster to optimize
(C) It makes it easier to visualize the data.
(D) It makes the parameter initialization faster.

74. What is the role of the temperature parameter in the context of knowledge distillation as a form of regularization?

(A) Controls the learning rate
(B) Adjusts the level of noise in the input data
(C) Regulates the softness of the target distribution
(D) Sets the threshold for dropout during training

75. In the context of neural networks, what does dropout rate refer to?

(A) The percentage of training samples used during each iteration
(B) The rate at which weight are decayed during training
(C) The probability of dropping out a unit in the hidden layers during training
(D) The learning rate for stochastic gradient descent.

76. Which of the following is a technique used for dynamic adjustment of the learning rate during training to improve convergence in deep learning?

(A) Adversarial training
(B) Learning rate annealing
(C) Batch Normalization
(D) Feature Scaling

77. What is the purpose of adding noise to the input data as a form of regularization?

(A) To make the training process deterministic
(B) To improve model interpretability
(C) To reduce the impact of outliers in the input data
(D) To prevent the model from memorizing the training data

78. In the context of regularization, what does the term "shrinkage" refer to?

(A) Reducing the size of the input data
(B) Reducing the number of hidden layers in the network
(C) Constraining the magnitude of the weights in the model
(D) Eliminating unnecessary features from the dataset

79. Which of the following statements is true about the dropout technique?

(A) Dropout is more effective in shallow networks than deep networks
(B) Dropout can be applied only to input layers
(C) Dropout introduces random variations only during testing
(D) Dropout helps prevent co-adaptation of hidden units

80. What is the primary goal of ensemble methods in machine learning?

(A) To reduce the computational complexity of models
(B) To increase the training time of individual models
(C) To improve the predictive performance of a model by combining multiple models
(D) To decrease the diversity among base models

81. Which of the following statements is true about bagging (Bootstrap Aggregating)?

- (A) It trains multiple models sequentially.
- (B) It trains multiple models independently on different subsets of the training data.
- (C) It combines models using a weighted average.
- (D) It is not suitable for high-variance models.

82. What is the purpose of random forests in ensemble learning?

- (A) To create a forest of decision trees with high correlation
- (B) To reduce the number of trees in the ensemble
- (C) To introduce randomness by considering a random subset of features for each tree
- (D) To eliminate the need for decision trees in the ensemble

83. In boosting, how are the weights assigned to misclassified instances during training?

- (A) Equally to all instances
- (B) Proportional to the difficulty of the instance
- (C) Sequentially, with higher weights for misclassified instances
- (D) Inversely proportional to the number of features

84. Which ensemble method combines the predictions of base models by taking a weighted average, where the weights are learned based on the performance of each model?

(A) Bagging (B) Stacking (C) Boosting (D) Random Forest

85. What is the primary advantage of ensemble methods over individual base models?

- (A) Ensemble methods are always faster than individual models.
- (B) Ensemble methods can handle only linear relationships.
- (C) Ensemble methods often generalize better and have improved robustness.
- (D) Ensemble methods are more prone to overfitting.

86. In the context of boosting, what does the term "weak learner" refer to?

- (A) A model with high training accuracy
- (B) A model that performs slightly better than random chance
- (C) A model with a large number of parameters
- (D) A model that is highly overfit

87. Which ensemble method trains multiple models independently on different subsets of the training data?

(A) Boosting (B) Stacking (C) Bagging (D) Random Forest

88. What is bagging short for in the context of ensemble methods?

(A) Bootstrap Aggregating (B) Boosting Algorithm (C) Bagged Aggregation (D) Batch Aggregation

89. Which ensemble method is known for building a sequence of weak learners, each correcting the errors of its predecessor?

(A) Bagging (B) AdaBoost (C) Random Forest (D) Gradient Boosting

90. What is the primary advantage of ensemble methods over individual base models?

- (A) Faster training time
- (B) Improved generalization and robustness
- (C) Lower computational complexity
- (D) Higher sensitivity to outliers

91. Which ensemble method is based on constructing a forest of decision trees with high diversity?

(A) Bagging (B) AdaBoost (C) Random Forest (D) Stacking

92. What does the acronym "LSTM" stand for in the context of deep learning?

- (A) Long Short-Term Memory
- (B) Linear Short-Term Memory
- (C) Limited Short-Term Memory
- (D) Lasting Short-Term Memory

93. In boosting, what is the purpose of the learning rate parameter?

- (A) It controls the number of weak learners It adjusts the amount by which weights are updated during each iteration
- (B) It determines the depth of decision trees
- (C) It sets the threshold for feature selection

94. What distinguishes Random Forest from traditional bagging techniques?

- (A) Random Forest uses a single decision tree
- (B) Random Forest trains models sequentially
- (C) Random Forest introduces randomness by considering a random subset of features for each tree
- (D) Random Forest assigns equal weights to all instances

95. How does stacking differ from bagging and boosting in ensemble methods?

- (A) Stacking trains models independently on different subsets
- (B) Stacking combines predictions using a weighted average
- (C) Stacking builds a sequence of weak learners
- (D) Stacking uses multiple base models to form a meta-model

96. What role does the concept of "bias-variance tradeoff" play in ensemble methods?

- (A) Ensemble methods eliminate the bias-variance tradeoff
- (B) Ensemble methods intensify the bias-variance tradeoff
- (C) Ensemble methods help balance bias and variance
- (D) Ensemble methods have no impact on bias and variance

97. What is the primary limitation of using too many weak learners in boosting?

- (A) Increased risk of overfitting
- (B) Decreased computational complexity
- (C) Improved generalization
- (D) Faster training time

98. In bagging, how are the subsets of the training data created for each base model?

- (A) Randomly and with replacement
- (B) Randomly and without replacement
- (C) Sequentially and with replacement
- (D) Sequentially and without replacement

99. What is the primary advantage of using gradient boosting over traditional AdaBoost?

- (A) Faster convergence
- (B) Better handling of outliers
- (C) Reduced risk of overfitting
- (D) Simplicity in implementation

100. Which ensemble method is prone to becoming computationally expensive as the number of models increases?

- (A) Bagging
- (B) Stacking
- (C) Boosting
- (D) Random Forest

101. What does the term "stacking" refer to in ensemble learning?

- (A) Combining models using a weighted average
- (B) Training models independently on different subsets
- (C) Constructing a sequence of weak learners
- (D) Using multiple base models to form a meta-model

102. Which ensemble method is known for its ability to handle both linear and non-linear relationships in the data?

Ⓐ Bagging Ⓑ Stacking Ⓒ Random Forest Ⓓ Gradient Boosting

103. Explain the concept of "out-of-bag" error in the context of bagging.

Ⓐ It is the error rate calculated on the training set
Ⓑ It is the error rate on the validation set
Ⓒ It is an estimate of the test error obtained from the unused samples during training
Ⓓ It is a measure of the model's performance on out-of-distribution data

104. What is the role of the hyperparameter "max depth" in decision trees within a Random Forest?

Ⓐ It controls the number of trees in the forest
Ⓑ It limits the maximum depth of individual decision trees
Ⓒ It sets the learning rate for boosting
Ⓓ It adjusts the weights assigned to misclassified instances

105. In the context of ensemble methods, what is "early stopping," and how does it contribute to regularization?

Ⓐ Early stopping involves terminating the training process when the model is underfitting, contributing to model simplicity.
Ⓑ Early stopping prevents overfitting by stopping the training process when the model starts to memorize the training data.
Ⓒ Early stopping introduces noise to the input data during training, preventing overfitting.
Ⓓ Early stopping is not related to regularization in ensemble methods.

106. What is the impact of increasing the number of base models on the computational complexity of stacking?

Ⓐ The computational complexity decreases linearly
Ⓑ The computational complexity increases linearly
Ⓒ The computational complexity remains constant
Ⓓ The computational complexity depends on the type of base models used

107. Explain the concept of "adversarial training" in the context of ensemble methods.

Ⓐ Adversarial training involves training models to be robust against adversarial attacks.
Ⓑ Adversarial training focuses on maximizing the accuracy on the training set.
Ⓒ Adversarial training eliminates the need for ensemble methods.
Ⓓ Adversarial training refers to using adversarial examples as additional training data.

108. How does the concept of "stacking with cross-validation" address the risk of overfitting in stacking?

Ⓐ It eliminates the need for cross-validation in stacking.
Ⓑ It uses multiple cross-validated models, reducing overfitting.
Ⓒ It increases the depth of individual base models.
Ⓓ It has no impact on the risk of overfitting.

109. What is the primary drawback of using a high learning rate in boosting algorithms?

Ⓐ Slower convergence Ⓑ Increased risk of overfitting
Ⓒ Decreased model performance Ⓓ Improved generalization

110. Explain the concept of "feature importance" in the context of Random Forest.

Ⓐ Feature importance represents the number of times a feature is selected by a base model.
Ⓑ Feature importance indicates the relevance of a feature in predicting the target variable.
Ⓒ Feature importance is not applicable to ensemble methods.
Ⓓ Feature importance measures the computational cost of using a specific feature.

111. What is the role of the "n estimators" hyperparameter in ensemble methods such as Random Forest and Gradient Boosting?

Ⓐ It controls the learning rate in boosting algorithms.
Ⓑ It sets the maximum depth of individual decision trees.
Ⓒ It specifies the number of base models in the ensemble.
Ⓓ It determines the subset of features considered for each base model.

112. Explain the concept of "stacking with meta-features" in the context of ensemble methods.

Ⓐ Stacking with meta-features involves using the output of base models as features for a meta-model.
Ⓑ Stacking with meta-features eliminates the need for multiple base models.
Ⓒ Stacking with meta-features refers to combining models using a weighted average.
Ⓓ Stacking with meta-features involves using only one type of base model in the ensemble.

113. What is Dropout in the context of neural networks?

Ⓐ Adding noise to input features
Ⓑ Removing random neurons during training
Ⓒ Reducing the learning rate
Ⓓ Increasing the number of hidden layers

114. What is the main purpose of Dropout in neural networks?

Ⓐ To increase overfitting
Ⓑ To speed up the training process
Ⓒ To prevent co-adaptation of neurons
Ⓓ To eliminate the need for activation functions

115. Which of the following statements is true about the application of Dropout during training?

Ⓐ Dropout is only applied to input layers
Ⓑ Dropout is applied to all layers except the output layer
Ⓒ Dropout is applied during both training and testing
Ⓓ Dropout is never applied to neural networks

116. How does Dropout contribute to regularization in neural networks?

Ⓐ By increasing the number of parameters
Ⓑ By introducing noise to the input data
Ⓒ By reducing the model's capacity
Ⓓ By promoting co-adaptation of neurons

117. In terms of training, what does it mean if a neuron is "dropped out"?

Ⓐ The neuron's weights are set to zero
Ⓑ The neuron is removed from the network temporarily
Ⓒ The neuron's activation function is bypassed
Ⓓ The neuron's output is squared

118. What challenge does Dropout aim to address in neural networks?

Ⓐ Underfitting Ⓑ Overfitting Ⓒ Vanishing gradients Ⓓ Exploding gradients

119. How does Dropout affect the training time of a neural network?

Ⓐ Slows down the training process
Ⓑ Speeds up the training process
Ⓒ No impact on training time
Ⓓ Depends on the type of activation function used

120. What is the recommended range for Dropout rates in neural networks?

Ⓐ 0.0 to 0.1 Ⓑ 0.2 to 0.5 Ⓒ 0.5 to 0.8 Ⓓ 0.9 to 1.0

121. How does Dropout contribute to model generalization?

Ⓐ By memorizing the training data
Ⓑ By promoting co-adaptation of neurons
Ⓒ By reducing the sensitivity of neurons to specific input features
Ⓓ By increasing the number of hidden layers

122. When applying Dropout, which phase is used for adjusting the weights of the neural network?

Ⓐ Training phase
Ⓑ Testing phase
Ⓒ Both training and testing phases
Ⓓ Neither training nor testing phases

123. Explain the term "co-adaptation of neurons" in the context of neural networks and how Dropout addresses it.

Ⓐ Co-adaptation refers to neurons relying too much on each other, and Dropout breaks these dependencies by randomly dropping neurons during training.
Ⓑ Co-adaptation is a form of regularization, and Dropout exacerbates co-adaptation by introducing noise.
Ⓒ Co-adaptation occurs when neurons are independent, and Dropout enforces co-adaptation by removing dependencies.
Ⓓ Co-adaptation is unrelated to Dropout; Dropout only affects the learning rate.

124. How does the effectiveness of Dropout vary with the size and complexity of a neural network?

Ⓐ Dropout is more effective in small and simple networks
Ⓑ Dropout is more effective in large and complex networks
Ⓒ Dropout is equally effective across all network sizes and complexities
Ⓓ Dropout is irrelevant to network size and complexity

125. What is the relationship between Dropout and the concept of ensemble learning?

Ⓐ Dropout is a type of ensemble learning
Ⓑ Ensemble learning and Dropout are unrelated concepts
Ⓒ Dropout and ensemble learning achieve the same result in terms of model diversity
Ⓓ Dropout eliminates the need for ensemble learning

126. Explain the trade-off between using a high Dropout rate and a low Dropout rate in neural networks.

Ⓐ High Dropout rates lead to overfitting, while low Dropout rates may result in underfitting.
Ⓑ High Dropout rates always improve model generalization, while low Dropout rates reduce model capacity.
Ⓒ There is no trade-off; the Dropout rate does not impact model performance.
Ⓓ The trade-off depends on the type of activation function used in the network.

127. How does Dropout contribute to mitigating the vanishing gradient problem in deep neural networks?

Ⓐ a. By increasing the learning rate
Ⓑ By preventing co-adaptation of neurons
Ⓒ By introducing noise to the input data
Ⓓ By reducing the sensitivity of neurons to specific input features

128. What is the primary goal of data augmentation in machine learning?

Ⓐ To decrease the size of the dataset
Ⓑ To increase the computational complexity
Ⓒ To improve model performance by increasing the diversity of the training data
Ⓓ To eliminate the need for validation data

129. Which of the following is a common technique used in data augmentation for image data?

Ⓐ Principal Component Analysis (PCA)
Ⓑ Feature scaling
Ⓒ Image rotation
Ⓓ Lasso regularization

130. How does data augmentation contribute to preventing overfitting in machine learning models?

Ⓐ By reducing the size of the training dataset
Ⓑ By increasing the number of layers in the model
Ⓒ By introducing noise to the input data
Ⓓ By providing a more diverse set of training examples

131. In text data augmentation, what technique involves replacing words with their synonyms?

Ⓐ Tokenization
Ⓑ Embedding
Ⓒ Word substitution
Ⓓ Lemmatization

132. Which of the following is a disadvantage of data augmentation?

Ⓐ Increased model generalization
Ⓑ Potential introduction of unrealistic patterns
Ⓒ Improved model robustness
Ⓓ Decreased computational efficiency

133. What is the purpose of random cropping in image data augmentation?

Ⓐ To decrease the image resolution
Ⓑ To remove irrelevant features from the image
Ⓒ To create variations in the spatial location of objects
Ⓓ To increase the image contrast

134. Which type of data augmentation is commonly used for time series data?

Ⓐ Image rotation
Ⓑ Time warping
Ⓒ Word substitution
Ⓓ Feature scaling

135. Explain the concept of "jittering" in the context of data augmentation.

Ⓐ Jittering refers to the introduction of noise to input features
Ⓑ Jittering involves the random selection of a subset of data points
Ⓒ Jittering is a synonym for image rotation
Ⓓ Jittering is irrelevant to data augmentation

136. In the context of image data augmentation, what is the purpose of horizontal flipping?

Ⓐ To rotate images clockwise
Ⓑ To create mirror images
Ⓒ To adjust the image brightness
Ⓓ To resize images

137. How does data augmentation differ from feature engineering?

Ⓐ Data augmentation focuses on creating new samples, while feature engineering manipulates existing features.
Ⓑ Feature engineering is limited to image data, while data augmentation is applicable to all data types.
Ⓒ Data augmentation involves scaling features, while feature engineering involves randomization.
Ⓓ Feature engineering and data augmentation are synonymous terms.

138. What is the role of dropout in the context of data augmentation?

Ⓐ Dropout is not related to data augmentation
Ⓑ Dropout enhances data augmentation by randomly removing features during training
Ⓒ Dropout is a type of data augmentation technique
Ⓓ Dropout prevents data augmentation from introducing unrealistic patterns

139. Which data augmentation technique is commonly used for audio data to introduce variations in pitch?

(A) Time warping
(B) Spectrogram augmentation
(C) Random cropping
(D) Jittering

140. What is the purpose of elastic deformation in image data augmentation?

(A) To adjust the image contrast
(B) To introduce non-linear distortions to the image
(C) To resize the image
(D) To rotate the image

141. In natural language processing, which technique involves randomly removing words from sentences during data augmentation?

(A) Tokenization
(B) Word substitution
(C) Sentence splitting
(D) Sentence dropout

142. Explain the concept of "adversarial training" in the context of data augmentation and how it addresses robustness.

(A) Adversarial training focuses on creating adversarial examples to test the model's robustness against unseen patterns introduced by data augmentation.
(B) Adversarial training is irrelevant to data augmentation.
(C) Adversarial training involves increasing the size of the training set.
(D) Adversarial training enhances data augmentation by introducing adversarial noise during the augmentation process.

143. How does data augmentation contribute to handling class imbalance in classification tasks?

(A) Data augmentation exacerbates class imbalance
(B) Data augmentation is not related to class imbalance
(C) Data augmentation generates additional samples for minority classes, addressing class imbalance
(D) Data augmentation reduces the need for addressing class imbalance

144. What challenges might arise when applying data augmentation to non-image data types, such as tabular data?

(A) Difficulty in implementing data augmentation for non-image data
(B) Limited applicability of data augmentation to non-image data
(C) The potential introduction of unrealistic patterns
(D) No challenges; data augmentation is equally effective for all data types

145. Explain the term "mixup" in the context of data augmentation and how it differs from traditional augmentation techniques.

(A) Mixup involves blending two or more samples, creating new synthetic samples with averaged labels.
(B) Mixup is a synonym for image rotation.
(C) Mixup refers to the addition of random noise to input features.
(D) Mixup is irrelevant to data augmentation.

146. How does data augmentation impact the interpretability of machine learning models?

(A) Data augmentation improves model interpretability by providing more diverse training examples.
(B) Data augmentation has no impact on model interpretability.
(C) Data augmentation reduces model interpretability due to the introduction of synthetic samples.
(D) Data augmentation improves model interpretability by eliminating the need for validation data.

147. What is the role of "cutout" in image data augmentation?

(A) To remove random portions from images
(B) To blur the edges of images
(C) To rotate images
(D) To resize images

148. In the context of data augmentation, explain how the technique of "shearing" is applied to image data.

(A) Shearing involves adjusting the brightness of images.
(B) Shearing is irrelevant to data augmentation.
(C) Shearing introduces non-linear distortions to the image by tilting it along one of its axes.
(D) Shearing is a synonym for image rotation.

149. Which ensemble learning algorithm can be applied to both regression and classification tasks?

(A) Bagging (B) AdaBoost (C) Random Forest (D) Stacking

150. Ensemble learning algorithms can be computationally expensive when:

(A) The dataset is small (B) The base models are simple
(C) The ensemble size is small (D) The dataset is large

151. Which ensemble learning algorithm can be used to identify important features in a dataset?

(A) Bagging (B) AdaBoost (C) Gradient Boosting (D) Stacking

152. The difference between deep learning and machine learning algorithms is that there is no need of feature engineering in machine learning algorithms, whereas, it is recommended to do feature engineering first and then apply deep learning.

(A) True (B) False

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

(A) Neural Network (B) Random Forest (C) k-Nearest neighbor (D) None of the above

154. Which of the following option is correct for the below-mentioned techniques?

1. AdaGrad uses first order differentiation
2. L-BFGS uses second order differentiation
3. AdaGrad uses second order differentiation
4. L-BFGS uses first order differentiation

(A) 1 and 2 (B) 3 and 4 (C) 1 and 4 (D) 2 and 3

155. Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional neural network.

(A) True (B) False

156. Suppose we have a deep neural network model which was trained on a vehicle detection problem. The dataset consisted of images on cars and trucks and the aim was to detect name of the vehicle (the number of classes of vehicles are 10). Now you want to use this model on different dataset which has images of only Ford Mustangs (aka car) and the task is to locate the car in an image.
Which of the following categories would be suitable for this type of problem?

(A) Fine tune only the last couple of layers and change the last layer (classification layer) to regression layer
(B) Freeze all the layers except the last, re-train the last layer
(C) Re-train the model for the new dataset
(D) None of these

157. Suppose you have 5 convolutional kernel of size 7 x 7 with zero padding and stride 1 in the first layer of a convolutional neural network. You pass an input of dimension 224 x 224 x 3 through this layer. What are the dimensions of the data which the next layer will receive?

(A) 217x217x3 (B) 217x217x8 (C) 218x218x5 (D) 220x220x7

158. Suppose you replace the ReLU activation function with linear activation in a neural network that was originally able to approximate an XNOR function with ReLU activations. Will the new neural network be able to approximate an XNOR function?

(A) Yes (B) No

159. If a 5-layer neural network takes 3 hours to train on a GPU with 4GB VRAM and at test time, it takes 2 seconds for a single data point, what would be the testing time for the new architecture if dropout is added after the 2nd and 4th layers with rates 0.2 and 0.3, respectively?

(A) Less than 2 secs (B) Exactly 2 secs (C) Greater than 2 secs (D) Can not Say

160. Which of the following options can be used to reduce overfitting in deep learning models?

(A) Add more data
(B) Use data augmentation
(C) Use architecture that generalizes well
(D) Add regularization
(E) Reduce architectural complexity
(F) All of these

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

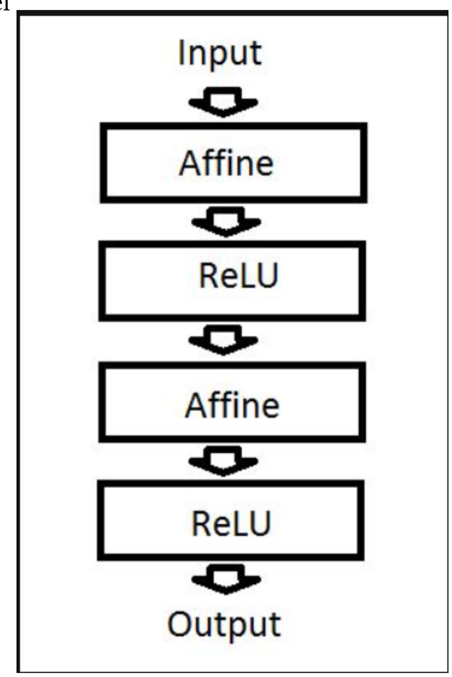
(A) Higher the perplexity the better
(B) Lower the perplexity the better

162. Suppose an input to Max-Pooling layer is given above. The pooling size of neurons in the layer is (3, 3).

3	4	5
4	5	6
5	6	7

(A) 3
(B) 5
(C) 5.5
(D) 7

163. If we remove the ReLU layers, we can still use this neural network to model non-linear functions.



(A) True
(B) False

164. 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

165. Scenario 1: You are given data of the map of Arcadia city, with aerial photographs of the city and its outskirts. The task is to segment the areas into industrial land, farmland, and natural landmarks like rivers, mountains, etc.

Scenario 2: You are given data of the map of Arcadia city, with detailed roads and distances between landmarks. This is represented as a graph structure. The task is to find out the nearest distance between two landmarks.

Can deep learning be applied to Scenario 1 but not Scenario 2?

(A) TRUE
(B) FALSE

166. Which of the following is a data augmentation technique used in image recognition tasks?

(A) Horizontal flipping (B) Random cropping (C) Random scaling (D) Color jittering
(E) Random translation (F) Random shearing (G) All of these

167. Given an n-character word, we want to predict which character would be the n+1th character in the sequence. For example, our input is "predictio" (which is a 9-character word) and we have to predict what would be the 10th character.

Which neural network architecture would be suitable to complete this task?

(A) Fully-Connected Neural Network
(B) Convolutional Neural Network
(C) Recurrent Neural Network
(D) Restricted Boltzmann Machine

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

(A) Convolutional network on input and deconvolutional network on output
(B) Deconvolutional network on input and convolutional network on output

169. A ReLU unit in neural network never gets saturated.

(A) True
(B) False

170. What is the relationship between dropout rate and regularization?

(A) Higher the dropout rate, higher is the regularization
(B) Higher the dropout rate, lower is the regularization

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

(A) Unlike backprop, in BPTT we sum up gradients for corresponding weight for each time step
(B) Unlike backprop, in BPTT we subtract gradients for corresponding weight for each time step

172. Exploding gradient problem is an issue in training deep networks where the gradient gets so large that the loss goes to an infinitely high value and then explodes. What is the probable approach when dealing with the "Exploding Gradient" problem in RNNs?

(A) Use modified architectures like LSTM and GRUs
(B) Gradient clipping
(C) Dropout
(D) None of these

173. There are many types of gradient descent algorithms. Two of the most notable ones are l-BFGS and SGD. l-BFGS is a second-order gradient descent technique whereas SGD is a first-order gradient descent technique.
In which of the following scenarios would you prefer l-BFGS over SGD?

(A) Data is sparse
(B) Number of parameters of neural network are small
(C) Both of them
(D) None of these

174. Which of the following is not a direct prediction technique for NLP tasks?

(A) Recurrent Neural Network
(B) Skip-gram model
(C) PCA
(D) Convolutional Neural Network

175. Which of the following would be the best for a non-continuous objective during optimization in deep neural net?

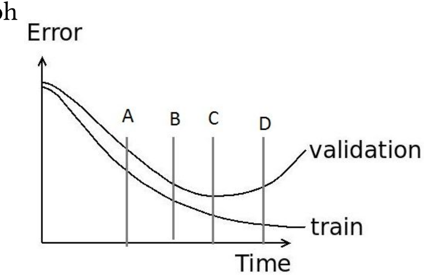
(A) L-BFGS
(B) SGD
(C) AdaGrad
(D) Subgradient method

176. Which of the following is correct?

(A) Dropout randomly masks the input weights to a neuron
(B) Dropconnect randomly masks both input and output weights to a neuron
(C) 1 is False and 2 is True
(D) Both 1 and 2 are True

177. While training a neural network for image recognition task, we plot the graph of training error and validation error for debugging.

(A) A
(B) B
(C) C
(D) D



178. Backpropagation works by first calculating the gradient of . . . and then propagating it backward.

(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

179. Mini-Batch sizes when defining a neural network are preferred to be multiples of 2's such as 256 or 512. What is the reason behind it?

(A) Gradient descent optimizes best when you use an even number
(B) Parallelization of the neural network is best when the memory is used optimally
(C) Losses are erratic when you don't use an even number
(D) None of these

180. As the length of a sentence increases, it becomes harder for a neural translation machine to perform as sentence meaning is represented by a fixed dimensional vector. To solve this, which of the following could we do?

(A) Use recursive units instead of recurrent
(B) Use attention mechanism
(C) Use character-level translation
(D) None of these

181. A recurrent neural network can be unfolded into a fully connected neural network with infinite length.

(A) TRUE
(B) FALSE

182. Which of the following is a bottleneck for deep learning algorithms?

(A) Data related to the problem
(B) CPU to GPU communication
(C) GPU memory
(D) All of the above

183. When deriving a memory cell in memory networks, we choose to read values as vector values instead of scalars. Which type of addressing would this entail?

(A) Content-based addressing
(B) Location-based addressing

184. It is generally recommended to replace pooling layers in the generator part of convolutional generative adversarial nets with …?

(A) Affine layer
(B) Strided convolutional layer
(C) Fractional strided convolutional layer
(D) ReLU layer

185. Which of the following statements is true with respect to GRU?

(A) Units with short-term dependencies have a very active reset gate.
(B) Units with long-term dependencies have a very active update gate.
(C) None of them
(D) Both 1 and 2

186. If the calculation of the reset gate in a GRU unit is close to 0, which of the following would occur?

(A) Previous hidden state would be ignored
(B) Previous hidden state would not be ignored

187. If the calculation of the update gate in a GRU unit is close to 1, which of the following would occur?

(A) Forgets the information for future time steps
(B) Copies the information through many time steps

188. Dropout technique is not an advantageous technique for which of the following layers?

(A) Affine layer
(B) Convolutional layer
(C) RNN layer
(D) None of these

189. Suppose your task is to predict the next few notes of a song when you are given the preceding segment of the song. Which architecture of a neural network would be better suited to solve the problem?

(A) End-to-End fully connected neural network
(B) CNN followed by recurrent units
(C) Neural Turing Machine
(D) None of these

190. What is the primary purpose of a Convolutional Neural Network (CNN)?

(A) Object detection
(B) Image classification
(C) Text generation
(D) Reinforcement learning

191. Which layer type is typically used to extract local features in a CNN?

(A) Convolutional layer
(B) Pooling layer
(C) Fully connected layer
(D) Activation layer

192. What is the advantage of using convolutional layers in a CNN?

- (A) They can capture local spatial patterns in the input data
- (B) They can handle sequential data
- (C) They can generate synthetic data
- (D) They can capture local spatial patterns in the input data

193. What is the purpose of the pooling layer in a CNN?

- (A) To introduce non-linearity to the network
- (B) To reduce the spatial dimensions of the feature maps
- (C) To adjust the weights and biases of the network
- (D) To compute the gradients for backpropagation

194. Which activation function is commonly used in the convolutional layers of a CNN?

- (A) Sigmoid
- (B) ReLU (Rectified Linear Unit)
- (C) Tanh (Hyperbolic Tangent)
- (D) Softmax

195. What is the purpose of the stride parameter in a convolutional layer?

- (A) To determine the size of the receptive field
- (B) To control the step size of the convolution operation
- (C) To adjust the learning rate during training
- (D) None of the above

196. Which layer type is used to reduce the spatial dimensions in a CNN?

- (A) Convolutional layer
- (B) Pooling layer
- (C) Fully connected layer
- (D) Activation layer

197. What is the purpose of the padding parameter in a convolutional layer?

- (A) To adjust the learning rate during training
- (B) To prevent the reduction of spatial dimensions
- (C) To regularize the network and prevent overfitting
- (D) None of the above

198. Which layer type is responsible for making final predictions in a CNN?

- (A) Convolutional layer
- (B) Pooling layer
- (C) Fully connected layer
- (D) Activation layer

199. What is the purpose of the fully connected layers in a CNN?

- (A) To capture global patterns and make predictions
- (B) To reduce the spatial dimensions of the input data
- (C) To apply non-linear transformations to the feature maps
- (D) To initialize the weights and biases of the network

200. Which layer type is responsible for applying non-linear transformations to the feature maps in a CNN?

- (A) Convolutional layer
- (B) Pooling layer
- (C) Fully connected layer
- (D) Activation layer

201. What is the purpose of dropout regularization in a CNN?

- (A) To randomly disable neurons during training to prevent overfitting
- (B) To adjust the learning rate during training
- (C) To increase the number of layers in the network
- (D) None of the above

202. Which layer type is responsible for backpropagating the gradients and updating the network's parameters in a CNN?

- (A) Convolutional layer
- (B) Pooling layer
- (C) Fully connected layer
- (D) Activation layer

203. What is the primary advantage of using a CNN over a fully connected neural network for image processing tasks?

- (A) CNNs have a higher training speed
- (B) CNNs can handle sequential data
- (C) CNNs have a higher number of neurons
- (D) CNNs can capture local spatial patterns in the input data

204. Which layer type is responsible for parameter sharing in a CNN?

- (A) Convolutional layer
- (B) Pooling layer
- (C) Fully connected layer
- (D) Activation layer

205. What is the purpose of the receptive field in a convolutional layer?

- (A) To determine the number of filters in the layer
- (B) To determine the size of the feature maps
- (C) To specify the size of the local region for the convolution operation
- (D) None of the above

206. Which layer type is responsible for spatial downsampling in a CNN?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) Activation layer

207. What is the purpose of the filter/kernel in a convolutional layer?

(A) To determine the number of neurons in the layer
(B) To specify the size of the feature maps
(C) To extract local features from the input data
(D) None of the above

208. Which layer type is commonly used in CNNs to normalize the input data?

(A) Convolutional layer (B) Pooling layer (C) Batch normalization layer (D) Activation layer

209. What is the primary goal of training a CNN?

(A) To minimize the prediction error on the training data
(B) To maximize the number of layers in the network
(C) To achieve 100
(D) None of the above

210. Which layer type is responsible for introducing translation invariance in a CNN?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) Activation layer

211. What is the purpose of the output layer in a CNN?

(A) To compute the predicted output based on the final feature representation
(B) To reduce the spatial dimensions of the input data
(C) To apply non-linear transformations to the feature maps
(D) To initialize the weights and biases of the network

212. What is the purpose of zero-padding in a CNN?

(A) To adjust the learning rate during training
(B) To prevent the reduction of spatial dimensions
(C) To regularize the network and prevent overfitting
(D) None of the above

213. Which layer type is commonly used in CNNs for semantic segmentation tasks?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) Upsampling layer

214. What is the purpose of the loss function in CNN training?

(A) To measure the prediction error and guide the learning process
(B) To initialize the weights and biases of the network
(C) To adjust the learning rate during training
(D) None of the above

215. Which layer type is commonly used in CNNs to introduce non-linearity?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) Activation layer

216. What is the purpose of the learning rate in CNN training?

(A) To control the step size of the parameter updates during optimization
(B) To adjust the size of the filters in the convolutional layers
(C) To increase the number of layers in the network
(D) None of the above

217. Which layer type is responsible for feature extraction in a CNN?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) Activation layer

218. What is the purpose of data augmentation in CNN training?

(A) To increase the number of layers in the network
(B) To introduce noise and variations in the training data
(C) To adjust the learning rate during training
(D) None of the above

219. Which layer type is commonly used in CNNs to handle variable-sized inputs?

(A) Convolutional layer (B) Pooling layer (C) Fully connected layer (D) None of the above

220. What is the primary purpose of a Recurrent Neural Network (RNN)?

(A) Image classification (B) Text generation (C) Reinforcement learning (D) Object detection

221. Which layer type is typically used to capture sequential dependencies in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) Activation layer

222. What is the advantage of using recurrent layers in an RNN?

(A) They can handle non-linear transformations
(B) They can handle variable-length inputs
(C) They can generate synthetic data
(D) They can capture temporal dependencies in the input data

223. What is the purpose of the hidden state in an RNN?

(A) To store the information from the previous time step
(B) To adjust the learning rate during training
(C) To compute the gradients for backpropagation
(D) None of the above

224. Which activation function is commonly used in the recurrent layers of an RNN?

(A) ReLU (Rectified Linear Unit)
(B) Sigmoid
(C) Tanh (Hyperbolic Tangent)
(D) Softmax

225. What is the purpose of the time step parameter in an RNN?

(A) To determine the number of recurrent layers in the network
(B) To adjust the learning rate during training
(C) To specify the length of the input sequence
(D) None of the above

226. Which layer type is commonly used to initialize the hidden state in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) Activation layer

227. What is the purpose of the bidirectional RNN architecture?

(A) To handle sequential data in both forward and backward directions
(B) To reduce the computational complexity of the network
(C) To adjust the learning rate during training
(D) None of the above

228. Which layer type is responsible for making final predictions in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) Activation layer

229. What is the purpose of the recurrent connection in an RNN?

(A) To propagate the hidden state across different time steps
(B) To adjust the weights and biases of the network
(C) To reduce the dimensionality of the input data
(D) None of the above

230. Which layer type is commonly used in RNNs for sequence-to-sequence tasks?

(A) Input layer (B) Hidden layer (C) Output layer (D) Attention layer

231. What is the purpose of the backpropagation through time (BPTT) algorithm in RNN training?

(A) To compute the gradients and update the network's parameters
(B) To adjust the learning rate during training
(C) To prevent overfitting by regularizing the model
(D) None of the above

232. Which layer type is commonly used in RNNs to handle variable-length inputs?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

233. What is the purpose of the initial hidden state in an RNN?

(A) To provide the starting point for the recurrent computation
(B) To adjust the learning rate during training
(C) To compute the gradients for backpropagation
(D) None of the above

234. Which layer type is responsible for handling the output at each time step in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) Activation layer

235. What is the purpose of the teacher forcing technique in RNN training?

(A) To adjust the learning rate during training
(B) To propagate the gradients through time
(C) To reduce the computational complexity of the network
(D) None of the above

236. Which layer type is commonly used in RNNs for language modeling tasks?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

237. What is the purpose of the sequence-to-vector architecture in an RNN?

(A) To process an input sequence and produce a fixed-length representation
(B) To adjust the weights and biases of the network
(C) To reduce the dimensionality of the input data
(D) None of the above

238. Which layer type is responsible for introducing non-linearity in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) Activation layer

239. What is the purpose of the forget gate in a Gated Recurrent Unit (GRU)?

(A) To control the flow of information from the previous hidden state (B) To adjust the learning rate during training
(C) To compute the gradients for backpropagation (D) None of the above

240. Which layer type is commonly used in RNNs for machine translation tasks?

(A) Input layer (B) Hidden layer (C) Output layer (D) Attention layer

241. What is the purpose of the peephole connections in a Long Short-Term Memory (LSTM) network?

(A) To allow the cell state to influence the gating mechanisms (B) To adjust the learning rate during training
(C) To introduce non-linearity to the network (D) None of the above

242. Which layer type is responsible for handling variable-length outputs in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

243. What is the purpose of the cell state in an LSTM network?

(A) To store long-term dependencies in the input sequence (B) To adjust the learning rate during training
(C) To compute the gradients for backpropagation (D) None of the above

244. Which layer type is commonly used in RNNs for speech recognition tasks?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

245. What is the purpose of the input gate in an LSTM network?

(A) To control the flow of information from the current input (B) To adjust the learning rate during training
(C) To introduce non-linearity to the network (D) None of the above

246. Which layer type is responsible for handling variable-length inputs and outputs in an RNN?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

247. What is the purpose of the output gate in an LSTM network?

(A) To control the flow of information to the current output (B) To adjust the learning rate during training
(C) To introduce non-linearity to the network (D) None of the above

248. Which layer type is commonly used in RNNs for time series prediction tasks?

(A) Input layer (B) Hidden layer (C) Output layer (D) None of the above

249. What is the purpose of the reset gate in a Gated Recurrent Unit (GRU)?

(A) To reset the hidden state based on the current input (B) To adjust the learning rate during training
(C) To introduce non-linearity to the network (D) None of the above

Solutions to the Exercises

1.(**B**) L1
2.(**B**) Trains multiple models on different subsets of the data
3.(**C**) Both A and B
4.(**A**) Shallow neural network
5.(**B**) unstructured data
6.(**C**) convolutional neural networks
7.(**D**) All of the above
8.(**C**) Constructs an ensemble by iteratively updating weights
9.(**E**) All of the previous
10.(**C**) Random Forest
11.(**B**) Boosting algorithm
12.(**C**) Constructs an ensemble by iteratively updating weights
13.(**B**) Boosting algorithm
14.(**D**) Trains a meta-model to make predictions based on outputs of base models
15.(**A**) Random Forest
16.(**A**) Reduce overfitting and improve generalization
17.(**A**) Handling imbalanced datasets
18.(**A**) AdaBoost
19.(**D**) Stacking
20.(**B**) Boosting
21.(**E**) All of the previous
22.(**B**) The dataset is large and high-dimensional
23.(**B**) Reducing variance
24.(**D**) Stacking
25.(**A**) Bagging
26.(**B**) Combining predictions by taking the mode of their classes
27.(**C**) Gradient Boosting
28.(**A**) Improving model stability
29.(**D**) Stacking
30.(**C**) Handling unbalanced datasets
31.(**D**) Stacking
32.(**C**) Random Forest
33.(**B**) AdaBoost
34.(**C**) Gradient Boosting
35.(**B**) False
36.(**D**) All of the previous
37.(**A**) Softmax
38.(**A**) Weight between input and hidden layer
39.(**C**) 96
40.(**C**) 22×22
41.(**A**) 50
42.(**D**) all of the previous
43.(**D**) 41%
44.(**D**) All of the above
45.(**D**) all of the previous
46.(**C**) ReLU
47.(**A**) True
48.(**E**) All of the previous
49.(**A**) True
50.(**B**) False
51.(**B**) False
52.(**D**) all of the previous
53.(**D**) All of the above
54.(**A**) True
55.(**B**) False
56.(**C**) to prevent overfitting
57.(**A**) L1 regularization
58.(**C**) small weight values
59.(**B**) It randomly drops entire layers during training
60.(**C**) Elastic Net
61.(**B**) To prevent the model from memorizing the training data
62.(**C**) Regularization can help balance bias and variance
63.(**B**) The gradual decrease in weight values during training
64.(**D**) Reduced capacity to capture complex patterns
65.(**A**) A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration.
66.(**A**) 98% train. 1% dev. 1% test
67.(**A**) Come from the same distribution
68.(**B**) Get more training data
(**E**) Add regularization
69.(**A**) Increase the regularization parameter lambda
(**C**) get more training data
70.(**A**) Weights are pushed twoard becoming smaller (closer to 0)
71.(**B**) You don't apply dropout (do not randomly eliminate units) and do not keep the `1/keep_prob` factor in the calculations usd in the training
72.(**A**) Dropout
(**C**) Data augmentation
(**F**) L2 regularization
73.(**B**) It makes the cost function faster to optimize
74.(**C**) Regulates the softness of the target distribution
75.(**C**) The probability of dropping out a unit in the hidden layers during training
76.(**B**) Learning rate annealing
77.(**D**) To prevent the model from memorizing the training data
78.(**C**) Constraining the magnitude of the weights in the model
79.(**D**) Dropout helps prevent co-adaptation of hidden units
80.(**C**) To improve the predictive performance of a model by combining multiple models
81.(**B**) It trains multiple models independently on different subsets of the training data.
82.(**C**) To introduce randomness by considering a random subset of features for each tree
83.(**C**) Sequentially, with higher weights for misclassified instances
84.(**B**) Stacking
85.(**C**) Ensemble methods often generalize better and have improved robustness.
86.(**B**) A model that performs slightly better than random chance
87.(**C**) Bagging
88.(**A**) Bootstrap Aggregating
89.(**B**) AdaBoost
90.(**B**) Improved generalization and robustness
91.(**C**) Random Forest
92.(**A**) Long Short-Term Memory
93.(**A**) It adjusts the amount by which weights are updated during each iteration
94.(**C**) Random Forest introduces randomness by considering a random subset of features for each tree
95.(**D**) Stacking uses multiple base models to form a meta-model
96.(**C**) Ensemble methods help balance bias and variance
97.(**A**) Increased risk of overfitting
98.(**A**) Randomly and with replacement
99.(**B**) Better handling of outliers
100.(**C**) Boosting
101.(**D**) Using multiple base models to form a meta-model

102.(**C**) Random Forest
103.(**C**) It is an estimate of the test error obtained from the unused samples during training
104.(**B**) It limits the maximum depth of individual decision trees
105.(**B**) Early stopping prevents overfitting by stopping the training process when the model starts to memorize the training data.
106.(**B**) The computational complexity increases linearly
107.(**A**) Adversarial training involves training models to be robust against adversarial attacks.
108.(**B**) It uses multiple cross-validated models, reducing overfitting.
109.(**B**) Increased risk of overfitting
110.(**B**) Feature importance indicates the relevance of a feature in predicting the target variable.
111.(**C**) It specifies the number of base models in the ensemble.
112.(**A**) Stacking with meta-features involves using the output of base models as features for a meta-model.
113.(**B**) Removing random neurons during training
114.(**C**) To prevent co-adaptation of neurons
115.(**B**) Dropout is applied to all layers except the output layer
116.(**C**) By reducing the model's capacity
117.(**B**) The neuron is removed from the network temporarily
118.(**B**) Overfitting
119.(**A**) Slows down the training process
120.(**B**) 0.2 to 0.5
121.(**C**) By reducing the sensitivity of neurons to specific input features
122.(**A**) Training phase
123.(**A**) Co-adaptation refers to neurons relying too much on each other, and Dropout breaks these dependencies by randomly dropping neurons during training.
124.(**B**) Dropout is more effective in large and complex networks
125.(**C**) Dropout and ensemble learning achieve the same result in terms of model diversity
126.(**A**) High Dropout rates lead to overfitting, while low Dropout rates may result in underfitting.
127.(**C**) By introducing noise to the input data
128.(**C**) To improve model performance by increasing the diversity of the training data
129.(**C**) Image rotation
130.(**D**) By providing a more diverse set of training examples
131.(**C**) Word substitution
132.(**B**) Potential introduction of unrealistic patterns
133.(**C**) To create variations in the spatial location of objects
134.(**B**) Time warping
135.(**A**) Jittering refers to the introduction of noise to input features
136.(**B**) To create mirror images
137.(**A**) Data augmentation focuses on creating new samples, while feature engineering manipulates existing features.
138.(**B**) Dropout enhances data augmentation by randomly removing features during training
139.(**B**) Spectrogram augmentation
140.(**B**) To introduce non-linear distortions to the image
141.(**D**) Sentence dropout
142.(**A**) Adversarial training focuses on creating adversarial examples to test the model's robustness against unseen patterns introduced by data augmentation.
143.(**C**) Data augmentation generates additional samples for minority classes, addressing class imbalance
144.(**C**) The potential introduction of unrealistic patterns
145.(**A**) Mixup involves blending two or more samples, creating new synthetic samples with averaged labels.
146.(**C**) Data augmentation reduces model interpretability due to the introduction of synthetic samples.
147.(**A**) To remove random portions from images
148.(**C**) Shearing introduces non-linear distortions to the image by tilting it along one of its axes.
149.(**C**) Random Forest
150.(**D**) The dataset is large
151.(**C**) Gradient Boosting
152.(**B**) False
153.(**A**) Neural Network
154.(**A**) 1 and 2
155.(**B**) False
156.(**A**) Fine tune only the last couple of layers and change the last layer (classification layer) to regression layer
157.(**C**) 218x218x5
158.(**B**) No
159.(**B**) Exactly 2 secs
160.(**F**) All of these
161.(**B**) Lower the perplexity the better
162.(**D**) 7
163.(**B**) False
164.(**D**) All of the above
165.(**B**) FALSE
166.(**G**) All of these
167.(**C**) Recurrent Neural Network
168.(**A**) Convolutional network on input and deconvolutional network on output
169.(**B**) False
170.(**B**) Higher the dropout rate, lower is the regularization
171.(**A**) Unlike backprop, in BPTT we sum up gradients for corresponding weight for each time step
172.(**B**) Gradient clipping
173.(**C**) Both of them
174.(**C**) PCA
175.(**D**) Subgradient method
176.(**C**) 1 is False and 2 is True In dropout, neurons are dropped, whereas in dropconnect, connections are dropped. So, both input and output weights will be rendered useless in dropconnect, while only one of them should be dropped in dropconnect.
177.(**C**) C In dropout, neurons are dropped, whereas in dropconnect, connections are dropped. So, both input and output weights will be rendered useless in dropconnect, while only one of them should be dropped in dropconnect.
178.(**C**) Sum of squared error with respect to outputs
179.(**B**) Parallelization of the neural network is best when the memory is used optimally
180.(**B**) Use attention mechanism
181.(**A**) TRUE
182.(**D**) All of the above
183.(**A**) Content-based addressing
184.(**C**) Fractional strided convolutional layer
185.(**D**) Both 1 and 2
186.(**A**) Previous hidden state would be ignored
187.(**B**) Copies the information through many time steps
188.(**C**) RNN layer
189.(**B**) CNN followed by recurrent units
190.(**B**) Image classification

191.(**A**) Convolutional layer
192.(**A**) They can capture local spatial patterns in the input data
193.(**B**) To reduce the spatial dimensions of the feature maps
194.(**B**) ReLU (Rectified Linear Unit)
195.(**B**) To control the step size of the convolution operation
196.(**B**) Pooling layer
197.(**B**) To prevent the reduction of spatial dimensions
198.(**C**) Fully connected layer
199.(**A**) To capture global patterns and make predictions
200.(**D**) Activation layer
201.(**A**) To randomly disable neurons during training to prevent overfitting
202.(**C**) Fully connected layer
203.(**D**) CNNs can capture local spatial patterns in the input data
204.(**A**) Convolutional layer
205.(**C**) To specify the size of the local region for the convolution operation
206.(**B**) Pooling layer
207.(**C**) To extract local features from the input data
208.(**C**) Batch normalization layer
209.(**A**) To minimize the prediction error on the training data
210.(**A**) Convolutional layer
211.(**A**) To compute the predicted output based on the final feature representation
212.(**B**) To prevent the reduction of spatial dimensions
213.(**D**) Upsampling layer
214.(**A**) To measure the prediction error and guide the learning process
215.(**D**) Activation layer
216.(**A**) To control the step size of the parameter updates during optimization
217.(**A**) Convolutional layer
218.(**B**) To introduce noise and variations in the training data
219.(**D**) None of the above
220.(**B**) Text generation
221.(**B**) Hidden layer
222.(**D**) They can capture temporal dependencies in the input data
223.(**A**) To store the information from the previous time step
224.(**C**) Tanh (Hyperbolic Tangent)
225.(**C**) To specify the length of the input sequence
226.(**B**) Hidden layer
227.(**A**) To handle sequential data in both forward and backward directions
228.(**C**) Output layer
229.(**A**) To propagate the hidden state across different time steps
230.(**D**) Attention layer
231.(**A**) To compute the gradients and update the network's parameters
232.(**A**) Input layer
233.(**A**) To provide the starting point for the recurrent computation
234.(**C**) Output layer
235.(**B**) To propagate the gradients through time
236.(**C**) Output layer
237.(**A**) To process an input sequence and produce a fixed-length representation
238.(**D**) Activation layer
239.(**A**) To control the flow of information from the previous hidden state
240.(**D**) Attention layer
241.(**A**) To allow the cell state to influence the gating mechanisms
242.(**C**) Output layer
243.(**A**) To store long-term dependencies in the input sequence
244.(**C**) Output layer
245.(**A**) To control the flow of information from the current input
246.(**D**) None of the above
247.(**A**) To control the flow of information to the current output
248.(**C**) Output layer
249.(**A**) To reset the hidden state based on the current input