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. 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

4. 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

5. 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

6. 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.

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

(a) Dropout (b) Ride regression
(c) Elastic Net (d) Batch Normalization

8. 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

9. 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

10. 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

11. 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

12. 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.

13. 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

14. 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

15. 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

16. 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

17. 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)

18. 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

19. 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

20. 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.

21. 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

22. 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.

23. 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

24. 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

25. 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

26. 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

27. 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

28. 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.

29. 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

30. 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

31. 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

32. 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.

33. 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

34. 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

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

(a) Boosting (b) Stacking (c) Bagging (d) Random Forest

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

(a) Bootstrap Aggregating (b) Boosting Algorithm (c) Bagged Aggregation (d) Batch Aggregation

37. 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) Randomly assigned

38. Which ensemble method combines the predictions of base models by taking a weighted average?

- (a) Bagging
- (b) Stacking
- (c) Boosting
- (d) Random Forest

39. 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

40. 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

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

(a) Bagging (b) AdaBoost (c) Random Forest (d) Stacking

42. 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

43. 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

44. 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

45. 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

46. 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

47. 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

48. 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

49. 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

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

- (a) Bagging
- (b) Stacking
- (c) Boosting
- (d) Random Forest

51. 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

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

(a) Bagging (b) Stacking (c) Random Forest (d) Gradient Boosting

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

(a) It is the error rate calculated on the training set
(b) It is the error rate on the validation set
(c) It is an estimate of the test error obtained from the unused samples during training
(d) It is a measure of the model's performance on out-of-distribution data

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

(a) It controls the number of trees in the forest
(b) It limits the maximum depth of individual decision trees
(c) It sets the learning rate for boosting
(d) It adjusts the weights assigned to misclassified instances

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

(a) Early stopping involves terminating the training process when the model is underfitting, contributing to model simplicity.
(b) Early stopping prevents overfitting by stopping the training process when the model starts to memorize the training data.
(c) Early stopping introduces noise to the input data during training, preventing overfitting.
(d) Early stopping is not related to regularization in ensemble methods.

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

(a) The computational complexity decreases linearly
(b) The computational complexity increases linearly
(c) The computational complexity remains constant
(d) The computational complexity depends on the type of base models used

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

(a) Adversarial training involves training models to be robust against adversarial attacks.
(b) Adversarial training focuses on maximizing the accuracy on the training set.
(c) Adversarial training eliminates the need for ensemble methods.
(d) Adversarial training refers to using adversarial examples as additional training data.

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

(a) It eliminates the need for cross-validation in stacking.
(b) It uses multiple cross-validated models, reducing overfitting.
(c) It increases the depth of individual base models.
(d) It has no impact on the risk of overfitting.

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

- (a) Slower convergence
- (b) Increased risk of overfitting
- (c) Decreased model performance
- (d) Improved generalization

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

- (a) Feature importance represents the number of times a feature is selected by a base model.
- (b) Feature importance indicates the relevance of a feature in predicting the target variable.
- (c) Feature importance is not applicable to ensemble methods.
- (d) Feature importance measures the computational cost of using a specific feature.

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

- (a) It controls the learning rate in boosting algorithms.
- (b) It sets the maximum depth of individual decision trees.
- (c) It specifies the number of base models in the ensemble.
- (d) It determines the subset of features considered for each base model.

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

- (a) Stacking with meta-features involves using the output of base models as features for a meta-model.
- (b) Stacking with meta-features eliminates the need for multiple base models.
- (c) Stacking with meta-features refers to combining models using a weighted average.
- (d) Stacking with meta-features involves using only one type of base model in the ensemble.

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

- (a) Adding noise to input features
- (b) Removing random neurons during training
- (c) Reducing the learning rate
- (d) Increasing the number of hidden layers

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

- (a) To increase overfitting
- (b) To speed up the training process
- (c) To prevent co-adaptation of neurons
- (d) To eliminate the need for activation functions

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

- (a) Dropout is only applied to input layers
- (b) Dropout is applied to all layers except the output layer
- (c) Dropout is applied during both training and testing
- (d) Dropout is never applied to neural networks

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

- (a) By increasing the number of parameters
- (b) By introducing noise to the input data
- (c) By reducing the model's capacity
- (d) By promoting co-adaptation of neurons

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

- (a) The neuron's weights are set to zero
- (b) The neuron is removed from the network temporarily
- (c) The neuron's activation function is bypassed
- (d) The neuron's output is squared

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

(a) Underfitting (b) Overfitting (c) Vanishing gradients (d) Exploding gradients

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

- (a) Slows down the training process
- (b) Speeds up the training process
- (c) No impact on training time
- (d) Depends on the type of activation function used

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

(a) 0.0 to 0.1 (b) 0.2 to 0.5 (c) 0.5 to 0.8 (d) 0.9 to 1.0

71. How does Dropout contribute to model generalization?

- (a) By memorizing the training data
- (b) By promoting co-adaptation of neurons
- (c) By reducing the sensitivity of neurons to specific input features
- (d) By increasing the number of hidden layers

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

- (a) Training phase
- (b) Testing phase
- (c) Both training and testing phases
- (d) Neither training nor testing phases

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

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

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

- (a) Dropout is more effective in small and simple networks
- (b) Dropout is more effective in large and complex networks
- (c) Dropout is equally effective across all network sizes and complexities
- (d) Dropout is irrelevant to network size and complexity

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

- (a) Dropout is a type of ensemble learning
- (b) Ensemble learning and Dropout are unrelated concepts
- (c) Dropout and ensemble learning achieve the same result in terms of model diversity
- (d) Dropout eliminates the need for ensemble learning

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

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

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

- (a) a. By increasing the learning rate
- (b) By preventing co-adaptation of neurons
- (c) By introducing noise to the input data
- (d) By reducing the sensitivity of neurons to specific input features

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

- (a) To decrease the size of the dataset
- (b) To increase the computational complexity
- (c) To improve model performance by increasing the diversity of the training data
- (d) To eliminate the need for validation data

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

- (a) Principal Component Analysis (PCA)
- (b) Feature scaling
- (c) Image rotation
- (d) Lasso regularization

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

- (a) By reducing the size of the training dataset
- (b) By increasing the number of layers in the model
- (c) By introducing noise to the input data
- (d) By providing a more diverse set of training examples

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

- (a) Tokenization
- (b) Embedding
- (c) Word substitution
- (d) Lemmatization

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

- (a) Increased model generalization
- (b) Potential introduction of unrealistic patterns
- (c) Improved model robustness
- (d) Decreased computational efficiency

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

- (a) To decrease the image resolution
- (b) To remove irrelevant features from the image
- (c) To create variations in the spatial location of objects
- (d) To increase the image contrast

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

- (a) Image rotation
- (b) Time warping
- (c) Word substitution
- (d) Feature scaling

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

- (a) Jittering refers to the introduction of noise to input features
- (b) Jittering involves the random selection of a subset of data points
- (c) Jittering is a synonym for image rotation
- (d) Jittering is irrelevant to data augmentation

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

- (a) To rotate images clockwise
- (b) To create mirror images
- (c) To adjust the image brightness
- (d) To resize images

87. How does data augmentation differ from feature engineering?

- (a) Data augmentation focuses on creating new samples, while feature engineering manipulates existing features.
- (b) Feature engineering is limited to image data, while data augmentation is applicable to all data types.
- (c) Data augmentation involves scaling features, while feature engineering involves randomization.
- (d) Feature engineering and data augmentation are synonymous terms.

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

- (a) Dropout is not related to data augmentation
- (b) Dropout enhances data augmentation by randomly removing features during training
- (c) Dropout is a type of data augmentation technique
- (d) Dropout prevents data augmentation from introducing unrealistic patterns

89. 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

90. 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

91. 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

92. 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.

93. 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

94. 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

95. 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.

96. 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.

97. 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

98. 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.

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

(a) Bagging (b) AdaBoost (c) Random Forest (d) Stacking

100. 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

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

(a) Bagging (b) AdaBoost (c) Gradient Boosting (d) Stacking

Solutions to the Exercises

1.(**b**) L1
2.(**b**) Trains multiple models on different subsets of the data
3.(**c**) to prevent overfitting
4.(**a**) L1 regularization
5.(**c**) small weight values
6.(**b**) It randomly drops entire layers during training
7.(**c**) Elastic Net
8.(**b**) To prevent the model from memorizing the training data
9.(**c**) Regularization can help balance bias and variance
10.(**b**) The gradual decrease in weight values during training
11.(**d**) Reduced capacity to capture complex patterns
12.(**a**) A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration.
13.(**a**) 98% train. 1% dev. 1% test
14.(**a**) Come from the same distribution
15.(**b**) Get more training data
(**e**) Add regularization
16.(**a**) Increase the regularization parameter lambda
(**c**) get more training data
17.(**a**) Weights are pushed twoard becoming smaller (closer to 0)
18.(**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
19.(**a**) Dropout
(**c**) Data augmentation
(**f**) L2 regularization
20.(**b**) It makes the cost function faster to optimize
21.(**c**) Regulates the softness of the target distribution
22.(**c**) The probability of dropping out a unit in the hidden layers during training
23.(**b**) Learning rate annealing
24.(**c**) To reduce the impact of outliers in the input data
25.(**c**) Constraining the magnitude of the weights in the model
26.(**d**) Dropout helps prevent co-adaptation of hidden units
27.(**c**) To improve the predictive performance of a model by combining multiple models
28.(**b**) It trains multiple models independently on different subsets of the training data.
29.(**c**) To introduce randomness by considering a random subset of features for each tree
30.(**c**) Sequentially, with higher weights for misclassified instances
31.(**b**) Stacking
32.(**c**) Ensemble methods often generalize better and have improved robustness.
33.(**b**) A model that performs slightly better than random chance
34.(**b**) AdaBoost
35.(**c**) Bagging
36.(**a**) Bootstrap Aggregating
37.(**c**) Sequentially, with higher weights for misclassified instances
38.(**b**) Stacking
39.(**b**) AdaBoost
40.(**b**) Improved generalization and robustness
41.(**c**) Random Forest
42.(**a**) Long Short-Term Memory
43.(**a**) It adjusts the amount by which weights are updated during each iteration
44.(**c**) Random Forest introduces randomness by considering a random subset of features for each tree
45.(**d**) Stacking uses multiple base models to form a meta-model
46.(**c**) Ensemble methods help balance bias and variance
47.(**a**) Increased risk of overfitting
48.(**a**) Randomly and with replacement
49.(**b**) Better handling of outliers
50.(**c**) Boosting
51.(**d**) Using multiple base models to form a meta-model
52.(**c**) Random Forest
53.(**c**) It is an estimate of the test error obtained from the unused samples during training
54.(**b**) It limits the maximum depth of individual decision trees
55.(**b**) Early stopping prevents overfitting by stopping the training process when the model starts to memorize the training data.
56.(**b**) The computational complexity increases linearly
57.(**a**) Adversarial training involves training models to be robust against adversarial attacks.
58.(**b**) It uses multiple cross-validated models, reducing overfitting.
59.(**b**) Increased risk of overfitting
60.(**b**) Feature importance indicates the relevance of a feature in predicting the target variable.
61.(**c**) It specifies the number of base models in the ensemble.
62.(**a**) Stacking with meta-features involves using the output of base models as features for a meta-model.
63.(**b**) Removing random neurons during training
64.(**c**) To prevent co-adaptation of neurons
65.(**b**) Dropout is applied to all layers except the output layer
66.(**c**) By reducing the model's capacity
67.(**b**) The neuron is removed from the network temporarily
68.(**b**) Overfitting
69.(**a**) Slows down the training process
70.(**b**) 0.2 to 0.5
71.(**c**) By reducing the sensitivity of neurons to specific input features
72.(**a**) Training phase

73.(**a**) Co-adaptation refers to neurons relying too much on each other, and Dropout breaks these dependencies by randomly dropping neurons during training.
74.(**b**) Dropout is more effective in large and complex networks
75.(**c**) Dropout and ensemble learning achieve the same result in terms of model diversity
76.(**a**) High Dropout rates lead to overfitting, while low Dropout rates may result in underfitting.
77.(**c**) By introducing noise to the input data
78.(**c**) To improve model performance by increasing the diversity of the training data
79.(**c**) Image rotation
80.(**d**) By providing a more diverse set of training examples
81.(**c**) Word substitution
82.(**b**) Potential introduction of unrealistic patterns
83.(**c**) To create variations in the spatial location of objects
84.(**b**) Time warping
85.(**a**) Jittering refers to the introduction of noise to input features
86.(**b**) To create mirror images
87.(**a**) Data augmentation focuses on creating new samples, while feature engineering manipulates existing features.
88.(**b**) Dropout enhances data augmentation by randomly removing features during training
89.(**b**) Spectrogram augmentation
90.(**b**) To introduce non-linear distortions to the image
91.(**d**) Sentence dropout
92.(**a**) Adversarial training focuses on creating adversarial examples to test the model's robustness against unseen patterns introduced by data augmentation.
93.(**c**) Data augmentation generates additional samples for minority classes, addressing class imbalance
94.(**c**) The potential introduction of unrealistic patterns
95.(**a**) Mixup involves blending two or more samples, creating new synthetic samples with averaged labels.
96.(**c**) Data augmentation reduces model interpretability due to the introduction of synthetic samples.
97.(**a**) To remove random portions from images
98.(**c**) Shearing introduces non-linear distortions to the image by tilting it along one of its axes.
99.(**c**) Random Forest
100.(**d**) The dataset is large