**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) $20 \times 20$ (B) $21 \times 21$ (C) $22 \times 22$ (D) $25 \times 25$

**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 $1 \times 1$ 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?

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

**103**. 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

**104**. 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

**105**. 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.

**106**. 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

**107**. 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.

**108**. 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.

**109**. 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

**110**. 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.

**111**. 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.

**112**. 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.

**113**. 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

**114**. 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

**115**. 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

**116**. 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

**117**. 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

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

(A) Underfitting (B) Overfitting (C) Vanishing gradients (D) Exploding gradients

**119**. 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

**120**. 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

**121**. 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

**122**. 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

**123**. 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.

**124**. 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

**125**. 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

**126**. 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.

**127**. 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

**128**. 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

**129**. 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

**130**. 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

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

- (A) Tokenization
- (B) Embedding
- (C) Word substitution
- (D) Lemmatization

**132**. 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

**133**. 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

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

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

**135**. 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

**136**. 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

**137**. 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.

**138**. 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

**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

## 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**.(**A**) Bagging
**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**) $22 \times 22$
**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