20 Exam Questions Part B

1. Which of the following is a classification algorithm?

(a) Linear regression
(b) Logistic regression
(c) Ridge regression
(d) Lasso regression

Answer: (b) Logistic regression

Explanation: Logistic regression is a classification algorithm that models the probability of a binary or categorical outcome.

2. Which of the following metrics is most suitable for evaluating a binary classification model?

(a) RMSE
(b) R^2
(c) ROC AUC
(d) Log loss

Answer: (c) ROC AUC

Explanation: ROC AUC is a widely used metric for evaluating binary classification models as it measures the ability of the model to distinguish between positive and negative samples across all possible threshold values.

3. Which of the following statements is true about the bias-variance tradeoff?

(a) Increasing model complexity always leads to overfitting.
(b) Increasing model complexity always leads to underfitting.
(c) Increasing model complexity can lead to both overfitting and underfitting.
(d) Model complexity has no effect on overfitting or underfitting.

Answer: (c) Increasing model complexity can lead to both overfitting and underfitting.

Explanation: The bias-variance tradeoff is the balance between underfitting (high bias) and overfitting (high variance). Increasing model complexity can lead to overfitting if the model becomes too complex and starts to fit to noise in the data, or underfitting if the model is too simple and doesn't capture the true underlying relationship.

4. Which of the following is a common regularization technique used in logistic regression?

a) Ridge regression

b) Lasso regression

c) Elastic net regression

d) None of the above

Answer: a) Ridge regression

Explanation: Ridge regression is a common regularization technique used in logistic regression to prevent overfitting. It adds a penalty term to the cost function that shrinks the coefficients towards zero, making the model less complex and more generalizable.

5. Which of the following statements is true about regularization?

(a) Regularization increases model complexity.
(b) Regularization decreases model complexity.
(c) Regularization has no effect on model complexity.
(d) Regularization only applies to linear regression models.

Answer: (b) Regularization decreases model complexity.

Explanation: Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function that discourages large parameter values. This effectively shrinks the parameters towards zero and reduces model complexity.

6. Which of the following metrics is most suitable for evaluating a regression model?
(a) RMSE
(b) ROC AUC
(c) Log loss
(d) F1 score

Answer: (a) RMSE

Explanation: RMSE (Root Mean Squared Error) is a commonly used metric for evaluating regression models as it measures the average distance between the predicted and actual values. It is sensitive to outliers and gives more weight to larger errors, making it suitable for continuous target variables.

7. Which of the following statements is true regarding the bias-variance tradeoff in machine learning?

a) High bias models tend to overfit the data
b) High variance models tend to underfit the data
c) Increasing model complexity tends to increase bias and decrease variance
d) Decreasing model complexity tends to increase bias and decrease variance

Answer: d) Decreasing model complexity tends to increase bias and decrease variance

Explanation: The bias-variance tradeoff is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. Bias refers to the error caused by approximating a real-life problem with a simplified model. High bias models are those that are too simple, leading to underfitting the data. Variance refers to the error caused by the model's sensitivity to small fluctuations in the training set. High variance models are those that are too complex, leading to overfitting the data. Therefore, decreasing model complexity tends to increase bias and decrease variance, leading to better generalization.

8. What is the Bayes error in classification?

a) The minimum achievable error rate on a given classification problem
b) The error rate of the optimal classifier that has access to the true conditional probability distribution of the data
c) The error rate of the optimal classifier that has access to the training data only
d) The error rate of the suboptimal classifier that is used in practice

Answer: b) The error rate of the optimal classifier that has access to the true conditional probability distribution of the data

Explanation: The Bayes error is the minimum achievable error rate on a given classification problem, assuming that we have access to the true conditional probability distribution of the data. It represents the best possible performance that any classifier can achieve on the given problem. In practice, we do not know the true distribution of the data, so we have to estimate it from the training data. As a result, the error rate of any classifier that we use in practice is always higher than the Bayes error.

9. In linear regression, what is the objective function that is optimized by the least squares method?

a) Mean absolute error
b) Mean squared error
c) Root mean squared error
d) Maximum likelihood

Answer: b) Mean squared error

Explanation: In linear regression, the least squares method is a common approach for estimating the parameters of a linear model that best fit the training data. The method minimizes the sum of squared residuals between the predicted values and the actual values in the training set. The objective function that is optimized by the least squares method is the mean squared error (MSE), which is the average of the squared residuals. The method is equivalent to maximizing the likelihood of the data under a Gaussian noise assumption.

10. Which of the following statements is true regarding the bias term in linear regression?

a) The bias term controls the slope of the linear regression line
b) The bias term controls the intercept of the linear regression line
c) The bias term is always positive
d) The bias term is always negative

Answer: b) The bias term controls the intercept of the linear regression line

Explanation: In linear regression, the bias term (also known as the intercept or the offset) is the coefficient that is multiplied by the constant 1 to account for the y-axis intercept of the regression line. It represents the value of the dependent variable when all independent variables are zero. The bias term controls the position of the regression line along the y-axis, whereas the other coefficients control its slope.

11. Which of the following statements is true regarding ridge regression?

A) Ridge regression does not reduce the magnitude of coefficients as much as LASSO regression.
B) Ridge regression can lead to the problem of overfitting in high-dimensional datasets.
C) Ridge regression is preferred over LASSO regression when we have a large number of features and we suspect that only a small subset of features are relevant.
D) Ridge regression is not suitable when the number of features is larger than the number of samples.

Answer: A) Ridge regression does not reduce the magnitude of coefficients as much as LASSO regression.

Explanation: Ridge regression is a form of regularization that adds a penalty term to the cost function of linear regression. This penalty term is the L2 norm of the coefficient vector multiplied by a hyperparameter lambda. Ridge regression shrinks the magnitude of the coefficients towards zero, but not to exactly zero. LASSO regression, on the other hand, uses the L1 norm of the coefficient vector as the penalty term and can drive some of the coefficients to exactly zero, effectively performing feature selection. Therefore, statement A is true.

Statement B is incorrect as ridge regression is actually used to prevent overfitting in high-dimensional datasets.

Statement C is also incorrect as LASSO regression is preferred over ridge regression when we have a large number of features and we suspect that only a small subset of features are relevant.

Statement D is also incorrect as ridge regression can handle high-dimensional datasets even when the number of features is larger than the number of samples.

12. Which of the following statements is true about logistic regression?

A) It can only handle binary classification problems.
B) It assumes that the relationship between the dependent variable and independent variables is linear.
C) It outputs a probability score for each possible class label.
D) It is not affected by multicollinearity among the independent variables.

Answer: C) It outputs a probability score for each possible class label.

Explanation: Logistic regression is a type of regression analysis used for predicting the probability of a binary outcome (i.e., two classes). It models the probability of the dependent variable (i.e., the binary outcome) using a logistic function, which outputs a probability score between 0 and 1. Therefore, statement C is true.

Statement A is incorrect as logistic regression can handle both binary and multiclass classification problems.

Statement B is incorrect as logistic regression does not assume a linear relationship between the dependent variable and independent variables. Instead, it assumes a logit function of the independent variables.

Statement D is incorrect as logistic regression assumes that there is no perfect multicollinearity among the independent variables. Perfect multicollinearity occurs when two or more independent variables are perfectly correlated, making it impossible to estimate the regression coefficients.

13. Which of the following statements is true about L1 regularization?

A) It can drive some of the coefficients to exactly zero, effectively performing feature selection.
B) It shrinks the magnitude of the coefficients towards zero, but not to exactly zero.
C) It is also known as ridge regularization.
D) It is more suitable for handling high-dimensional datasets with a large number of features.

Answer: A) It can drive some of the coefficients to exactly zero, effectively performing feature selection.

Explanation: L1 regularization, also known as LASSO regularization, uses the L1 norm of the coefficient vector as the penalty term in the cost function of a regression model. This penalty term can drive some of the coefficients to exactly zero, effectively performing feature selection. Therefore, statement A is true.

Statement B is incorrect as L2 regularization (ridge regularization) shrinks the magnitude of the coefficients towards zero, but not to exactly zero.

Statement C is incorrect as ridge regularization uses the L2 norm of the coefficient vector as the penalty term.

Statement D is incorrect as both L1 and L2 regularization can handle high-dimensional datasets with a large number of features.

14. Which of the following is a benefit of standardizing or normalizing the input data?

A) It increases the interpretability of the coefficients.
B) It reduces the computational complexity of the model.
C) It prevents overfitting in high-dimensional datasets.
D) It ensures that all variables have equal influence on the model.

Answer: D) It ensures that all variables have equal influence on the model.

Explanation: Standardizing or normalizing the input data can help ensure that all variables have equal influence on the model. This is important because variables with larger scales or variances can dominate the model and bias the coefficients towards those variables. Standardization or normalization scales the variables to have zero mean and unit variance (in the case of standardization) or to fall within a specified range (in the case of normalization), which can prevent bias in the coefficients. Therefore, statement D is true.

Statement A is incorrect as standardization or normalization does not affect the interpretability of the coefficients.

Statement B is incorrect as standardization or normalization does not affect the computational complexity of the model.

Statement C is incorrect as standardization or normalization does not directly prevent overfitting in high-dimensional datasets, but it can help prevent bias in the coefficients and improve model performance.

15. Which of the following statements about regularization in machine learning is true?

a) Regularization increases the variance of the model
b) Regularization increases the bias of the model
c) Regularization can help prevent overfitting
d) Regularization has no effect on the model's performance

Answer: c) Regularization can help prevent overfitting. Regularization is a technique used to reduce the complexity of a model and prevent overfitting. It adds a penalty term to the loss function, which encourages the model to have smaller weights and reduces the likelihood of overfitting.

Explanation: Regularization helps in reducing overfitting by discouraging the model from learning complex relationships between features and target variable, which may not generalize well to new data. By adding a penalty term to the loss function, the model is encouraged to learn simpler relationships between features and target variable. This reduces the likelihood of overfitting and improves the model's generalization performance.

16. Which of the following techniques can be used to standardize or normalize data in machine learning?

a) Min-max scaling
b) Z-score normalization
c) Log transformation
d) All of the above

Answer: d) All of the above. There are multiple techniques that can be used to standardize or normalize data in machine learning, including min-max scaling, z-score normalization, and log transformation.

Explanation: Standardization or normalization of data is an important preprocessing step in machine learning. It involves transforming the data to have a standard scale and/or distribution. This can help improve the performance of many machine learning algorithms. Min-max scaling scales the data to a fixed range (usually 0 to 1), z-score normalization scales the data to have a mean of 0 and standard deviation of 1, and log transformation can be used to transform skewed data to a more normal distribution.

17. Which of the following statements is true regarding the Gini index and the entropy criterion for building decision trees?

a) The Gini index tends to create more balanced trees than the entropy criterion.

b) The entropy criterion tends to create more balanced trees than the Gini index.

c) The Gini index and the entropy criterion always result in the same tree.

d) The Gini index and the entropy criterion are not commonly used for building decision trees.

Answer: a) The Gini index tends to create more balanced trees than the entropy criterion.

Explanation: The Gini index measures the probability of misclassifying a random sample from a given class, while the entropy criterion measures the degree of disorder in a set of data. In general, the Gini index tends to create more balanced trees (i.e., trees with more equally-sized branches) than the entropy criterion.

18. Which of the following statements is true regarding the support vector machine (SVM)?

a) The SVM algorithm always finds the global minimum for the optimization problem.

b) The SVM algorithm is sensitive to the choice of kernel function.

c) The SVM algorithm does not work well with non-linearly separable data.

d) The SVM algorithm does not require any hyperparameters to be tuned.

Answer: b) The SVM algorithm is sensitive to the choice of kernel function.

Explanation: The SVM algorithm works by finding a hyperplane that separates the data into two classes with the maximum possible margin. This can be done using different kernel functions (such as linear, polynomial, or radial basis function kernels). The choice of kernel function can have a significant impact on the performance of the SVM algorithm.

19. Which of the following is an advantage of resampling methods (such as bootstrapping and cross-validation) over a single train/test split for evaluating a supervised learning model?

a) Resampling methods are faster and less computationally expensive.

b) Resampling methods provide a more accurate estimate of model performance on unseen data.

c) Resampling methods are more robust to overfitting.

d) Resampling methods are only useful for large datasets.

Answer: b) Resampling methods provide a more accurate estimate of model performance on unseen data.

Explanation: Resampling methods such as bootstrapping and cross-validation involve repeatedly splitting the data into training and testing sets, and then averaging the performance across multiple iterations. This provides a more accurate estimate of model performance on unseen data than a single train/test split, as it uses more data and reduces the impact of random variations in the split.

20. Which of the following statements is true regarding standardization and normalization?

a) Standardization and normalization always result in the same scaling of the data.

b) Standardization and normalization are only useful for continuous variables.

c) Standardization scales the data to have zero mean and unit variance, while normalization scales the data to have values between 0 and 1.

d) Normalization is only useful for variables with a normal distribution.

Answer: c) Standardization scales the data to have zero mean and unit variance, while normalization scales the data to have values between 0 and 1.

Explanation: Standardization and normalization are two common methods used to transform the features in a dataset to a common scale. Standardization rescales the data so that it has a mean of zero and a standard deviation of one. Normalization, on the other hand, rescales the data so that it falls between 0 and 1.

Option a) is incorrect because standardization and normalization can result in different scaling of the data depending on the original distribution of the data.

Option b) is incorrect because standardization and normalization can be applied to both continuous and categorical variables.

Option d) is incorrect because normalization can be applied to any variable regardless of its distribution.

Therefore, option c) is the correct answer as it accurately describes the differences between standardization and normalization.