DPP 01

DS & AI

Machine Learning

DPP: 1

Supervised Learning

Q1 In linear regression, how can feature selection impact the model?

S_1 : Increase overfitting risk.

S_2 : Reduce model interpretability.

S_3 : Improve model perfomance.

S_4 : No impact on model.

(A) 1 only (B) 3 only

(C) 1 and 2 (D) 2 and 3

Q2 What is the purpose of regularization in linear regression?

S_1 :To increase model complexity.

S_2 :To prevent overfitting.

S_3 :To handle missing data.

S_4 :To enhance model interpretability

(A) 1 and 3 (B) 2 only

(C) 2 and 4 (D) 1, 2, and 4

Q3 How does multicollinearity affect linear regression models?

s_1 : Increases model accuracy.

s_2 : Makes the model's estimates very sensitive to changes in the model.

s_3 : Reduces the speed of the algorithm.

s_4 : Improves model generalization.

(A) 1 and 3 (B) 2 only

(C) 1, 3, and 4 (D) 2 and 4

Q4 In linear regression, what does a coefficient of an independent variable represent?

s_1 : The mean of the dependent variable.

s_2 : The variance of the independent variable.

s_3 : The change in the dependent variable for a one-unit change in the independent variable.

s_4 : The correlation between the dependent and independent variables.

(A) 1 and 2 (B) 3 only

(C) 4 only (D) 2 and 3

Q5 What is the primary characteristic of Ridge Regression in the context of linear regression?

s_1 : It eliminates coefficients of less important features.

s_2 : It minimizes a penalized version of the least squares cost function with L1 penalty.

s_3 : It minimizes a penalized version of the least squares cost function with L2 penalty.

s_4 : It automatically selects the best features.

(A) 1 only (B) 2 only

(C) 3 only (D) 1 and 4

Q6 Which of the following statements correctly describes Lasso Regression?

s_1 : It can lead to the elimination of some coefficients, effectively performing feature selection.

s_2 : It uses an L2 penalty to reduce model complexity.

s_3 : It cannot handle multicollinearity in data.

s_4 :It always performs better than Ridge Regression.

(A) 1 and 3 (B) 1 only

(C) 2 and 4 (D) 3 and 4

Q7 What is the primary advantage of Elastic Net Regression?

s_1 : It combines the penalties of Ridge and Lasso regression.

s_2 : It is computationally faster than both Ridge and Lasso.

s_3 : It can only shrink coefficients but not eliminate them.

s_4 : It performs well only on large datasets.

(A) 1 only (B) 1 and 4

(C) 2 and 3 (D) 3 and 4

Q8 In which scenario is it more appropriate to use Lasso Regression over Ridge Regression?

s_1 : When there are a few significant variables and many insignificant ones.

s_2 : When multicollinearity is present in the dataset.

s_3 : When the number of features is much smaller than the number of observations.

s_4 : When the data is evenly distributed without outliers.

(A) 1 and 4 (B) 1 only

(C) 2 and 3 (D) 3 and 4

Q9 What does the R-squared metric indicate in a linear regression model?

s_1 : The proportion of the variance in the dependent variable that is predictable from the independent variables.

s_2 :The absolute fit of the model to the data.

s_3 :The ratio of correctly predicted instances to the total instances.

s_4 : The likelihood of the model to overfit.

(A) 1 only (B) 2 and 3

(C) 1 and 4 (D) 3 only

Q10 Why is adjusted R-squared considered a better measure than R-squared in multiple regression models?

s_1 : It penalizes the model for adding independent variables that do not improve the model.

s_2 : It always increases as more variables are added to the model.

s_3 : It measures the absolute error of the model.

s_4 : It is computationally less intensive than R-squared.

(A) 1 only (B) 2 and 3

(C) 1 and 4 (D) 3 only

Q11 When using Ridge or Lasso regression, which performance metric would be most appropriate to assess the model's explanatory power?

s_1 : Accuracy

s_2 : F1-score

s_3 : R-squared or Adjusted R-squared

s_4 : Precision and Recall

(A) 1 and 2

(B) 3 only

(C) 2 and 4

(D) All of the above

Q12 In which of the following scenarios is it more critical to use Adjusted R-squared instead of R-squared?

s_1 : When the model has a large number of independent variables.

s_2 : When the dataset is small.

s_3 : When there is high multicollinearity in the model.

s_4 : When using a simple linear regression model.

(A) 1 only

(B) 1 and 3

(C) 2 and 4

(D) All of the above

Q13 In the context of logistic regression, what does a confusion matrix represent?

s_1 : The coefficients of the logistic regression model.

s_2 : The number of instances where the predicted class matches the actual class.

s_3 : The matrix of precision and recall values for each class.

s_4 : The count of true positives, false positives, true negatives, and false negatives.

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

(C) 4 only (D) 2 only

Q14 What do precision and recall measure in a logistic regression model?

s_1 : Precision measures the ratio of correctly predicted positive observations to the total predicted positives. Recall measures the ratio of correctly predicted positive observations to all observations in actual class.

s_2 : Precision is the same as the R-squared value.

s_3 : Recall is the same as accuracy.

s_4 : Precision and recall measure the model's ability to handle imbalanced data.

(A) 1 only (B) 2 and 3

(C) 1 and 4 (D) 4 only

Q15 Why are precision, recall, and the confusion matrix often preferred over accuracy in evaluating logistic regression models?

S$_1$: They provide a more detailed view of the model's performance, especially in imbalanced datasets.

S$_2$: Accuracy always gives a misleading view.

S$_3$: They are computationally less intensive.

S$_4$: They are the only metrics available for logistic regression.

(A) 1 only

(B) 1 and 3

(C) 2 and 4

(D) All of the above

Q16 In logistic regression, what does the coefficient of an independent variable indicate?

S$_1$: The likelihood of the dependent variable being one.

S$_2$: The change in the log odds of the dependent variable for a one-unit change in the independent variable.

S$_3$: The exact probability of the dependent variable.

S$_4$: The variance explained by the independent variable.

(A) 2 only (B) 1 and 3

(C) 3 and 4 (D) 1, 2 and 4

Q17 What is a key difference between logistic regression and linear regression?

S$_1$: Logistic regression can only be used for binary outcomes, while linear regression can be used for any type of outcome.

S$_2$: Logistic regression is a type of linear regression.

S$_3$: Logistic regression models the probability of a specific outcome, while linear regression models the relationship between variables.

S$_4$: Logistic regression uses a different loss function compared to linear regression.

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

(C) 3 only (D) 3 and 4

Q18 In the context of logistic regression, what do the ROC Curve and AUC represent?

S$_1$: ROC Curve represents the relationship between sensitivity and specificity. AUC is the area under the ROC Curve.

S_2 : ROC Curve is a graphical representation of the confusion matrix. AUC measures accuracy.

S_3 : AUC represents the model's precision, while the ROC Curve represents recall.

S_4 : Both represent different forms of the R-squared metric.

(A) 1 only (B) 2 and 3

(C) 1 and 4 (D) 3 and 4

Q19 How does class imbalance affect the performance of a logistic regression model?

S_1 : It does not affect the performance.

S_2 : It can lead to a model that overly favors the majority class.

S_3 : It improves the model's ability to predict minority class instances.

S_4 : It affects the model's speed but not its accuracy.

(A) 1 only (B) 2 only

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

Q20 In logistic regression, what does a statistically significant coefficient suggest?

S_1 : The variable is important for predicting the outcome.

S_2 : The variable has a causal relationship with the outcome.

S_3 : The model is overfitting.

S_4 : The variable improves the model's accuracy.

(A) 1 only (B) 2 and 4

(C) 1, 2 and 4 (D) 3 only

Q21 What role does entropy play in building a decision tree classifier?

S_1 : It measures the purity of a split.

S_2 : It is used to calculate the depth of the tree.

S_3 : It defines the maximum number of leaves in the tree.

S_4 : It measures the homogeneity of a dataset.

(A) 1 and 4 (B) 2 only

(C) 1, 2 and 3 (D) 4 only

Q22 What does Gini impurity measure in a decision tree?

S_1 : The probability of misclassifying a randomly chosen element in the dataset.

S_2 : The overall accuracy of the decision tree.

S_3 : The depth of the decision tree.

S_4 : The computational complexity of the decision tree.

(A) 1 only (B) 1 and 2

(C) 2, 3 and 4 (D) 1 and 3

Q23 In decision tree algorithms, what is information gain used for?

S_1 : To measure the effectiveness of a split.

S_2 : To calculate the accuracy of the tree.

S_3 : To determine the number of trees in a random forest.

S_4 : To reduce the computational complexity.

(A) 1 only (B) 2 and 3

(C) 1, 2 and 4 (D) 4 only

Q24 How can overfitting be addressed in decision tree models?

S_1 : By increasing the depth of the tree.

S_2 : By pruning the tree.

S_3 : By using a larger training dataset.

S_4 : By adding more features to the dataset.

(A) 2 only

(B) 2 and 3

(C) 1 and 4

(D) All of the above

Q25 What does the bias-variance trade-off refer to in machine learning models?

S_1 : The trade-off between the model's complexity and its accuracy on the training data.

S_2 : The compromise between underfitting and overfitting a model.

S_3 : The balance between the precision and recall of a model.

S_4 : The correlation between the number of features and the model's performance.

(A) 1 and 2 (B) 2 only

(C) 1, 2 and 4 (D) 3 only

Q26 Which of the following are indicative of a model with high bias?

S_1 : The model performs poorly on both training and test data.

S_2 : The model is overly complex and has too many parameters.

S_3 : The model is too simplistic and overlooks the data's underlying trends.

S_4 : The model has a very high accuracy on training data but performs poorly on test data.

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

(C) 1, 2 and 4 (D) 3 only

Q27 What characteristics are associated with a high variance model?

S_1 : The model captures the training data very well, often at the expense of generalization.

S_2 : The model is overly simplistic and does not fit the training data well.

S_3 : The model performs equally well on training and unseen data.

S_4 : The model has too few parameters or features.

(A) 1 only (B) 2 and 4

(C) 1 and 3 (D) 3 only

Q28 Which techniques can help in managing the bias-variance trade-off in machine learning models?

S_1 : Increasing the model complexity.

S_2 : Simplifying the model.

S_3 : Regularization techniques.

S_4 : Using more training data.

(A) 1 and 4

(B) 2 and 3

(C) 3 and 4

(D) All of the above

Q29 What is the primary principle behind bagging in machine learning?

S_1 : It creates multiple models sequentially, each correcting the errors of the previous one.

S_2 : It involves dividing the dataset into random subsets without replacement and training a model on each subset.

S_3 : It builds multiple models in parallel, each on a random subset of the data with replacement, and aggregates their

predictions.

S_4 : It increases the model complexity to reduce bias.

(A) 1 only (B) 3 only

(C) 2 and 4 (D) 1 and 4

Q30 Which statement best describes the boosting technique in machine learning?

S_1 : It builds multiple models sequentially, each model attempting to correct the errors of the previous model.

S_2 : It combines models that are built in parallel on the same dataset.

S_3 : It focuses on reducing bias by increasing model complexity.

S_4 : It creates a single, highly complex model for better performance.

(A) 1 only (B) 2 and 3

(C) 1 and 3 (D) 4 only

Q31 How is a Random Forest model related to bagging?

S_1 : Random Forest is a specific type of bagging applied to decision trees.

S_2 : Random Forest uses boosting on decision trees.

S_3 : Random Forest builds a single, deep decision tree on the entire dataset.

S_4 : Random Forest and bagging are entirely different with no relation.

(A) 1 only (B) 2 and 3

(C) 3 only (D) 4 only

Q32 How does boosting potentially lead to overfitting?

S_1 : By focusing too much on the data's noise and outliers.

S_2 : By creating too many models in parallel.

S_3 : By under-emphasizing instances that were previously misclassified.

S_4 : By ignoring the variance in the dataset.

(A) 1 only (B) 2 and 4

(C) 3 only (D) 1 and 4

Answer Key

Q1	(B)
Q2	(B)
Q3	(B)
Q4	(B)
Q5	(C)
Q6	(B)
Q7	(A)
Q8	(B)
Q9	(A)
Q10	(A)
Q11	(B)
Q12	(A)
Q13	(C)
Q14	(A)
Q15	(A)
Q16	(A)
Q17	(C)
Q18	(A)
Q19	(B)
Q20	(A)
Q21	(A)
Q22	(A)
Q23	(A)
Q24	(B)
Q25	(B)
Q26	(A)
Q27	(A)
Q28	(D)
Q29	(B)
Q30	(C)
Q31	(A)
Q32	(A)

Hints & Solutions

Q1 Text Solution:

Explanation (B) : Proper feature selection can improve model performance by reducing noise and focusing on relevant variables. It doesn't inherently increase overfitting or reduce interpretability. In fact, it often enhances interpretability by simplifying the model.

Q2 Text Solution:

Explanation (B) : Regularization, such as L1 (Lasso) and L2 (Ridge) regularization, is used to prevent overfitting by adding a penalty to the loss function for large coefficients.

Q3 Text Solution:

Explanation 2 only: Multicollinearity, where two or more independent variables are highly correlated, makes the model's estimates highly sensitive and unreliable, as small changes in the data can lead to large changes in the model coefficients.

Q4 Text Solution:

Explanation (B): The coefficient of an independent variable in a linear regression model indicates the expected change in the dependent variable for a one-unit increase in that independent variable, assuming all other variables remain constant.

Q5 Text Solution:

Explanation (C): Ridge Regression introduces an L2 penalty (squared magnitude of coefficients) to the cost function. This method doesn't eliminate coefficients but shrinks them closer to zero, which helps in reducing model complexity and overfitting.

Q6 Text Solution:

Explanation (B): Lasso Regression introduces an L1 penalty (absolute value of the magnitude of coefficients), which can reduce some coefficients to zero, thus performing feature selection. It does not always perform better than Ridge Regression, and the choice between Lasso and Ridge depends on the specific dataset and problem.

Q7 Text Solution:

Explanation (A): Elastic Net is a regularization technique that combines both L1 and L2 penalties (from Lasso and Ridge). This allows it to inherit the benefits of both methods: feature selection from Lasso and regularization from Ridge, making it effective in cases where there are correlations between parameters.

Q8 Text Solution:

Explanation (1 only): Lasso Regression is particularly useful when you suspect that only a few features are actually important and many features are irrelevant or redundant. This is because Lasso can shrink the coefficients of less important features to zero, effectively performing feature selection. Ridge Regression, on the other hand, is better suited when most features contribute to the output.

Q9 Text Solution:

Explanation (a): R-squared, or the coefficient of determination, is a statistical measure that represents

the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a

regression model. It does not directly indicate the absolute fit, the prediction accuracy ratio, or the likelihood of overfitting.

Q10 Text Solution:

Explanation (A): Adjusted R-squared adjusts the statistic based on the number of independent variables

in the model. Unlike R-squared, which can increase with the addition of non-significant variables, adjusted R-squared will

decrease if the added variables don't improve the model significantly. This makes it a more reliable measure, especially

for multiple regression models with several independent variables.

Q11 Text Solution:

Explanation (B): In the context of Ridge and Lasso regression, which are linear models, R-squared or

Adjusted R-squared are the most appropriate metrics to assess the model's explanatory power, as they measure how

well the model captures the variance in the dependent variable. Accuracy, F1-score, Precision, and Recall are more

suitable for classification problems, not for regression.

Q12 Text Solution:

Explanation (A): Adjusted R-squared is particularly important in models with a large number of independent variables. It adjusts for the number of predictors in the model, providing a more accurate measure of the goodness of fit when compared to R-squared, especially in complex models. It is less critical in simple linear regression models with few predictors, where R-squared alone can be a sufficient measure of model fit.

Q13 Text Solution:

Explanation (C): In logistic regression, a confusion matrix is a table used to describe the performance of

a classification model. It contains counts of true positives, false positives, true negatives, and false negatives. This helps

in assessing the accuracy and effectiveness of the classification model.

Q14 Text Solution:

Explanation (A): In the context of logistic regression, precision is the ratio of correctly predicted positive observations to the total predicted positive observations, while recall (or sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. They do not equate to R-squared or accuracy, though they are important metrics, especially in cases of imbalanced datasets.

Q15 Text Solution:

Explanation(A): Precision, recall, and the confusion matrix are often preferred over accuracy because

they provide a more nuanced view of the model's performance, particularly in cases of imbalanced class distributions.

Accuracy can be misleading in these scenarios as it might be skewed by the majority class. These metrics are not the only ones available for logistic regression, nor are they preferred because of computational efficiency.

Q16 Text Solution:

Explanation (A): In logistic regression, the coefficient of an independent variable indicates the change in the log odds of the dependent variable (usually a binary outcome) for a one-unit change in the independent variable. This coefficient helps in understanding the influence of each independent variable on the likelihood of the dependent variable being one.

Q17 Text Solution:

Explanation(C): Logistic regression is used for binary classification problems and models the probability of a specific outcome. It differs from linear regression, which is used to model the relationship between variables and can be used for continuous outcomes. The logistic regression model outputs probabilities, using a logistic function to map predicted values to probabilities.

Q18 Text Solution:

Explanation(A): In logistic regression, the Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. The Area Under the Curve (AUC) represents the measure of the ability of the classifier to distinguish between classes and is used as a summary of the ROC curve

Q19 Text Solution:

Explanation(B): Class imbalance can significantly affect the performance of a logistic regression model by leading it to favor the majority class. In such cases, even a naive model that always predicts the majority class can appear to have high accuracy, but it would fail to meaningfully predict the minority class, which is often of greater interest.

Q20 Text Solution:

Explanation(A): In logistic regression, a statistically significant coefficient indicates that the variable is important for predicting the outcome. However, it does not imply a causal relationship, nor does it directly speak to the model's accuracy or overfitting. Statistical significance suggests that the association observed in the data is not likely due to random chance.

Q21 Text Solution:

Explanation(A): In the context of decision trees, entropy is a measure of the disorder or uncertainty in the data. It is used to quantify the impurity or randomness in a dataset and is crucial in the decision-making process of a decision tree, particularly in determining how a dataset should be split.

Q22 Text Solution:

Explanation (A): Gini impurity is a measure used in decision trees to quantify the likelihood of a random sample being incorrectly labeled if it was randomly labeled according to the distribution of labels in the dataset. It reflects the probability of misclassification and is a criterion for splitting nodes in a tree, aiming to minimize this impurity.

Q23 Text Solution:

Explanation(A): Information gain in decision trees is used to measure how well a given attribute separates the training examples according to their target classification. It is a key metric for deciding the node at which the dataset is split. A higher information gain implies a more effective split, reducing the impurity in the resulting groups.

Q24 Text Solution:

Explanation(B): Overfitting in decision trees can be addressed by pruning the tree, which involves cutting back branches of the tree to reduce its complexity and to make it more generalizable to new data. Additionally, using a larger training dataset can also help in reducing overfitting by providing more varied data for the model to learn from. Increasing the depth of the tree and adding more features, on the other hand, can potentially lead to more overfitting.

Q25 Text Solution:

Explanation(B): The bias-variance trade-off is a fundamental concept in machine learning that deals with the compromise between underfitting and overfitting a model. Bias refers to the error due to overly simplistic assumptions in the learning algorithm, leading to underfitting. Variance refers to the error due to too much complexity in the learning algorithm, leading to overfitting. A good model must balance these two to achieve good predictive performance.

Q26 Text Solution:

Explanation(A): High bias in a model indicates that the model is too simplistic and is not capturing the underlying trends and patterns in the data, leading to poor performance on both training and test datasets. It is not related to the complexity of the model or its performance on just the training data.

Q27 Text Solution:

Explanation(A): A high variance model captures the training data very well, including its noise and outliers, which often leads to poor generalization to unseen data (overfitting). It is not characterized by simplicity or a lack of parameters, which are traits of high bias.

Q28 Text Solution:

Explanation(D): Managing the bias-variance trade-off can involve different strategies based on the problem at hand. Increasing model complexity can reduce bias but might increase variance. Simplifying the model can reduce variance but might increase bias. Regularization techniques are used to reduce overfitting (high variance) without significantly increasing bias. Using more training data can help reduce variance without affecting bias.

Q29 Text Solution:

Explanation(B): Bagging, short for bootstrap aggregating, involves creating multiple models in parallel, each trained on a random subset of the data with replacement (bootstrap samples). The final prediction is typically an average (for regression) or a majority vote (for classification) of the predictions from all models, reducing variance and improving robustness.

Q30 Text Solution:

Explanation(C): Boosting builds models sequentially, with each subsequent model focusing on the errors made by the previous models. This technique aims to reduce bias and improve the accuracy of the model. Unlike bagging, which focuses on reducing variance by building models in parallel, boosting reduces bias by emphasizing the instances that previous models misclassified.

Q31 Text Solution:

Explanation(A): A Random Forest is a type of bagging ensemble method where multiple decision trees are created and trained on different subsets of the dataset (with replacement), and their predictions are averaged (in regression) or voted on (in classification). This method effectively reduces variance and avoids overfitting, which is a common problem with individual, deep decision trees.

Q32 Text Solution:

Explanation(A): Boosting can lead to overfitting if it focuses too intensely on the training data, especially on the noise and outliers. This happens because each new model in the sequence places more emphasis on the instances that previous models misclassified, potentially leading to a model that is too complex and too tailored to the training data. This overemphasis on correcting errors can cause the model.