

BUSINESS INTELLIGENCE AND ANALYTICS

ASSIGNMENT WEEK 7:

Total marks = 17 Marks
14 Qns * 1 marks = 14 marks
1 Qns * 3 marks = 3 marks

1. What is a key advantage of decision trees in knowledge representation? (1 Mark)
 - a) They possess a complex representation
 - b) They have a slow learning process
 - c) They represent acquired knowledge in an intuitive tree form
 - d) They handle only unidimensional data

Answer: C) They represent acquired knowledge in an intuitive tree form

2. What does an internal node in a decision tree represent? (1 Mark)
 - a) Class label
 - b) Outcome of a test on an attribute
 - c) Unknown tuple
 - d) Root node

Answer: B) Outcome of a test on an attribute

3. Why might a decision tree, resulting from the described process, perform poorly on a test set? (1 Mark)
 - a) Due to too few splits leading to underfitting
 - b) Because it has too few leaves
 - c) It's likely to have too many splits, causing overfitting
 - d) It has high bias but low variance

Answer: C) It's likely to be to have too many splits, causing overfitting

4. What might a smaller tree with fewer splits achieve in terms of variance and bias? (1 Mark)
 - a) It reduces both variance and bias
 - b) It reduces variance but possibly increases bias
 - c) It increases both variance and bias
 - d) It doesn't affect variance or bias

Answer: B) It reduces variance but possibly increases bias

5. How does a classification tree differ from a regression tree? (1 Mark)
 - a) It predicts a quantitative response, not a qualitative one.
 - b) It predicts a qualitative response, not a quantitative one.
 - c) It's less accurate than a regression tree.
 - d) It always predicts the mean response.

Answer: B) It predicts a qualitative response, not a quantitative one.

6. Which of the following is an advantage of decision trees compared to linear regression? (1 Mark)
- a) Decision trees are more accurate in predicting continuous outcomes.
 - b) Decision trees are easier to explain and interpret.
 - c) Decision trees require the creation of dummy variables for qualitative predictors.
 - d) Decision trees are less affected by outliers.

Answer: B) Decision trees are easier to explain and interpret.

7. Ensemble methods are used to improve prediction performance of decision trees (T/F) (1 Mark)

Answer: True

8. Bagging primarily addresses which issue within statistical learning methods like decision trees? (1 Mark)
- a) Increases the computational complexity of the models.
 - b) Reduces the need for accurate parameter tuning in models.
 - c) Deals with high variance and improves prediction accuracy.
 - d) Reduces high bias and improves prediction accuracy.

Answer: C) Deals with high variance and improves prediction accuracy.

9. Which technique involves averaging predictions from multiple models built on bootstrapped training sets? (1 Mark)
- a) Decision tree pruning
 - b) Bootstrap aggregation (bagging)
 - c) Recursive feature elimination
 - d) Ridge regression

Answer: b) Bootstrap aggregation (bagging)

10. How does bagging handle classification problems? (1 Mark)
- a) By averaging predictions from different classifiers
 - b) By constructing decision trees without bootstrapping
 - c) By taking a majority vote from predictions of multiple trees
 - d) By using only a single tree for classification

Answer: c) By taking a majority vote from predictions of multiple trees

11. In Bagging, each individual tree is independent of each other because they consider different subset of features and samples. (T/F) (1 Mark)

Ans: True

12. What are some common techniques for handling imbalanced data in classification tasks? (1 Mark)
- a) Oversampling the minority class to create a more balanced dataset.
 - b) Under sampling the majority class to reduce its dominance.
 - c) Only a is correct
 - d) Both a and b are correct

Ans: d) Both a and b are correct

13. In Random forest you can generate hundreds of trees (say T_1, T_2, \dots, T_n) and then aggregate the results of these trees. Which of the following is true about an individual (T_k) tree in Random Forest? (1 Mark)

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

- a) 1 and 3
- b) 1 and 4
- c) 2 and 3
- d) 2 and 4

Solution: A) 1 and 3

14. Consider a dataset with a binary target variable (0 or 1) and a split based on a feature resulting in two child nodes after the split.

- Node 1 (left child): Out of 40 samples, 30 belong to class 0 and 10 belong to class 1.
- Node 2 (right child): Out of 60 samples, 20 belong to class 0 and 40 belong to class 1.

which option has the correct Gini indices of the child nodes? (3 Marks)

- a) Gini index for Node 1: 0.375, Gini index for Node 2: 0.444
- b) Gini index for Node 1: 0.375, Gini index for Node 2: 0.320
- c) Gini index for Node 1: 0.425, Gini index for Node 2: 0.320
- d) Gini index for Node 1: 0.444, Gini index for Node 2: 0.375

Solution: b) Gini index for Node 1: 0.375, Gini index for Node 2: 0.444

15. How does Random Forest aim to reduce correlation among trees? (1 Mark)

- a) By constructing trees sequentially based on the residuals.
- b) By growing trees independently with a random subset of predictors at each split.
- c) By fitting trees to the residuals from the current model.
- d) By sequentially building trees using information from previously grown trees.

Answer: b) By growing trees independently with a random subset of predictors at each split.