

day=13

## Assessment -1

Answers

- (1) a) Large K with noisy data.
- (2) (c) They overfit
- (3) (c) Reduce variance
- (4) (c) All features are considered at each split
- (5) (c) Features are independent
- (6) (c) Sigmoid
- (7) (d) F1-score
- (8) (d) Overfitting
- 9) a) To reduce bias
- 10) c) Logistic regression

### 11) Overfitting in Decision Trees :

- overfitting is a condition where a model starts learning a ~~data~~ training data rather than target patterns, logic and relationship with unseen features and can't perform well on new data.
- Decision trees are unstable learners because they are prone to overfitting means low bias but high variance
- generally, ~~the~~ max-depth is a hyperparameter in decision trees which decides that how deep the splits are gone, if we select the high value for it the model can easily be prone to overfitting because all the focus will be biased towards splitting maximum as possible and which cause overfitting on training data cause overfitting on testing data.

→ solution, use bagging technique → random forest  
It solves the problem of overfitting (low bias, high variance) to (low bias, low variance)  
by training n-estimators → no. of decision trees  
parallelly on n-samples of data by using  
technique called as bootstrapping which  
will send samples of the original data  
by replacing rows and features depend on us.

→ and n-estimators learns on diff samples  
of training data and produce outputs

→ and we use techniques like voting for classification  
and average for numerical outputs of diff trees  
and aggregate their result and produce  
best model from no. of weak learners.  
that's how bagging resolve the decision tree  
problem

### Aus: 12) Random Forest Working:

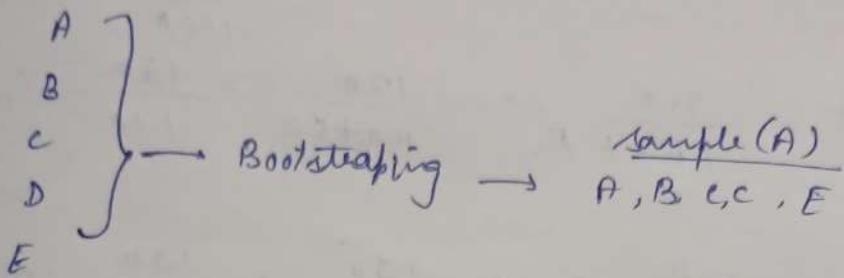
Random Forest is Ensemble learning bagging  
technique, which train no. of decision  
trees on samples of data called weak  
learners and aggregate their outputs  
to produce best model.

→ Bootstrap sampling: It is a sampling  
data technique used by the random  
forest to train no. of decision trees on  
diff samples of same of training data.

it uses row sampling and column sampling  
to produce a data frame in a way that not  
same data will repeat for all decision trees

like with row replacement  $\rightarrow$  some rows can be  
repeated in single sample  
similarly about column replacement

fg: Row id



Some rows can be included  $\approx 68\%$  to repeat or some can never be included  $\approx 32\%$ . This can be taken as for unseen data generally.  
 i.e. without replacement  $\rightarrow$  If a row is included in a bag, it will not be considered out of bag means one can't repeat.

→ Random Feature selection: similarly some features are included ~~or~~ or can't be included for different samples, features are selected randomly

Majority Voting: It is technique when we have classification problem output select that one for which most of the models produce maximum same output Max Voting

Ex:

```

graph LR
    DT1[Decision Tree 1] -- output --> Class0[Class 0]
    DT2[DT2] --> Class1_1[Class 1 ✓]
    DT3[DT3] --> Class1_2[Class 1 ✗]
    DT4[DT4] --> Class1_3[Class 1 ✓]
    DT5[DT5] --> Class0[Class 0]
    DT6[DT6] --> Class1_4[Class 1 ✓]
  
```

$l = 4$

$0 = 2$

∴ Best model will seek to output  $\rightarrow$  class 1, selected as max week 1 learns produce class 1.

Ans - 13

$$\rightarrow \text{(a) Accuracy} : \frac{T.P + T.N}{T.P + T.N + F.N + F.P} = \frac{120 + 800}{120 + 800 + 50 + 30} = \frac{920}{1000} = 0.92$$

$$\text{b) Precision} : \frac{T.P}{T.P + F.P} = \frac{120}{120 + 50} = \frac{120}{170} = 0.70$$

$$\text{c) Recall} : \frac{T.P}{T.P + F.N} = \frac{120}{120 + 30} = \frac{120}{150} = 0.80$$

$$\text{(d) F1-score} \rightarrow \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.70 * 0.80}{0.70 + 0.80} = \frac{1.12}{1.5} = 0.74$$

Actually the accuracy about 92%, but according to confusion matrix actually 150 are actually fraud out of 120 are positively classified as fraud by model, 30 are misclassified.

So model can be considered but can be improved for fraud detection.

Ans - 14

(i) if max\_features = None.  $\rightarrow$  all features are considered at each split

(ii) model will overfit more

(iii) If we choose n\_estimators = 200, the model will perform well on training data and produce predictions with more ~~accuracy~~ ~~value~~.

it will accept low variance and few bids  $\rightarrow$  bias,  
variance trade-off.