DWM

ASSIGNMENT - 2

CLASSIFICATION

JUNAID. GIRKAR
60004190057
TE COMPS A4

08/11/2021

Q1 For the dataset given below, perform NB classifier to predict X = < Young, Myope, Yes, Reduced >

RECORD ID	AGE	SPECTACLE PRESCRIPTION	ASTIGMATIC	TEAR PRODUCTION RATE	CLASS LABEL LENSES
1	Young	Myope	No	Reduced	Noncontact
2	Young	Myope	No	Normal	Soft contact
3	Pre-presbyopic	Myope	No	Normal	Soft contact
4	Pre-presbyopic	Myope	Yes	Reduced	Noncontact
5	Presbyopic	Myope	No	Normal	Noncontact
6	Presbyopic	Myope	Yes	Reduced	Noncontact
7	Young	Hypermetrope	Yes	Reduced	Noncontact
8	Young	Hypermetrope	Yes	Normal	Hard contact
9	Pre-Presbyopic	Myope	No	Reduced	Noncontact
10	Young	Hypermetrope	No	Reduced	Noncontact
11	Young	Hypermetrope	No	Normal	Soft contact
12	Pre-Presbyopic	Myope	Yes	Normal	Hard contact
13	Pre-Presbyopic	Hypermetrope	No	Reduced	Noncontact
14	Pre-Presbyopic	Hypermetrope	Yes	Normal	Noncontact
15	Presbyopic	Myope	No	Reduced	Noncontact
16	Pre-presbyopic	Hypermetrope	No	Normal	Soft contact
17	Pre-presbyopic	Hypermetrope	Yes	Reduced	Noncontact
18	Presbyopic	Myope	Yes	Normal	Hard contact
19	Young	Myope	Yes	Reduced	Noncontact
20	Young	Myope	Yes	Normal	Hard contact

Let A → Age
S → Spectacle Prescription
B → Astigmatism
T → Tear production rate
L → class label lense
X → (Young, Myope, Yes, Reduced)

(i) $P(\text{Young} \mid \text{Non Contact}) = \frac{P(\text{Non contact} \mid \text{Young}) \cdot P(\text{Young})}{P(\text{Non-contact})}$

$= \frac{\frac{4}{12} \times \frac{8}{12}}{\frac{12}{20}} = \frac{\frac{4}{8} \times \frac{8}{20}}{\frac{12}{20}} = \frac{4}{12} = \frac{1}{3}$

$\therefore P(\text{Young} \mid \text{Non-contact}) = \frac{1}{3}$

(ii) $P(\text{Myope} \mid \text{Non-contact}) = \frac{P(\text{Non-Contact} \mid \text{Myope}) \cdot P(\text{Myope})}{P(\text{Non-contact})}$

$= \frac{\frac{7}{12} \times \frac{12}{20}}{\frac{12}{20}}$

$= \frac{7}{12}$

$\therefore P(\text{Myope} \mid \text{Non-contact}) = \frac{7}{12}$.

(iii) $P(\text{Yes} \mid \text{Non-Contact}) = \dfrac{P(\text{Non-Contact} \mid \text{Yes}) \cdot P(\text{Yes})}{P(\text{Non-contact})}$

$= \dfrac{\frac{6}{10} \times \frac{10}{20}}{\frac{12}{20}} = \dfrac{1}{2}$

$\therefore P(\text{Yes} \mid \text{Non-Contact}) = \dfrac{1}{2}$

(iv) $P(\text{Reduced} \mid \text{Non-Contact}) = \dfrac{P(\text{Non-contact} \mid \text{Reduced}) \cdot P(\text{Reduced})}{P(\text{Non-Contact})}$

$= \dfrac{\frac{10}{10} \times \frac{10}{20}}{\frac{12}{20}} = \dfrac{5}{6}$

$\therefore P(\text{Reduced} \mid \text{Non-Contact}) = \dfrac{5}{6}$

(v) $P(X \mid \text{Non-Contact}) = P(\text{Non-contact} \mid X)'$

$= \dfrac{1}{3} \times \dfrac{7}{12} \times \dfrac{1}{2} \times \dfrac{5}{6}$

$= \dfrac{35}{432}$

$= 0.081$

$\therefore P(X \mid \text{Non-Contact}) = 0.081$

(vi) $P(\text{Young} \mid \text{Soft contact}) = \dfrac{P(\text{soft contact} \mid \text{young}) \cdot P(\text{young})}{P(\text{soft contact})}$

$$= \frac{\frac{2}{8} \times \frac{8}{20}}{\frac{4}{20}} = \frac{1}{2}$$

$\therefore P(\text{Young} \mid \text{Soft Contact}) = \frac{1}{2}$.

(vii) $P(\text{Myopic} \mid \text{soft Contact}) = \dfrac{P(\text{Soft Contact} \mid \text{Myopic}) \cdot P(\text{Myopic})}{P(\text{soft contact})}$

$$= \frac{\frac{2}{12} \times \frac{12}{20}}{\frac{4}{20}} = \frac{1}{2}$$

$\therefore P(\text{Myopic} \mid \text{soft contact}) = \frac{1}{2}$.

(viii) $P(\text{Yes} \mid \text{soft contact}) = \dfrac{P(\text{soft contact} \mid \text{Yes}) \cdot P(\text{Yes})}{P(\text{soft contact})}$

$$= \frac{\frac{0}{10} \times \frac{10}{20}}{\frac{4}{20}} = 0$$

$\therefore P(\text{Yes} \mid \text{soft contact}) = 0$

3

(ix) $P(\text{Young} \mid \text{Hard Contact}) = \dfrac{P(\text{Hard contact} \mid \text{Young}) \cdot P(\text{Young})}{P(\text{Hard Contact})}$

$$= \frac{\frac{2}{8} \times \frac{8}{20}}{\frac{4}{20}} = \frac{1}{2}$$

$\therefore P(\text{Young} \mid \text{Hard Contact}) = \frac{1}{2}$.

(x) $P(\text{Myopic} \mid \text{Hard contact}) = \dfrac{P(\text{Hard Contact} \mid \text{Myopic}) \cdot P(\text{Myopic})}{P(\text{Hard contact})}$

$$= \frac{\frac{3}{12} \times \frac{12}{20}}{\frac{4}{20}} = \frac{3}{4}$$

$\therefore P(\text{Myopic} \mid \text{Hard Contact}) = \frac{3}{4}$

(xi) $P(\text{Yes} \mid \text{Hard contact}) = \dfrac{P(\text{Hard Contact} \mid \text{Yes}) \cdot P(\text{Yes})}{P(\text{Hard Contact})}$

$$= \frac{\frac{4}{10} \times \frac{10}{20}}{\frac{4}{20}} = 1$$

$\therefore P(\text{Yes} \mid \text{Hard Contact}) = 1$

(xii) $P(\text{Reduced} \mid \text{Hard Contact}) = \dfrac{P(\text{Hard Contact} \mid \text{Reduced}) \cdot P(\text{Reduced})}{P(\text{Hard contact})}$

$$= \frac{\frac{0}{10} \times \frac{10}{20}}{\frac{4}{20}} = 0$$

$\therefore P(\text{Reduced} \mid \text{Hard Contact}) = 0$

(xiii) $P(X \mid \text{Hard Contact}) = 0$

$\therefore P(X \mid \text{Non Contact}) > P(X \mid \text{Soft contact}) = P(X \mid \text{Hard Contact})$

$\therefore 0.081 > 0 > 0$

$\therefore$ Prediction of $\langle \text{Young}, \text{Myope}, \text{Yes}, \text{Reduced} \rangle \rightarrow$ Non Contact

Q2 The following table consists training data from an employee database. The data has been generalized. For example, "31 ::: 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age and salary given in that row.

DEPARTMENT	STATUS	AGE	SALARY	COUNT
sales	senior	31 ::: 35	46K to 50K	30
sales	junior	26 ::: 30	26K to 30K	40
sales	junior	31 ::: 35	31K to 35K	40
systems	junior	21 ::: 25	46K to 50K	20
systems	senior	31 ::: 35	66K to 70K	5
systems	junior	26 ::: 30	46K to 50K	3
systems	senior	41 ::: 45	66K to 70K	3
marketing	senior	36 ::: 40	46K to 50K	10
marketing	junior	31 ::: 35	41K to 45K	4
secretary	senior	46 ::: 50	36K to 40K	4
secretary	junior	26 ::: 30	26K to 30K	6

Let status be class label attribute.

a How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple.

b use your algorithm to construct a decision tree

c Given a data tuple having the values "systems", "26 ... 30" and "46K to 50K" for the attributes department, age, and salary respectively, what would a NB classification of the status for the tuple be?

ANS a] The basic decision tree algorithm should be modified as follows to take into consideration the count of each generalized data tuple :-

- The count of each tuple must be integrated into the calculation of attribute selection measuring
- Take the count into consideration to determine the most common class among the tuples.

ANS b] $P = 113$

$N = 52$

$$DE = -\frac{113}{52}\log_2\left(\frac{113}{11}\right) - \frac{52}{165}\log_2\left(\frac{52}{165}\right)$$

$DE = 0.8990$

DEPARTMENT

	P_i	N_i	Entropy
sales	80	30	0.8454
System	23	8	0.8238
Marketing	4	10	0.8631
Secretary	6	4	0.9709

$$\text{Entropy for department} = \frac{80+30}{165}(0.8454) + \frac{23+8}{165}(0.8238) + \frac{4+10}{165}(0.8631) + \frac{6+4}{165}(0.9709)$$

$$= \frac{110}{165}(0.8454) + \frac{31}{165}(0.8238) + \frac{14}{165}(0.8631) + \frac{10}{165}(0.9709)$$

$$= 0.8504$$

Gain = 0.899 − 0.8504

= 0.049

AGE	P	N	Entropy
21...25	20	0	0
26...30	49	0	0
31...35	44	35	0.9906
36...40	0	0	0
41...45	0	0	0
46...50	0	0	0

Age entropy = $\frac{44+35}{165}$ (0.9906)

= $\frac{79}{165}$

= 0.4743

∴ Gain = 0.8990 − 0.4743

= 0.4247

SALARY	P	N	Entropy
26K ... 30K	46	0	0
31K ... 35K	40	0	0
36K ... 40K	0	4	0
41K ... 45K	4	0	0
46K ... 50K	23	40	0.9468
66K ... 70K	0	8	0

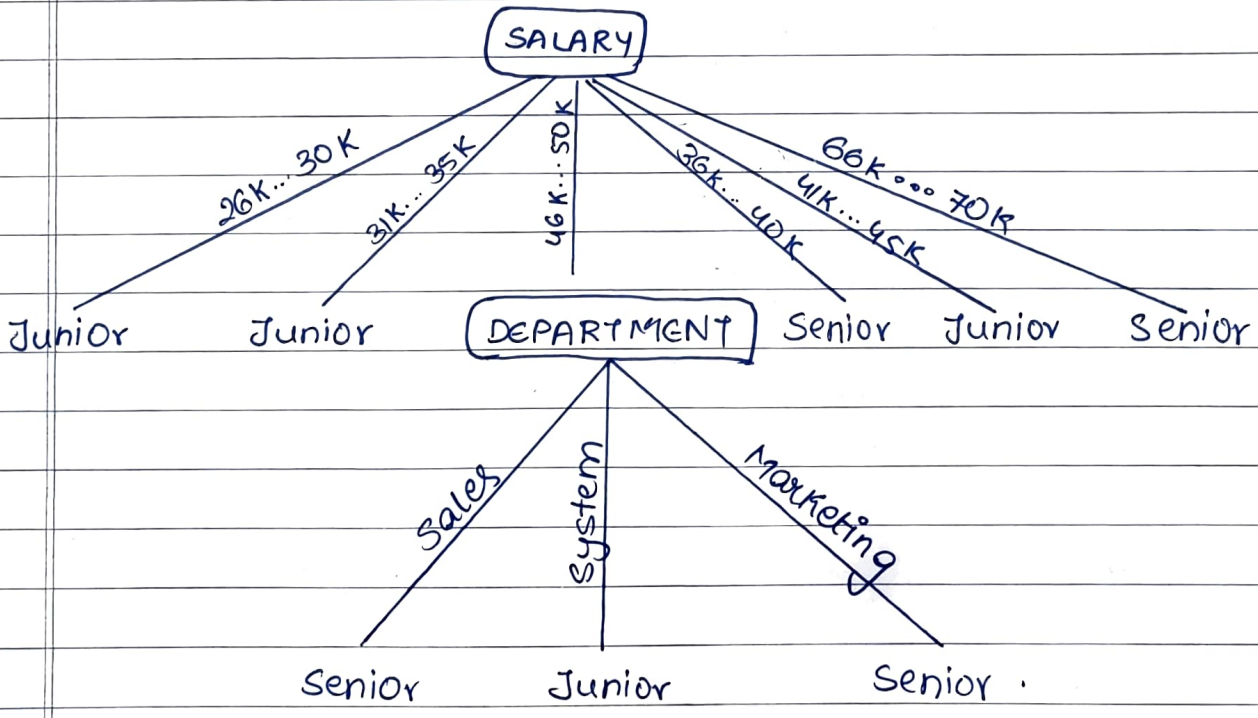
Salary Entropy $= \frac{23+40}{165} \times 0.9468$

$= \frac{63}{165} \times 0.9468$

$= 0.3615$

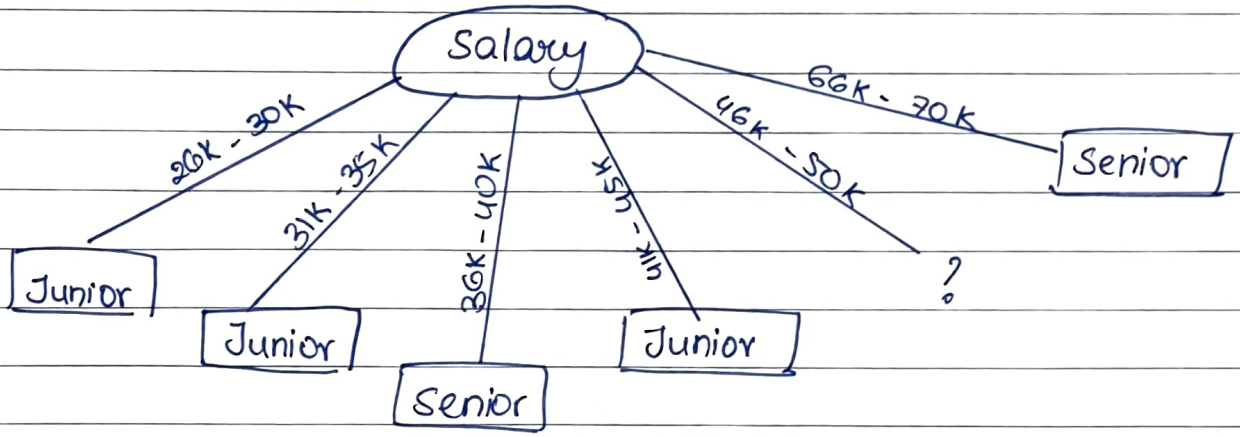
$\therefore$ Gain $= 0.5375$

DECISION TREE:



Now, Gain (salary) > Gain (age) > Gain (department)

$\therefore$ Salary is selected as splitting attribute



Now we consider data tuples having salary 46K-50K and find the splitting attribute for it.

$$Info(D_1) = \frac{-40}{63}\log_2\frac{40}{63} - \frac{23}{63}\log_2\frac{23}{63}$$

$$= 0.9468$$

consider department attribute

$$Info_{dept}(D_1) = \frac{30}{63} \times Info(Sales) + \frac{23}{63} Info(Systems) + \frac{10}{63} \times Info(Marketing)$$

$$= \frac{30}{63}\left(\frac{-30}{30}\log_2\frac{30}{30}\right) + \frac{23}{63}\left(\frac{-23}{23}\log_2\frac{23}{23}\right) + \frac{10}{63}\left(\frac{-10}{10}\log_2\frac{10}{10}\right)$$

$$= 0$$

$$Gain(dep) = 0.9468$$

consider age attribute

$$Info_{age}(D_1) = \frac{20}{63} Info(21\text{-}25) + \frac{3}{63} Info(26\text{-}30) + \frac{30}{63} Info(31\text{-}25) + \frac{10}{63} Info(36\text{-}40)$$

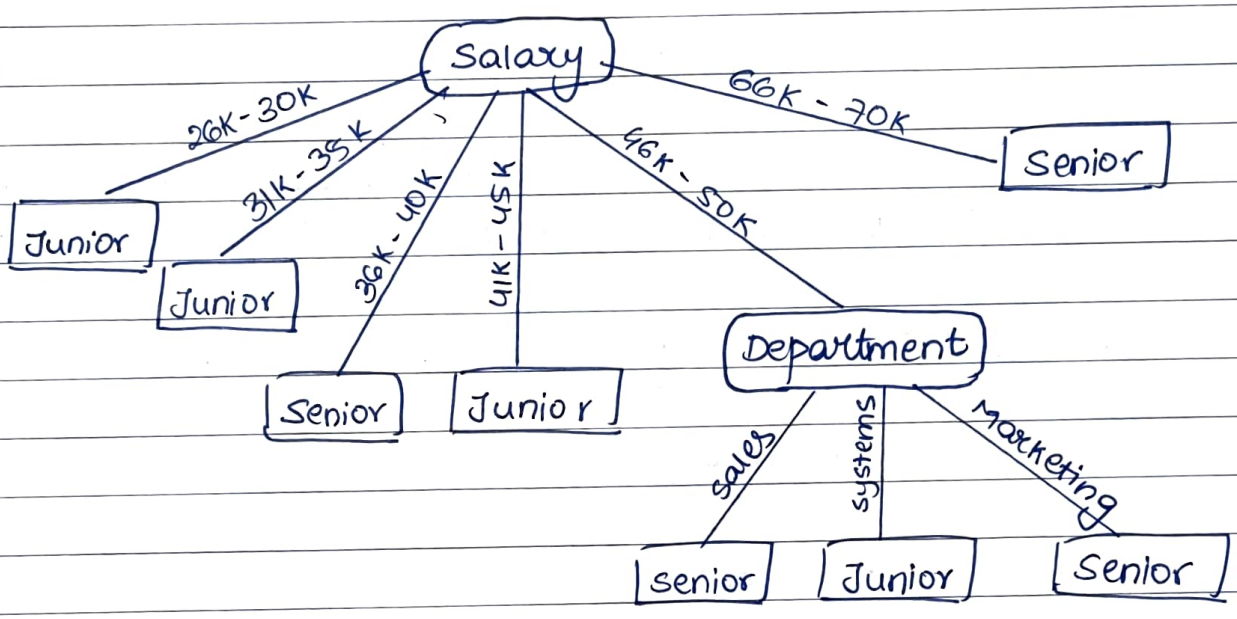
$$= \frac{20}{63}\left[\frac{-20}{20}\log_2\frac{20}{20}\right] + \frac{3}{63}\left[\frac{-3}{3}\log_2\frac{3}{3}\right] + \frac{30}{63}\left[\frac{-30}{30}\log_2\frac{30}{30}\right] + \frac{10}{63}\left[\frac{-10}{10}\log_2\frac{10}{10}\right]$$

$$= 0$$

Gain (age) = 0.9468

∵ Gain for both attributes are same we can choose any for splitting

∴ Decision Tree :-



Above tree is the final decision tree from Information Gain (ID3) algorithm.

ANS c] Prior probabilities :

$$P(\text{senior}) = \frac{52}{165}$$

$$P(\text{Junior}) = \frac{113}{165}$$

7

Consider attribute department

P (systems | senior) = 8/52
P (systems | junior) = 23/113

Consider attribute age
P(26-30 | senior) = 0/52
P(26-30 | junior) = 49/113

consider attribute salary
P(46-50K | senior) = 40/52
P(46-50K | junior) = 23/113

Posterior probabilities :

$P\left(\frac{X}{\text{junior}}\right) = \frac{23 \times 49 \times 23}{113 \times 113 \times 113}$

$= 0.018$

$P\left(\frac{X}{\text{junior}}\right) = \frac{8 \times 0 \times 40}{52 \times 52 \times 52}$

$= 0$

∴ Bayes theorem : $P\left(\frac{C_i}{X}\right) = \frac{P(C_i) \times P(X/C_i)}{P(X)}$

∴ $P\left(\frac{\text{junior}}{X}\right) = \frac{113}{165} \times 0.018$

$= 0.0123$

$P\left(\frac{\text{senior}}{X}\right) = 0 \times \frac{52}{165}$

$= 0$

∵ $0.0123 > 0$

∴ X = (systems, 26:::30, 46:::50K) will be predicted as status = junior.

Q3 Suppose a bank wants to develop a classifier that guards against fraudulent credit card transactions. Illustrate how you can induce a quality classifier based on a large set of non-fraudulent examples and a very small set of fraudulent cases.

ANS The case mentioned above is an example of imbalanced data from classification, where the non fraudulent cases from the majority class and the fraudulent cases from the minority class. To handle the imbalanced data in classifiers there are 2 approaches :-

1) Data based
2) Algorithm based

we can use them combined also.

(1) DATA BASED HANDLING

(i) Over sampling : In over sampling we increase the number of minority class keeping the number of majority class till balance is achieved.

EXAMPLE : Initially fraudulent : 10
Non-fraudulent : 990 ∞.

If we increase fraudulent to 10 times the original then fraudulent = 100, non-fraudulent = 990. we can increase number of fraudulent cases till balance is achieved.

(ii) Under sampling : In under sampling we decrease the number of majority class members keeping the number of minority class members till balance is achieved.

EXAMPLE : Initially Fraud : 10 , Non-fraud : 990

if we decrease the number of non-fraud cases by 50% then fraud : 10 and non-fraud : 495 ; we can decrease the number of non-fraud cases till balance is achieved.

(iii) Hybrid Sampling : In hybrid sampling we increase number of minority class members and decrease number of majority class members till balance of data is achieved.

EXAMPLE : Fraud : 10 ; Non-fraud : 990, if we increase fraud cases by 10 times and decrease non-fraud cases by 50% then fraud : 100, non-fraud : 495, we can do so till balance between data is achieved.

2) ALGORITHMIC BASED:

(i) Ensemble learning : Imbalanced data can be handled modifying existing classification algorithms to make them appropriate for handling imbalanced data. Ensemble learning involves involves constructing several 2 stage classifiers and aggregate their predictions, which gives higher performance than single classifiers.

BAGGING : It generates 'n' different training samples with replacement, and training the algorithm on each bootstrapped samples and then aggregating their predictions. This helps to balance the data and make predictions.

BOOSTING: It is a technique to combine weak learners to create a strong learner and make accurate predictions. At each iteration the new classifier places more weight on the cases which were predicted incorrect by previous classifier, thus giving more accurate predictions.

The above mentioned methods can be used to induce a quality classifier for the bank that wants to guard against fraudulent transactions where number of frauds <<< number of non-frauds.

Q4 Show that accuracy is a function of sensitivity and specificity $\text{accuracy} = \text{sensitivity} * \frac{P}{P+N} + \text{specificity} * \frac{N}{P+N}$

ANS $RHS = \text{sensitivity} * \frac{P}{P+N} + \text{specificity} * \frac{N}{P+N}$

Now, $\text{sensitivity} = \frac{TP}{P}$, $\text{specificity} = \frac{TN}{N}$

Substituting in RHS $= \frac{TP}{P} \times \frac{P}{P+N} + \frac{TN}{N} \times \frac{N}{P+N}$

$= \frac{TP + TN}{P+N} \longrightarrow (1)$

$= T$

Consider LHS = accuracy $= \frac{TP+TN}{P+N} \longrightarrow (2)$

From (1) and (2) LHS = RHS

$\therefore \text{accuracy} = \text{sensitivity} * \frac{P}{P+N} + \text{specificity} * \frac{N}{P+N}$

9

Q5 Compare and contrast bagging and boosting

ANS

BAGGING: Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here the objective is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

BAGGING STEPS:

- Suppose there are N observations and M features in training data set. A sample from training data set is taken randomly with replacement.
- A subset of M features are selected randomly and whichever feature gives the best split is used to split the node iteratively.
- The tree is grown to the largest.
- Above steps are repeated n times and prediction is given based on the aggregation of predictions from n number of trees.

ADVANTAGES:

- Reduces over-fitting of the model.
- Handles higher dimensionality data very well
- Maintains accuracy for missing data.

DISADVANTAGES:

- Since final prediction is based on the means predictions from subset trees, it won't give precise values for the classification and regression model.

BOOSTING: Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. consecutive trees are fit and at every step, the goal is to improve the accuracy from the prior tree. When an input is misclassified by a hypothesis, its weight is increased so that the next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model.

BOOSTING STEPS:

- Draw a random subset of training samples d1 without replacement from the training set D to train a weak learner C_1
- Draw a second random training subset d2 without replacement from the training set and add 50% of the samples that were previously falsely classified to train a weak learner C_2.
- Find the training sample d3 in the training set D on which C1 and C2 disagree to train a weak learner C3
- Combine all the weak learners via majority voting.

10

ADVANTAGES:

- Supports different loss functions
- Works well with interactions.

DISADVANTAGES:

- Prone to over-fitting
- Requires careful tuning of different hyper-parameters.

SIMILARITIES:

→ Both are ensemble methods to get N learners from 1 learner

→ Both generate several training data sets by random sampling

→ Both make the final decision by averaging the N learners

→ Both are good at reducing variance and provide higher stability.

BAGGING	BOOSTING
• The simplest way of combining predictions of same type.	• A way of combining predictions that belong to different types.
• Aims to decrease variance, not bias	• Aims to decrease bias, not variance
• Each model receives equal weight.	• Models are weighted according to their performance.
• Different training data sets are randomly taken without replacement from the entire training dataset.	• Every new subset contains the elements that were misclassified by previous models
• Example: Random forest	• Example: Adaboost.