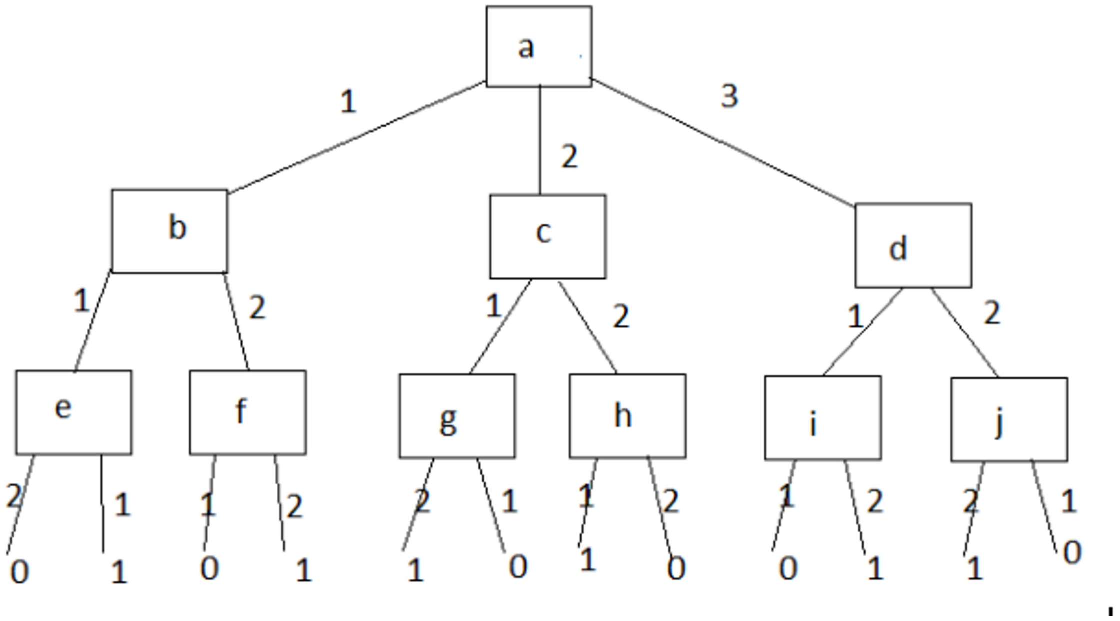
Q&A

01 March 2020 19:47

1.) Predict Classes
Consider the following tree:

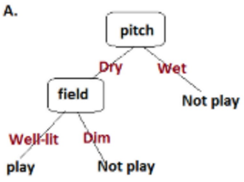


a, b, c, d, e, f, g, h, i, j are the attributes. A test on attribute 'a' can yield 3 possible values: 1, 2 and 3. Similarly, tests on other attributes can yield 2 values: 1 and 2. You have 2 class labels: 0 and 1.
Predict the class labels for the following cases:
Case1: a=2, c=2, h=2
Case2: a=1, b=2, f=1
Case3: a=3, d=1, i=2
0, 0, 1
Feedback :
For case1, from 'a' go to 'c' and then to 'h', following the lines corresponding to 'a=2' and 'c=2'. The class label associated with case1 will be at the end of the line, with 'h=2'. This will be similar for other cases as well.

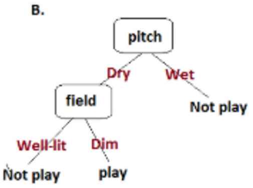
2.) True/False
An attribute can be present only in one test/node of a decision tree.
False
Feedback :*If you test on age, then you can have 'age < 3' as the first test, 'age < 21' as the second test, and so on.*

Comprehension - Interpreting a Decision Tree
Mithali plays cricket only when the pitch is dry, and the field is well lit. Based on past observations, you also know that she doesn't play cricket when either the pitch is wet, or the light is dim.
3.) Interpreting a decision tree
From the options given below, which decision tree would correctly predict whether Mithali will play or not?

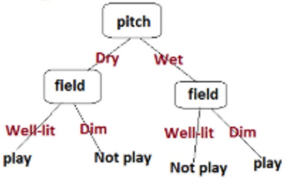
A.



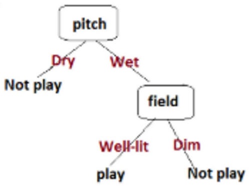
B.



C.



D.



A

Feedback :

Mithali plays cricket only when the pitch is dry, and the field is well lit. Otherwise, she doesn’t play.

4.)Interpreting a decision tree

Mithali didn’t play today. Identify the possible reason/s for this. More than one option may be correct.

The pitch was dry, and the lighting was dim.

Feedback : The lighting is dim. Hence Mithali wouldn't play.

Correct

The pitch was wet, and the field was well lit.

Feedback : The pitch is wet. Hence Mithali wouldn't play.

Correct

The pitch was wet, and the lighting was dim.

Feedback : The pitch is wet. Hence Mithali wouldn't play.

Correct

5.) Decision trees

Which of the following attributes are present on the left half of the tree? More than one option may be correct. (Note: The left half is the part along which Thal < 4.5 is true.)

Pain.type

Feedback :

On constructing the decision tree using the python code provided, you get a tree with pain.type attribute in the left half.

Correct

Flouroscopy.coloured

Feedback :

On constructing the decision tree using the code provided, you get a tree with flouroscopy .coloured attribute in the left half.

6.) Decision trees

Which of the following attributes is present in the right branch of the tree? More than one option may be correct. (Note: The right half is the part along which Thal < 4.5 is false.)

Decision trees

Which of the following attributes is present in the right branch of the tree? More than one option may be correct. (Note: The right half is the part along which Thal < 4.5 is false.)

Exercise.angina

Feedback :

On constructing the decision tree using the python code provided, you get angina with three attributes in the right half.

Correct

Flouroscopy.coloured

Feedback :

On constructing the decision tree using the python code provided, you get a tree with flouroscopy.coloured attribute in the right half.

7.) Decision trees

If the decision tree algorithm predicts that a person doesn’t have heart disease and ‘Thal’ < 4.5, which tests have been performed? More than one option may be correct.

Pain.type < 3.5

Feedback :
Since ‘Thal’ < 4.5, you go left on the tree. One of the tests that is performed, then, is on pain.typ.
CorrectYou missed this!
Flouroscopy.coloured < 0.5
Feedback :
Since ‘Thal’ < 4.5, you go left on the tree. One of the tests that is performed, then, is on flouroscopy.coloured.

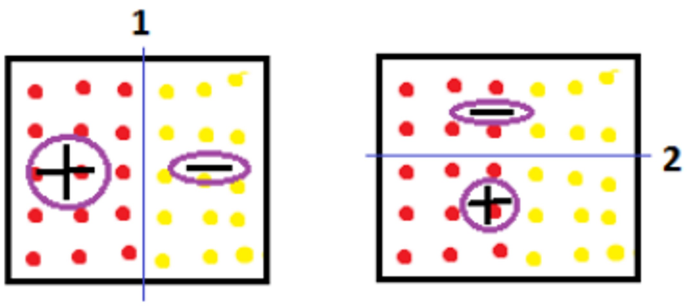
8.) Decision trees
How many leaves does the tree have?
8
Feedback :
Count the nodes that are not further splitting into branches.

9.) Decision trees
If the algorithm predicts that a person does not have heart disease, and it is known that she has ‘Thal’ > 4.5, then which of the following conditions has to be TRUE? More than one option may be correct.
‘Flouroscopy.coloured’ has to be less than 0.5.
Feedback :
Since 'Thal' is greater than 4.5, you go right. Then, the first test you encounter is on 'flouroscopy.coloured'.
Correct
‘Exercise.angina’ has to be less than 0.5.
Feedback :
Since 'Thal' is greater than 4.5, you go right. Then, the second test you encounter is on 'exercise.angina'.

10.) Regression
Select all that is correct about decision tree classification and regression models.
Leaves in classification contain labels.
Feedback : In classification, the target variable is discrete. Hence, each data point in a leaf has an associated class label.
Correct
Leaves in regression contain models.
Feedback : True. Each leaf in regression contains a model that is used to make predictions.

11.) Decision trees
Say you have a data set with lots of categorical variables and some numerical variables. The target variable is continuous, so it’s a regression problem. After some exploratory data analysis, you figure out that it will be best to perform decision tree regression instead of linear regression. Which of the following statements will be correct? More than one option may be correct.
It is hard to represent all the data via a single model; so you don’t want to use the linear regression model.
Feedback : Decision tree regression is performed because the entire data set cannot be represented by a single linear regression model.
Correct
Decision trees will split the data set into multiple sets and will apply linear regression to each set separately.
Feedback : Leaves obtained after splitting contain linear regression models to make predictions.

12.) Split dataset
Red dot => label = 1, yellow dot => label = 0



Given the scatter plots, which line would you choose to split the data such that a minimum number of data points is misclassified after splitting?

Split by line1 such that after the split, one partition has all the data points belonging to label '1', and the other partition has data points belonging to label '0'.

Feedback :

All labels are identical in the left half. Similarly, all labels belong to the same class in the right half of line 1. All the points are correctly classified in the first image, while the second image has a lot of misclassified points.

13.) Homogeneity

Given a dataset with two attributes, age and gender, you want to predict whether a person will purchase a product (yes/no). You know that among all those who have purchased the product in the past, 98% are females. Also, the age distribution of the customers is almost uniform. The homogeneity of the resultant nodes will be maximised if you split on:

Gender

Feedback :*Correct. 'Gender' will split the data such that one node will contain 98% of total observations that belong to 'product purchased' class and the other node will contain 2% of total observations belonging to 'product purchased' class. So, nodes will have high homogeneity here.*

14.) Homogeneity

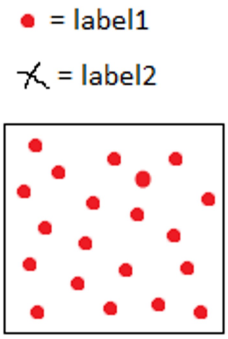
How do you identify a 'completely homogeneous' data set?

100% of the data points belong to one class label.

Feedback :*Since all the data points belong to only one label, the data set is completely homogeneous.*

15.) Homogeneity

What is the homogeneity of the following data set?



Here, d1, d2, d3, d4, d5, d6, and d7 represent the data points; a1, a2, a3, and a4 are the attributes.

Completely homogeneous

Feedback :

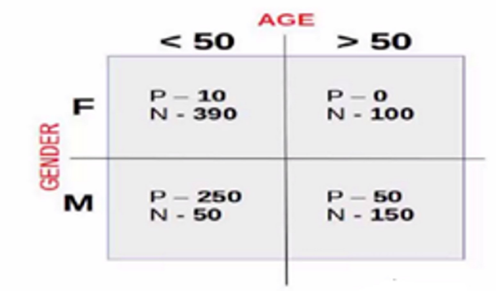
All the data points belong to only one class label, and so this data set is completely homogeneous.

16.) Homogeneity
The ultimate aim of decision tree splitting is to _____.
Increase homogeneity
Feedback :*More homogeneity will mean that most of the data points in the set belong to the same class label. Hence, classifying all the data points of that set, to get them to belong to that class, will result in lesser errors.*

17.) Splitting
Out of so many attributes, how does a decision tree select one for splitting? Select the best option.
It calculates the improvement in homogeneity associated with each attribute and picks the one that results in the maximum increase in homogeneity.
Feedback :*Out of all the attributes, the attribute that results in the maximum increase in homogeneity is chosen for splitting.*

18.) Consider the dataset of female and male employees of the company shown in the lecture. There are 1000 employees in total. Let's say, you split the data on gender. On splitting, you get two nodes with 500 observations each, call them male-node and female-node. In the male-node, 300/500 play football. In the female node, only 10/500 play football. Gini index of male-node = P(play football)^2 + P(doesn't play football)^2 = (0.6)^2 + (0.4)^2 = 0.52. What is the Gini index of female-node? (Note: The Gini index = ∑ki=1p2i, where pi is the probability of finding a point with the label *i*, and *k* is the number of different labels.)
(0.02)^2 + (0.98)^2
Feedback :*The probability that a female plays football = 0.02 and the probability that a female doesn't play football = 0.98*

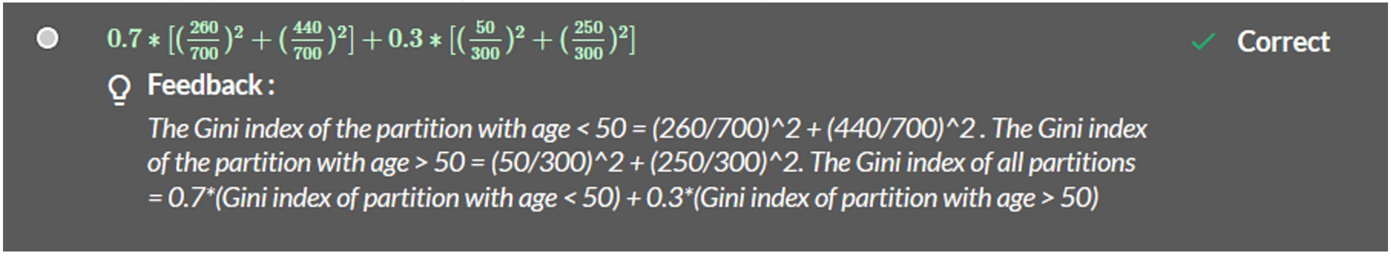
19.) Split on gender
What is the Gini index of the partitions if you split on ‘age’?



AGE

GENDER	< 50	> 50
F	P – 10 N - 390	P – 0 N - 100
M	P – 250 N - 50	P – 50 N - 150

To recall, there are 1000 employees in total, out of which 500 are females and 500 are males.Out of these 1000 employees, 700 are below age 50 and 300 are above age 50.



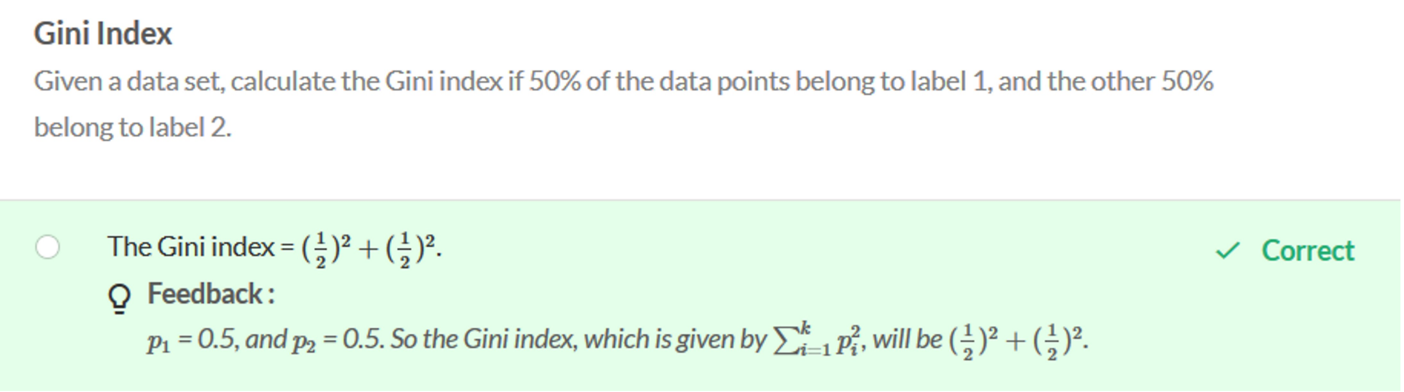
○ $0.7 * [(\frac{260}{700})^2 + (\frac{440}{700})^2] + 0.3 * [(\frac{50}{300})^2 + (\frac{250}{300})^2]$ ✓ Correct

Feedback :

The Gini index of the partition with age < 50 = (260/700)^2 + (440/700)^2 . The Gini index of the partition with age > 50 = (50/300)^2 + (250/300)^2. The Gini index of all partitions = 0.7(Gini index of partition with age < 50) + 0.3*(Gini index of partition with age > 50)*

20.) Gini Index
What is the Gini index if all the data points in a data set have the same label?
Gini Index = 1
Feedback :*The probability of exactly one class will be 1, and the probability of all the other classes will be 0. So, the Gini index, which is given by ∑ki=1p2i, will be 1.*

21.)



Gini Index

Given a data set, calculate the Gini index if 50% of the data points belong to label 1, and the other 50% belong to label 2.

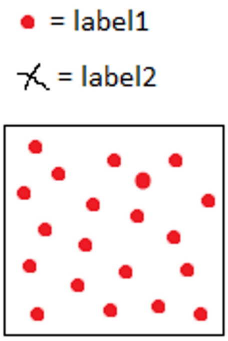
○ The Gini index = $(\frac{1}{2})^2 + (\frac{1}{2})^2$. ✓ Correct

Feedback :

$p_1 = 0.5$, and $p_2 = 0.5$. So the Gini index, which is given by $\sum_{i=1}^{k} p_i^2$, will be $(\frac{1}{2})^2 + (\frac{1}{2})^2$.

22.) Gini Index
When is the Gini index maximum?
When the homogeneity is maximum.
Feedback :*When all the points in the data set belong to one class label leading to the maximum homogeneity, the Gini index will be maximum.*

23.) Calculate p1, p2 and the entropy of the following data set. Please note that 0*log(0) = 0.



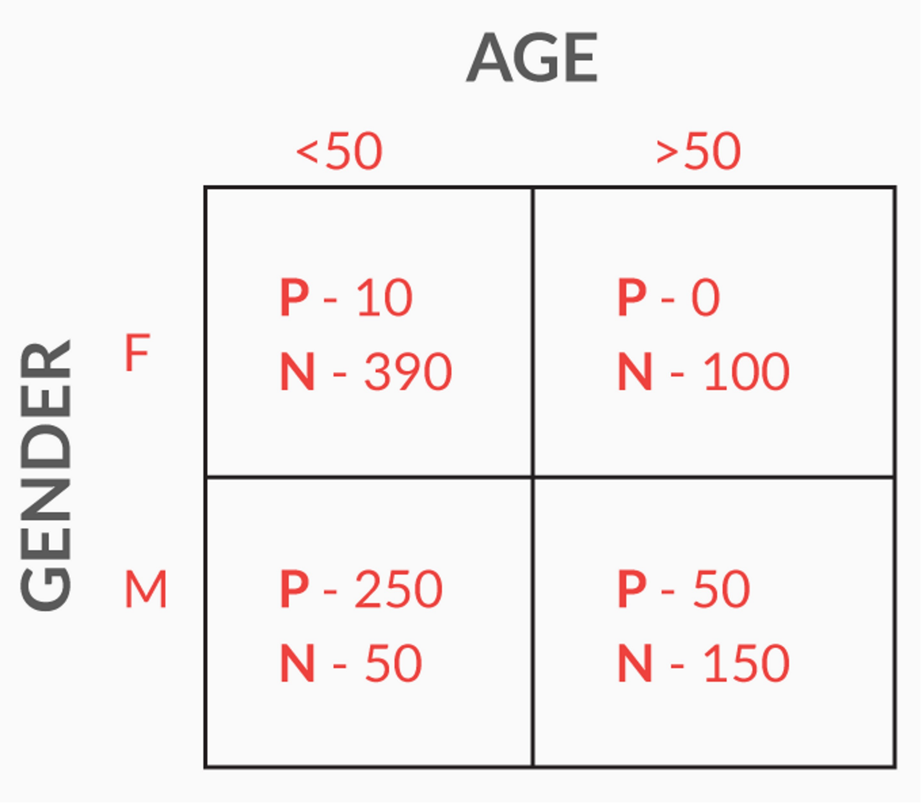
p1 = 1, p2 = 0, entropy = 0
Feedback :
*All the data points belong to exactly one class. So entropy = 1*log(1) + 0*log(0) = 0.*

24.) Entropy
How is entropy related to the Gini index?
The lower the entropy, the higher the Gini index.
Feedback :*The entropy is a measure of disorderness, while the Gini index is a measure of homogeneity in the data set. The lesser the disorder, the lower the entropy and the greater the homogeneity; and hence, the Gini index is higher.*

25.) Information Gain
When is the information gain maximum? (Select the most appropriate option.)
When the decrease in entropy, from the parent set to the partitions obtained after splitting, is maximum.
Feedback :*The information gain is equal to the entropy change from the parent set to the partitions. So it is maximum when the entropy of the parent set minus the entropy of the partitions is maximum.*

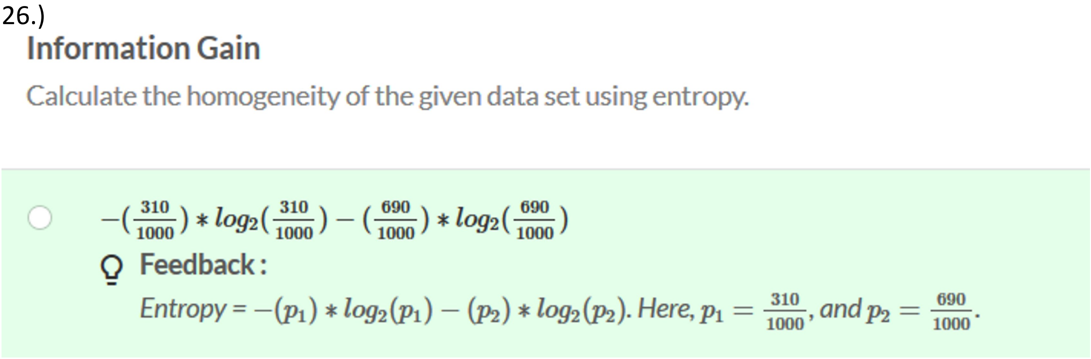
Details of the data set: There are 1000 employees in an organisation, of which 500 are females and 500 are males. The number of employees below 50 years of age is 700 and above 50 years is 300. Given the current data set of these 1000 employees, you want to predict whether in the future, a given employee will play football or not. Here,
P implies 'plays football' - class A = **label 1.**
N implies 'does not play football' - class B = **label 2**.

A total of 10 females and 300 males play football in the organisation. 260 people who are less than 50 and 50 people above the age of 50 play football, as shown in the figure above.



		AGE	
		<50	>50
GENDER	F	P - 10 N - 390	P - 0 N - 100
	M	P - 250 N - 50	P - 50 N - 150

26.)



Information Gain

Calculate the homogeneity of the given data set using entropy.

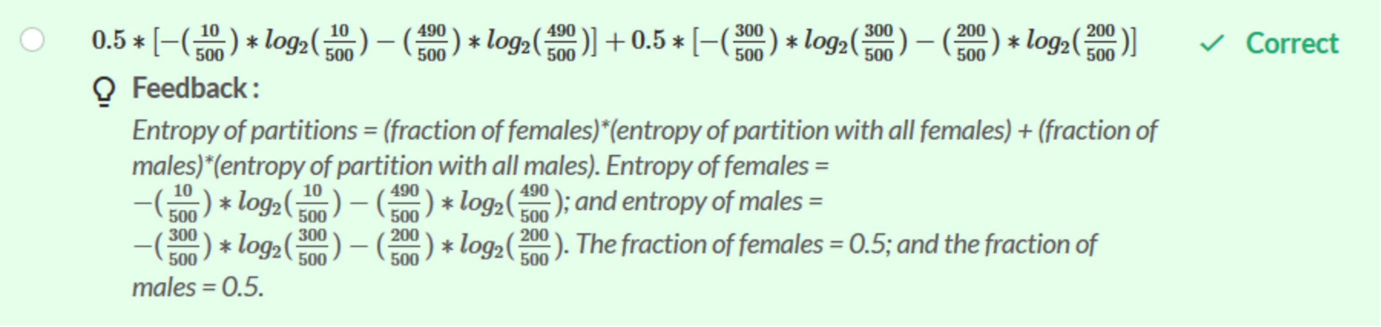
○ $-(\frac{310}{1000})*log_2(\frac{310}{1000})-(\frac{690}{1000})*log_2(\frac{690}{1000})$

Feedback :

Entropy = $-(p_1)*log_2(p_1)-(p_2)*log_2(p_2)$*. Here,* $p_1=\frac{310}{1000}$*, and* $p_2=\frac{690}{1000}$*.*

27.) Information Gain

What is the entropy of the partitions if you split on ‘gender’?



○ $0.5*[-(\frac{10}{500})*log_2(\frac{10}{500})-(\frac{490}{500})*log_2(\frac{490}{500})]+0.5*[-(\frac{300}{500})*log_2(\frac{300}{500})-(\frac{200}{500})*log_2(\frac{200}{500})]$ ✓ Correct

Feedback :

Entropy of partitions = (fraction of females)(entropy of partition with all females) + (fraction of males)*(entropy of partition with all males). Entropy of females =* $-(\frac{10}{500})*log_2(\frac{10}{500})-(\frac{490}{500})*log_2(\frac{490}{500})$*; and entropy of males =* $-(\frac{300}{500})*log_2(\frac{300}{500})-(\frac{200}{500})*log_2(\frac{200}{500})$*. The fraction of females = 0.5; and the fraction of males = 0.5.*

28.) Information Gain

What is the information gain if you split the original data set on ‘gender’?

Entropy of original data set - entropy of partitions obtained after splitting on ‘gender’ = 0.89317 - 0.5562.

Feedback :*Calculate the entropy of the original data set, and subtract the entropy of the partitions obtained after splitting, to get the information gain.*

29.) Information Gain

The information gain, if you split on ‘age’, is 0.031929, and it is 0.33697 if you split on ‘gender’. What attribute should you split the original data on?

Gender

Feedback :*You split on the attribute that maximises the information gain. The information gain on gender is greater than the information gain on age.*

30.)Decision Tree - Regression

You stop splitting further if the R2 is

High enough

Feedback :*The higher the R2, the better the regression model.*

31.) Regression

Which homogeneity measure is used in tree regression?

R2

Feedback :*R2 is used to measure the homogeneity in regression, where the target variable is continuous.*

32.) Regression

Select all that apply. One or more option may be correct.

Leaves in decision tree regression contain models.

Feedback : Each leaf in regression contains a model that is used for prediction.

Correct

Further split the data if the R2 measure is high enough.

Feedback : R2 is a measure of how close the data is to the fitted regression line.

Correct

Sometimes a single linear regression model is not good enough to perform the regression task, so you split the data into smaller chunks and assign one linear regression model to each chunk.

Feedback : As you saw in the lecture, people in different age brackets should be represented using different models.

Correct

Decision tree regression and classification are similar in the sense that both try to pick an attribute (for splitting) that maximises the homogeneity of the data set.

Feedback : A decision tree splits the data set on that attribute which results in the maximum increase in homogeneity.

33.) Regression

Let's go through the steps involved in decision tree construction. Arrange the following steps in the order of their occurrence:

1. Now that the R2 is sufficiently high, stop splitting.
2. You have a data set, D, with categorical as well as numerical attributes and continuous target variables. So it is a regression problem. Hence, you apply decision tree regression to it.
3. Split the original data set, D, on the selected attribute.
4. After selecting the homogeneity measure, you need to decide the first attribute to split the original data set, D, on.
5. Keep on splitting the subsequent data sets till you get a sufficiently high R2.
6. Select a homogeneity measure for splitting. Since it is regression, let's choose R2.
7. Each leaf will now represent a linear regression model.
8. Split 'D' on all the attributes one by one, and select the attribute that results in the maximum increase in homogeneity after splitting.

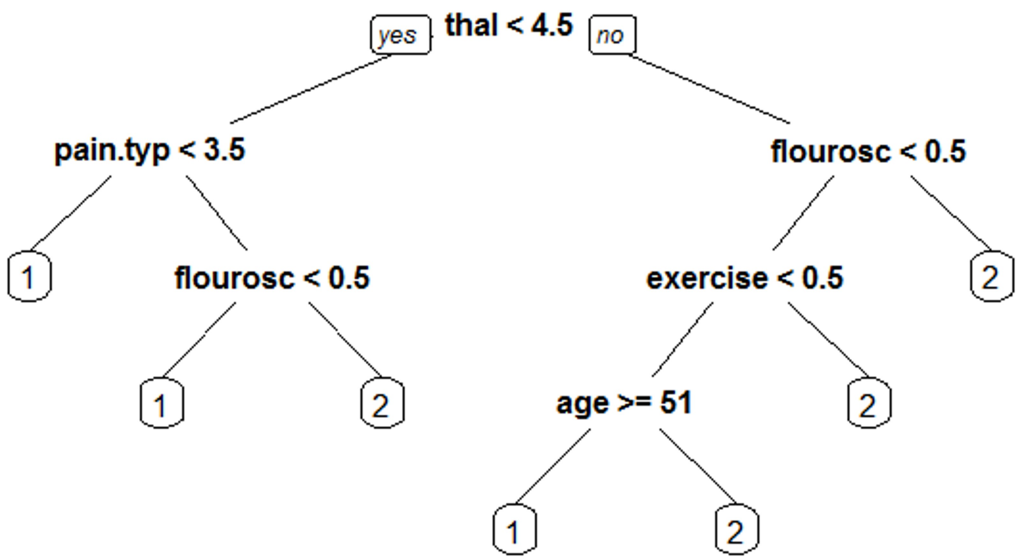
2, 6, 4, 8, 3, 5, 1, 7

Feedback :*First, decide if it's a classification problem or a regression problem. Then, select a homogeneity measure for splitting, and select the first attribute for splitting out of the many. After this, split the original data set on the selected attribute, and keep splitting till you get a sufficiently high R2. Once you stop splitting, you will get leaves containing linear regression models.*

34.)

Decision Tree

Consider the following tree.



Which is the most informative feature, as identified by the tree?

Thal

Feedback :

The most informative features are towards the top of a tree.

35.) Decision Tree

Suppose you are getting an accuracy of 40% on the test data and 98% on the training data. Select all of the following options that apply. (More than one may be correct.)

The model has a low variance and high bias.

The model has a high variance and a low bias.

Feedback : The model has memorised the data, giving you a 98% training accuracy and leading to a high variance. Since it can now represent the training set very well, it has a low bias.

Correct

The model is overfitting.

Feedback : The test accuracy is very low (40%). The model is unable to work well on unseen/test data. It has memorised the training set. Hence, it is overfitting.

36.) Comprehension - Truncation and Pruning

The process of splitting only when there is a sufficient number of data points in the node is called _______.

Truncation

Feedback :

Truncation lets you split the data only when the number of data points in the node is greater than or equal to the 'minsplit'.

37.) Comprehension - Truncation and Pruning

Which hyperparameter controls the minimum no. of samples required to split an internal node?

min_samples_split

Feedback :

The min_samples_split specifies the minimum number of data points a node should have for it to be considered for splitting.

38.) Comprehension - Truncation and Pruning

__________ takes care of the minimum number of samples required to be at a leaf node.

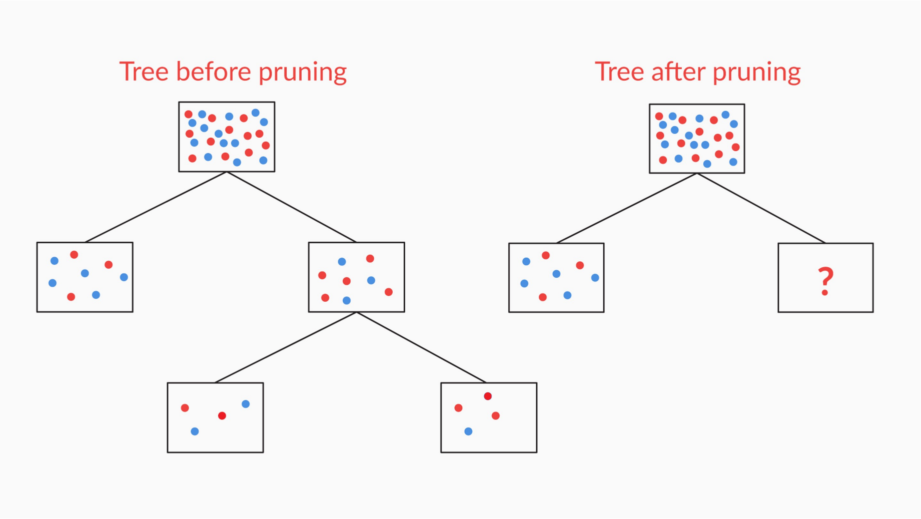
min_samples_leaf

Feedback :

The minimum number of samples required to be at a leaf node. If an integer value is taken then consider - -min_samples_leaf as the minimum no. If float, then it shows percentage. By default, it takes “1” value.

39.) Comprehension - Truncation and Pruning
Assume that you have set the min_samples_leaf as 3 and the min_samples_split as 6. Consider a node with 10 data points. On splitting on an attribute, one leaf gets 2 points, and the other one gets 8 data points. This split will not be executed. Why?
The number of data points in one of the leaves < min_samples_leaf.
Feedback :
The number of data points in one of the leaves is 2, which violates the condition that the number of data points in a leaf should be at least 3 (as specified by the min_samples_leaf)

Questions
Consider the tree 'before' pruning and the tree 'after' pruning. The red dots belong to the class 'Label 1'. The blue dots belong to the class 'Label 2'.



Pruning
Consider the tree on the left: 'Tree before pruning'. You decided to prune the bottom two branches such that you get the tree on the right: 'Tree after pruning'.
40.)Pruning
How many observations/data points will the new leaf have?
8
Feedback :*The observations in the leaf will be the same as before. Say, you have 4 observations in the node and you split it such that the left partition has 1 observation and the right partition has 3 observations. Now, if you chop these partitions off, the original node which is now a leaf, will still have 4 observations.*

41.) Pruning
Which label will be assigned to the new leaf?
Label 1
Feedback :*Since the number of points that belong to label 1 is greater than the number of points that belong to label 2, in the leaf, label 1 will be assigned to it.*

42.) Pruning
If the accuracy after pruning on the unseen data decreases significantly, then
You shouldn't prune the branch.
Feedback :*You should prune only if the accuracy after pruning does not decrease.*

43.) Decision Tree Hyperparameters
State whether true or false - the parameter min_samples_split specifies the minimum number of data points required in a node to be considered for further splitting.
True
Feedback :
Yes, as mentioned in the documentation, min_samples_split is the minimum number of data points required in a node to be considered for further splitting.

44.) Decision Tree Hyperparameters
Choose the correct option: As you increase the value of the hyperparameter min_samples_leaf, the

resulting tree will:
become less complex, and the depth will tend to reduce
Feedback :
Correct - min_samples_leaf is the minimum number of samples required in a (resulting) leaf for the split to happen. Thus, if you specify a high value of min_samples_leaf, the tree will be forced to stop splitting quite early.

45.) Min_sample_leaf
What will be the effect on the depth of the tree if min_sample_leaf is set to 1? Will the tree be overfitting the train data or the test data?
Suggested Answer
Yes, if the min_sample_leaf is set to 1 the tree will overfit the data as the number of the samples required at the leaf node can be 1.

46.) Difference
What is the difference between min_sample_split and min_sample_leaf?
Suggested Answer
min_sample_split tells above the minimum no. of samples reqd. to split an internal node. If an integer value is taken then consider min_samples_split as the minimum no. If float, then it shows percentage. By default, it takes “2” value.
min_sample_leaf is the minimum number of samples required to be at a leaf node. If an integer value is taken then consider - -min_samples_leaf as the minimum no. If float, then it shows percentage. By default, it takes “1” value.

Comprehension - Hyperparameters

Consider a decision tree classification model that has a very high training accuracy and a low test accuracy, i.e. the model has a high variance. The training accuracy is 98%, and the test accuracy is 55%. The ‘min_samples_split’ for this model is 5, and the ‘max_depth’ is 20.
47.) Hyperparameters
What does the min_samples_split = 5 imply? More than one option may be correct.
The minimum no. of samples required to split an internal node is equal to 5.
Feedback :
The hyperparameter **min_samples_split** is the minimum no. of samples required to split an internal node. Its default value is 2, which means that even if a node is having 2 samples it can be further divided into leaf nodes.
Even if a node is having 5 samples it can be further divided into leaf nodes.
Feedback :
The hyperparameter **min_samples_split** is the minimum no. of samples required to split an internal node. Its default value is 2, which means that even if a node is having 2 samples it can be further divided into leaf nodes.

48.)Hyperparameters
Select all that apply. (More than one option may be correct.)
min_samples_split = 5 implies that the node should have at least five data points for splitting.
Feedback :
min_samples_split = 5 indicates that splitting will not be performed if the number of data points in the node is less than 5. The min_samples_split specifies the minimum number of data points a node should have for splitting to be attempted.
The min_samples_split sets a lower bar on the number of data points a node should have.
Feedback :
The min_samples_split specifies the minimum number of data points a node should have for splitting to be attempted.

49.) Hyperparameters
Suppose you decide to tweak the hyperparameters so as to decrease the variance/overfitting. Which of the following steps will help? More than one option may be correct.
Increasing min_samples_split.
Feedback :

A low value of the min_samples_split will lead to a small number of data points in the nodes. This means that it is very likely that each leaf (obtained after splitting) is going to represent very few (or only one, in some cases) data points. So, you increase the min_samples_split.

Decreasing max_depth

Feedback :

Decreasing max_depth will stop the tree to grow deeper, in that way your tree will not overfit the data and you will have a decent accuracy in both test and train.

50.) Hyperparameters

What will the (likely) impact of increasing the value of min_sample_splits from 5 to 10?

The depth will decrease.

Feedback :

Since the node should now contain at least 10 data points before splitting, as opposed to 1, all the branches — where the nodes had less than 10 data points — will be chopped off, leading to a decrease in the tree depth.

51.) Random Models

If we consider a binary classification task, then better than a random model implies that the model under consideration

Makes correct predictions with a probability statistically better than that of a random model, i.e. 0.5

Feedback :*Acceptability means that a model is at least not making random guesses, whose P(success) is 0.50. Thus, we want models whose probability of success is > 50%.*

52.)Diversity

The diversity of models in an ensemble implies that

If two models give the same answers on a random data, it will be totally coincidental

Feedback :*Diversity represents independence, i.e. models are not correlated (and do not get influenced by) other models. This means that the answers (predictions) given by two models are independent of each other (you'll study how this is achieved in a short while).*

53.) Coin Toss Analogy

Which of the following assumptions make it possible to use the coin toss analogy? Mark all that apply

Each toss is independent of each other and so are the models in an ensemble

Feedback : We have assumed that 1) the models are independent and 2) they are all individually acceptable. The coin, with success mapped to heads, is biased towards heads. Thus, P(heads > 0.5) and since all models are acceptable, P(success) > 0.5.

Correct

The coin is biased towards and favors heads, and each model is acceptable

Feedback : We have assumed that 1) the models are independent and 2) they are all individually acceptable. The coin, with success mapped to heads, is biased towards heads. Thus, P(heads > 0.5) and since all models are acceptable, P(success) > 0.5.

54.) Biased Coin

We compare an ensemble to the toss of a biased coin with heads being favored, or P(heads) > 0.5. This is equivalent to saying that

Each model is acceptable

Feedback :*We have defined the notion of acceptability in terms of P(success/correct answer) > 0.5. Heads is equivalent to success.*

55.) Binary Classification

Consider a simple ensemble with 2 models - m1 and m2. To make a decision (binary classification), we take the majority vote of m1 and m2. Let's denote the probability of correct prediction, i.e. P(correct prediction) of each model by P(mi = correct). For this to be an ensemble, which of the following should hold true (mark all that apply)

P(m1 = correct) > 0.5

Feedback : For a model to be acceptable, the probability of it being correct should be more than 0.5 i.e. better than a random guess.

Correct

P(m2 = correct) > 0.5
Feedback : For a model to be acceptable, the probability of it being correct should be more than 0.5 i.e. better than a random guess.

56.) Random Feature Selection in Random Forests
Consider the heart disease data set where a few attributes such as Thal, blood pressure, etc., are prominent predictors for the target variable. If you were to build multiple decision trees on this as a part of an ensemble, considering all the attributes for all the individual trees, which of these violations would occur and be significant?
The models will not be diverse enough
Feedback :*If a few variables are prominent, a large number of trees will have them as important nodes, and they will look similar. Similar trees violate the condition of diversity.*

57.) Building a Random Forest
When do observation and feature sampling take place for trees inside a random forest? More than one option may be correct.
A random subset of observations is chosen every time a new tree is built in a forest.
Feedback : A different random subset of observations is chosen, which is called the bootstrap sample, for each tree that is to be built in the forest. This is called bootstrapping.
Correct
A random subset of features is chosen every time a node is being split inside a tree.
Feedback : After the bootstrap sample is selected, tree building starts, and a random subset of features is selected at each node in order to split it. This is what makes random forests even better than a simple bagging algorithm.

58.) Random Forest vs Bagging
How is a random forest different from bagging?
In a random forest, a random sample of features is chosen at each node split, which does not happen in bagging.
Feedback :*Bagging includes the creation of different bootstrap samples for different models, and aggregating the results of the models. Random forests use this technique along with randomly selecting features at each node while splitting it.*

59.) Aggregation in Random Forests
During bagging, or bootstrap aggregation, the test data point is passed through all the trees of the forest, and each tree makes its own prediction. How are these predictions aggregated in the case of a regression problem?
Mean
Feedback :*The final prediction is the mean of all the predictions of the individual decision trees.*

60.) Random Forests
Mark all the correct statements
Bootstrapping implies that each tree in a RF is built on randomly chosen observations
Feedback : The word 'random' in random forests refers to the random choice of bootstrapped observations
Correct
While considering split at a node, a random set of the attributes is considered
Feedback : Random choice of attributes at each split of a tree
Correct
Random choice of attributes while splitting at nodes ensures diversity in a random forest
Feedback : Random choice of attributes ensures that the prominent features do not appear in every tree, thus ensuring diversity.

61.) Bagging
The core idea behind bagging is that of a **majority score** rather than committing to set of assumptions made by a single model. The idea is particularly successful in random forests because trees:
Are typically unstable
Feedback :*If you have only one tree, you have to rely on the decision it makes. The decision a single*

tree makes (on unseen data) depend highly on the training data since trees are unstable. In a forest, even if a few trees are unstable, averaging out their decisions ensures that you are not making mistakes because of a few trees' unstable behaviour.

62.) Random Forests vs Decision Trees
In terms of accuracy, is a random forest always better than a decision tree?
False
Feedback :*While it is well known that random forests are better, in terms of accuracy, than a single decision tree, it cannot be said that they are better than every possible decision tree. It is just more difficult to build a decision tree that is better than a random forest. In fact, there may be several trees that provide better predictions on unseen data.*

63.) Variance in Random Forests
Which of the following statements is true?
A larger number of trees will result in a lower variance of the ensemble.
Feedback :*Variance means how much a model (ensemble here) changes with changes in the training data. If a large number of trees is at work, then even if some of them show a high instability (extreme variation in the trees and their predictions), the ensemble as a whole will reduce the variance by averaging out the results of each tree.*

64.) OOB Error
Which of the following dataset is used to calculate the OOB error?
Training set
Feedback :*Only the training set is used while calculating the OOB error, which is why it gives a good idea of model performance on the unseen data without using a test set.*

65.) OOB Error
Which of the following statements is true?
All the observations of the training set are used to calculate the OOB error.
Feedback :*Recall that all the observations of the training set are used to calculate the OOB error.*

66.) Time Required to Build a Forest
If there are S trees in a forest, M features (income, age etc.) and n observations (in the original training data), the time taken to build the forest depends on:
S, M and n
Feedback :*The time required will obviously depend on S. While building each of the S trees, time is spent in creating the levels of trees and time required to find splits among f features. Levels of trees are given by log(n). Finding the right split depends on both n observations and f features because homogeneity will be measured for all f features and n observations.*

67.) Time Spent on Splitting
Consider building a single individual tree in an ensemble by taking j = 40% observations randomly from the training set. There are M features and n observations in the original training data. The **time spent at each split** in this tree is proportional to:
sqrt(M).n.j
Feedback :*Each split is made by comparing the homogeneity across j= 40% of the n observations. Thus, it has to depend on j and n (more the observations, more the time required to compare homogeneity). The time required to find a split also depends upon the number of features being considered which is sqrt(M).*