Question bank on decision tree algorithm

Presentation · May 2023

DOI: 10.13140/RG.2.2.12228.09606

Abhishek D. Patange
ABB
115 PUBLICATIONS **753** CITATIONS

DECISION TREE – ENTROPY & INFORMATION GAIN

Problems for 2-4 marks

Problems 1

A decision tree classifier is used to predict the failure mode of a boiler system with a dataset of 300 samples, where 200 samples indicate normal operation and 100 samples indicate tube leakage. What is the entropy of the classifier?

Problems 2

If a decision tree classifier is used to predict the type of heat exchanger fouling in a system with a dataset of 180 samples, where 100 samples indicate deposition and 80 samples indicate scaling, what is the entropy of the classifier?

Problems 3

We want to predict the type of tire wear in a vehicle with a dataset of 120 samples, where 50 samples indicate wear due to underinflation, 30 samples indicate wear due to overinflation, and 40 samples indicate wear due to misalignment, what is the entropy of the classifier?

Problems 4

A dataset of 250 samples were collected to predict the type of suspension failure in a vehicle, where 100 samples indicate failure due to worn-out shock absorbers, 80 samples indicate failure due to broken springs, and 70 samples indicate failure due to worn-out bushings. What is the entropy of the classifier?

Problems 5

We need to split a set of energy efficiency data on the attribute "glazing area distribution" with two possible values (0.4 and 0.7) and the information gain is 0.214. What is the entropy of the original set?

Problems 6

The information gain for splitting a set of solar panel efficiency data on the attribute "angle of incidence" with three possible values (0°, 45°, and 90°) is 0.325. If the entropy of the original set is 0.95, what is the size of the original set?

Problems 7

A decision tree classifier is trained on a set of geothermal energy data with two attributes: "temperature" and "depth." The entropy of the original set is 0.81, and the information gain for splitting on the "depth" attribute is 0.251. What is the information gain for splitting on the "temperature" attribute?

Problems 8

Suppose a manufacturer of mechanical components is testing a new production process that produces components with two possible defects: crack or deformation. The manufacturer has collected data on 500 components produced by the new process, and has labeled each component as either defect-free, cracked, or deformed. The data is summarized in the following table:

Component status	Defect-free	Cracked	Deformed
Number of components	350	100	50

What is the entropy of the dataset?

Problems 9

A car manufacturer is testing a new engine design that can produce three types of failures: overheating, oil leakage, and poor fuel efficiency. A data has been collected on 500 engines produced by the new design, and has labeled each engine as either failure-free, overheating, oil leakage, or poor fuel efficiency. The data is summarized in the following table:

Engine status	Number of engines
Failure-free	350
Overheating	100
Oil leakage	30
Poor fuel efficiency	20

What is the entropy of the dataset?

Problems for 6-8 marks

Problems 10

A training dataset collected while navigating autonomous vehicle is shown below.

Speed	Distance to intersection	Traffic Control Device	Pedestrian Presence	Decision
0.5	0.7	Stop sign	Pedestrian crossing	Slow down
0.2	0.9	Green traffic light	No pedestrian crossing	Stop
0.8	0.4	Red traffic light	No pedestrian crossing	Go
0.6	0.1	Pedestrian crossing	No traffic light	Slow down
0.3	0.5	No traffic sign	Pedestrian crossing	Go
0.9	0.8	No traffic sign	No pedestrian crossing	Stop
0.7	0.2	Green traffic light	No pedestrian crossing	Stop
0.4	0.6	Stop sign	No pedestrian crossing	Slow down
0.1	0.3	No traffic sign	No pedestrian crossing	Stop
0.5	0.5	Red traffic light	Pedestrian crossing	Go

Identify discrete & continuous attributes along with their features. Calculate information gain of two discrete attributes individually.

Problems 11

A decision tree classifier is to be trained to predict the likelihood of a vehicle rolling over during a sharp turn based on input features: vehicle speed during the turn (mph), angle of the turn (degrees), type of vehicle (e.g., sedan, SUV, pickup truck), weight (Ibs), road surface conditions (e.g., dry, wet, icy), tire tread, driver behavior (e.g., aggressive, cautious).

- Calculate information gain of attribute ‘Vehicle Type’ & ‘Road Surface’.
- Compare them and comment on which one of these two is suitable for the best split.

Speed (mph)	Turn Angle	Vehicle Type	Weight (lbs)	Road Surface	Tire Tread	Driver Behavior	Rollover
35	50°	SUV	4000	Wet	Good	Aggressive	Yes
25	30°	Sedan	3200	Dry	Good	Cautious	No
40	70°	Pickup	5000	Icy	Poor	Aggressive	Yes
30	45°	SUV	4500	Wet	Good	Cautious	No
50	80°	Pickup	5500	Dry	Good	Aggressive	Yes
20	20°	Sedan	2800	Dry	Good	Cautious	No
30	60°	SUV	4200	Wet	Good	Aggressive	Yes
40	75°	Pickup	5100	Icy	Poor	Cautious	Yes
35	55°	SUV	4100	Wet	Good	Cautious	No
45	90°	Pickup	5800	Dry	Good	Aggressive	Yes

Problems 12

Consider following dataset

Roll	Pitch	Gyro Mode	Gyro Speed	Label
20.5	45.2	Stabilization mode	Low speed	Roll right
-15	60	Acrobatic mode	High speed	Roll left
10	30	Follow mode	Low speed	Pitch up
0	0	Stabilization mode	High speed	Pitch down
45	80	Acrobatic mode	Low speed	Roll right
-30	-60	Follow mode	High speed	Roll left
-10	-20	Stabilization mode	Low speed	Pitch down
60	-45	Acrobatic mode	High speed	Pitch up
-80	0	Follow mode	Low speed	Roll left
35	-70	Stabilization mode	High speed	Pitch down

Calculate entropy of dataset before splitting any attribute and after splitting Gyro Mode & Gyro Speed.

Problems 13

Calculate entropy of dataset before splitting any attribute and after air filter type.

Room Size (sqm)	Number of Occupants	Time of Day (hours)	Air Filter Type	Label
20	2	14	Basic Filter	Comfortable
25	4	12	HEPA Filter	Comfortable
30	6	16	Basic Filter	Too Hot
18	3	10	Carbon Filter	Comfortable
15	1	18	HEPA Filter	Too Cold
28	5	13	Carbon Filter	Comfortable
22	2	15	Basic Filter	Comfortable
16	2	11	HEPA Filter	Too Hot
21	4	14	Carbon Filter	Comfortable
24	3	17	Basic Filter	Too Cold

Problems 14

A government agency wants to investigate the factors that contribute to unsafe braking in cars equipped with an ABS system. They have collected data on vehicle speed, road surface type, brake pedal pressure, and ABS activation type for a set of cars involved in accidents where the ABS system failed to prevent unsafe braking. Built a decision tree based on attribute ‘ABS Activation Type’ only.

Vehicle Speed (km/h)	Road Surface Type	Brake Pedal Pressure (%)	Activation Type	Label
60	Dry	50	Active	Safe
80	Wet	60	Passive	Unsafe
70	Dry	70	Active	Safe
50	Wet	30	Passive	Unsafe
100	Dry	80	Active	Safe
40	Wet	40	Passive	Unsafe
90	Dry	90	Active	Safe
65	Wet	20	Passive	Unsafe
75	Dry	60	Active	Safe
55	Wet	50	Passive	Unsafe

Problems 15

Consider following dataset

Road Condition	Suspension Type	Speed (mph)	Temperature (°F)	Performance
Wet	Independent	50	70	Good
Dry	Dependent	70	80	Good
Wet	Independent	30	60	Poor
Dry	Dependent	60	90	Good
Wet	Independent	40	75	Poor
Dry	Dependent	80	85	Good
Wet	Independent	20	50	Poor
Dry	Dependent	50	70	Poor
Wet	Independent	60	80	Good
Dry	Dependent	40	65	Poor

In order to create a decision tree that accurately predicts the suspension system's performance, which attribute is the most important amongst Road Condition and Suspension Type?

Problems 16

Predict the Powertrain Output based on the Road Condition and Traction Control attributes.

Engine Speed (rpm)	Throttle Position (%)	Road Condition	Traction Control	Powertrain Output
2000	30	Dry	Off	Low
2500	50	Wet	On	Low
3500	80	Dry	Off	Medium
4000	90	Snowy	On	Low
3000	70	Wet	On	Medium
1500	20	Dry	Off	Low
4000	100	Dry	Off	High
2000	40	Snowy	On	Low
3000	60	Wet	Off	Medium
3500	80	Dry	On	High

Problems 17

Given the dataset, can you classify TATA car models as "Good", "Average", or "Poor" based on their overall performance? What features are the most important for making this classification?

Model	Transmission	Fuel Type	Drivetrain	Car Type	Outcome
Tiago	Manual	Petrol	Front-wheel drive	Hatchback	Good
Tigor	Manual	Petrol	Front-wheel drive	Sedan	Average
Nexon	Automatic	Diesel	All-wheel drive	SUV	Good
Harrier	Automatic	Diesel	Front-wheel drive	SUV	Good
Altroz	Manual	Petrol	Front-wheel drive	Hatchback	Good
Safari	Automatic	Diesel	All-wheel drive	SUV	Average
Hexa	Automatic	Diesel	All-wheel drive	SUV	Average
Bolt	Manual	Petrol	Front-wheel drive	Hatchback	Average
Zest	Manual	Diesel	Front-wheel drive	Sedan	Average
Sumo	Manual	Diesel	Rear-wheel drive	SUV	Poor

Problems 18

Calculate the entropy of the dataset for each attribute (altitude, wind, temperature, and humidity) and determine which attribute is the best choice for the root node of the decision tree.

Altitude	Wind	Temperature	Humidity	Outcome
High	Low	Hot	High	Crash
Low	High	Cold	Low	Safe
Low	Low	Mild	High	Safe
Medium	High	Hot	Low	Crash
High	Low	Mild	Low	Safe
Medium	High	Mild	High	Crash
High	Low	Cold	High	Crash
Low	Low	Cold	Low	Safe
Medium	Low	Mild	Low	Safe
Low	High	Hot	High	Crash

Problems 19

Given the training dataset, in order to build a decision tree to determine the optimal destination for a drone which attribute amongst altitude, speed, wind, temperature, and weather conditions.is most significant?

Altitude	Speed	Wind	Temperature	Weather	Destination
High	Fast	Weak	Warm	Sunny	City
Low	Slow	Strong	Cold	Rainy	Forest
Medium	Medium	Weak	Mild	Cloudy	Beach
High	Slow	Strong	Warm	Cloudy	City
Medium	Fast	Weak	Hot	Sunny	Beach
Low	Medium	Strong	Cold	Rainy	Forest
High	Slow	Weak	Warm	Cloudy	City
Low	Fast	Strong	Hot	Sunny	Forest
Medium	Medium	Weak	Mild	Cloudy	Beach
Low	Slow	Strong	Cold	Rainy	Forest

Problems 20

In this dataset, we have five discrete attributes: Object Shape, Object Size, Object Weight, Object Color, and Target Location, and the Target Location is the target variable we want to predict.

Object Shape	Object Size	Object Weight	Object Color	Target Location
Square	Small	Light	Red	Shelf 1
Circle	Medium	Heavy	Blue	Shelf 2
Rectangle	Small	Light	Green	Shelf 3
Triangle	Large	Heavy	Red	Shelf 4
Circle	Small	Light	Blue	Shelf 2
Square	Medium	Heavy	Green	Shelf 3
Rectangle	Small	Heavy	Red	Shelf 4
Triangle	Large	Light	Blue	Shelf 2
Circle	Small	Heavy	Green	Shelf 3
Square	Medium	Light	Red	Shelf 1

- What is the entropy of the Target Location attribute in the entire dataset?
- Given the Object Shape attribute, what is the entropy of the Target Location attribute?

- Given the Object Size attribute, what is the entropy of the Target Location attribute?
- Given the Object Weight attribute, what is the entropy of the Target Location attribute?
- Given the Object Color attribute, what is the entropy of the Target Location attribute?
- What is the best attribute to split the dataset on to maximize information gain?

Problems 21

What is the entropy of the target variable (successful landing or not)?
Which attribute in the dataset has the highest information gain for predicting the target variable?

Terrain Type	Crater Depth	Sunlight	Gravity	Obstacles	Landing Site	Dust Level	Communication	Successful Landing
Rocky	Shallow	Strong	Low	None	Near Equator	Low	Good	Y
Sandy	Deep	Weak	High	Few	Near Poles	High	Poor	N
Crater	Shallow	Strong	Low	None	Near Equator	Low	Good	Y
Rocky	Deep	Weak	High	Many	Near Poles	High	Poor	N
Flat	Shallow	Strong	Low	None	Near Equator	Low	Good	Y
Sandy	Deep	Strong	High	Few	Near Poles	High	Poor	N
Crater	Shallow	Weak	Low	None	Near Equator	Low	Good	Y
Rocky	Deep	Strong	High	Many	Near Poles	High	Poor	N
Flat	Shallow	Weak	Low	None	Near Equator	Low	Good	Y
Sandy	Deep	Strong	High	Few	Near Poles	High	Poor	N

Problems 22

A company is trying to develop a decision tree classifier to predict whether a customer's HVAC system needs a repair or not based on several features. The features are: outside temperature (in degrees Fahrenheit), inside temperature (in degrees Fahrenheit), humidity level (in percentage), and age of the HVAC system (in years). If the company splits the data on the "outside temperature" feature and calculates the information gain, what is the information gain value?

Outside Temp (F)	Inside Temp (F)	Humidity (%)	Age (Years)	Repair Needed?
<=60	<=62	<=40	<=5	No
>80	>75	>55	>5 and <=10	Yes
>80	>75	>55	>5 and <=10	Yes
>60 and <=70	>62 and <=68	<=40	<=5	No
>70 and <=80	>68 and <=75	>40 and <=55	<=5	No
>80	>75	>55	<=5	Yes
<=60	>62 and <=68	<=40	>5 and <=10	No
>80	>68 and <=75	>40 and <=55	>5 and <=10	Yes
>70 and <=80	>62 and <=68	>40 and <=55	<=5	No
>80	>75	>55	>10	Yes

Tricky problems for 1-2 marks

23. In a decision tree, if a node has 3 possible outcomes with probabilities 0.4, 0.3, and 0.3, what is its entropy?
24. Consider a decision tree with 5 levels and 32 leaves. How many nodes does it have?
25. What is the information gain if a binary split divides a dataset with 10 positive and 10 negative instances into two subsets, each with 5 positive and 5 negative instances?
26. In a decision tree, if a node has 5 possible outcomes with probabilities 0.1, 0.2, 0.3, 0.2, and 0.2, what is its Gini index?
27. Consider a decision tree with 4 levels and 16 leaves. How many branches does it have?
28. In a decision tree, if a node has 4 possible outcomes with probabilities 0.25, 0.25, 0.25, and 0.25, what is its information entropy?
29. What is the maximum possible value of the information gain for a binary split of a dataset with 10 positive and 10 negative instances?
30. Consider a decision tree with 3 levels and 8 leaves. How many decision nodes does it have?
31. In a decision tree, if a node has 6 possible outcomes with probabilities 0.1, 0.2, 0.1, 0.2, 0.2, and 0.2, what is its Gini index?
32. What is the minimum possible value of the information gain for a binary split of a dataset with 10 instances, all of which belong to the same class?
33. A fluid has a density of 800 kg/m^3 and a viscosity of 0.05 Pa·s. It flows through a 3-cm diameter pipe at a velocity of 5 m/s. What is the information gain of splitting the data based on the flow rate, which can either be 10 L/min or 20 L/min?
34. A tank contains water up to a height of 2 m. The tank has a diameter of 3 m. What is the information gain of splitting the data based on the tank material, which can either be steel or concrete?
35. Suppose we have a dataset with 100 examples, 40 of which belong to class A and 60 of which belong to class B. The dataset has one feature, and the feature has two possible values. If 30 examples have the first value of the feature and 70 examples have the second value, what is the information gain?
36. Suppose we have a dataset with 100 examples, and 80 of them belong to class A while 20 belong to class B. We split the data based on a feature that has 70 examples belonging to A and 30 belonging to B in one branch, and 10 belonging to A and 10 belonging to B in the other branch. What is the information gain of this split?
37. A decision tree is being developed to classify whether a bridge is safe or not based on various factors. Out of 100 bridges, 60 are safe and 40 are unsafe. One of the factors being considered is the material used to construct the bridge. Of the safe bridges, 30 are made of steel and 30 are made of concrete. Of the unsafe bridges, 20 are made of steel and 20 are made of concrete. Calculate the information gain for the material feature.
38. A materials engineer is developing a decision tree to classify different types of metals based on their mechanical properties. The engineer has collected a dataset with 100 samples of metals, where 40 samples are aluminum, 30 samples are steel, and 30 samples are copper. The engineer decides to split the dataset based on the tensile strength of the metals, where samples with a tensile strength greater than 500 MPa are classified as "strong" and those with less than 500 MPa are classified as "weak". Calculate the information gain for this split.
39. Suppose we have a dataset of 1000 instances of car accidents with corresponding values of accelerometer readings (in g units) and whether the airbag deployed or not. We split the dataset based on the accelerometer reading being above or below 10g. The split results in two subsets: 600 instances with readings below 10g (out of which 100 have airbag deployed) and 400 instances with readings above 10g (out of which 300 have

airbag deployed). What is the information gain of this split based on the airbag deployment status?

40. Suppose we have a dataset of 100 accelerometers, out of which 60 are MEMS and 40 are piezoelectric. We want to create a decision tree classifier based on two attributes: frequency response and sensitivity. What are possible discrete (categorical)features to these attributes?

Simple theory questions for 4-6 marks

41. What is the relationship between entropy and information in decision trees?
42. How is entropy used to measure the impurity of a decision tree node?
43. How does the information gain criterion help in selecting the best split in a decision tree?
44. Can decision trees handle continuous data? If so, how is entropy used to handle continuous data in decision trees?
45. What is overfitting in decision trees, and how can it be avoided?
46. Can decision trees handle missing data? If so, how is entropy used to handle missing data in decision trees?
47. What is the maximum possible entropy of a decision tree node, and when is it achieved?
48. Can decision trees be used for regression problems, and if so, how is entropy used in regression decision trees?
49. How can decision trees be used for feature selection, and how does entropy play a role in feature selection?
50. What are the limitations of decision trees, and how can entropy be used to overcome these limitations?