Quiz-4

Due Oct 27 at 11:59pm **Points** 36 **Questions** 20
Available Oct 19 at 11:59pm - Oct 27 at 11:59pm **Time Limit** 60 Minutes

Instructions

Preparation:

- The quiz content is drawn from the lecture slides, shared codes, homeworks, and the things we discuss during the lecture & lab.
- To prepare for the quiz, make sure you understand the content in the lecture slides, and pay attention and take notes during the lecture.
- Lecture slides can be found here
  - https://drive.google.com/drive/folders/1xq-9W-PRDtUZHqiyRfVZUZ8iN0iVTQ86?usp=sharing (https://drive.google.com/drive/folders/1xq-9W-PRDtUZHqiyRfVZUZ8iN0iVTQ86?usp=sharing)
- (https://drive.google.com/drive/folders/1xq-9W-PRDtUZHqiyRfVZUZ8iN0iVTQ86?usp=sharing) If

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	12 minutes	32.8 out of 36

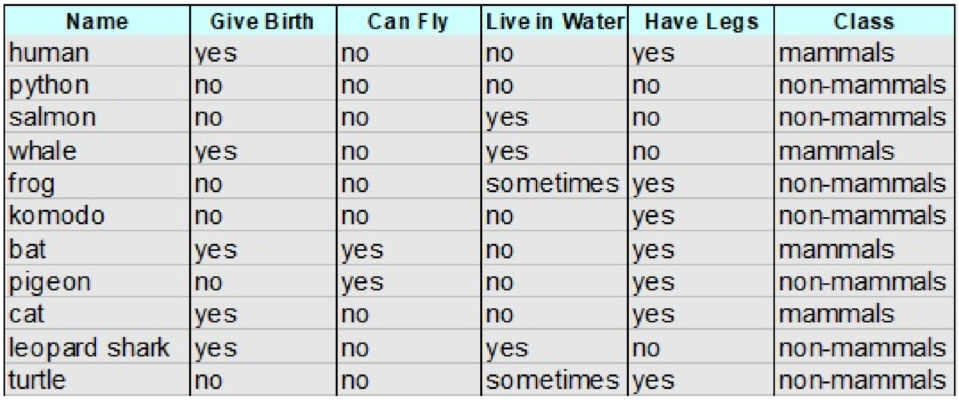
Correct answers are hidden.

Score for this quiz: **32.8** out of 36
Submitted Oct 25 at 5:30pm
This attempt took 12 minutes.

Question 1

2 / 2 pts

You will be using the following dataset to answer this question and other questions below.



Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals

Using paper, build a decision tree for this data.

Use the variable "HaveLegs" as the root node and "class" as the second layer in the tree

Split the node using "yes" and "no" using From there, you can see that neither of the resulting nodes are pure.

Choose the option that best describes their current distribution of class groups for the second layer.

○ YES [Mammals: 4/7 , Non-Mammals: 4/7] AND NO [Mammals: 2/4 , Non-Mammals: 3/4]

○ YES [Mammals: 3/11 , Non-Mammals: 4/11] AND NO [Mammals: 1/11 , Non-Mammals: 3/11]

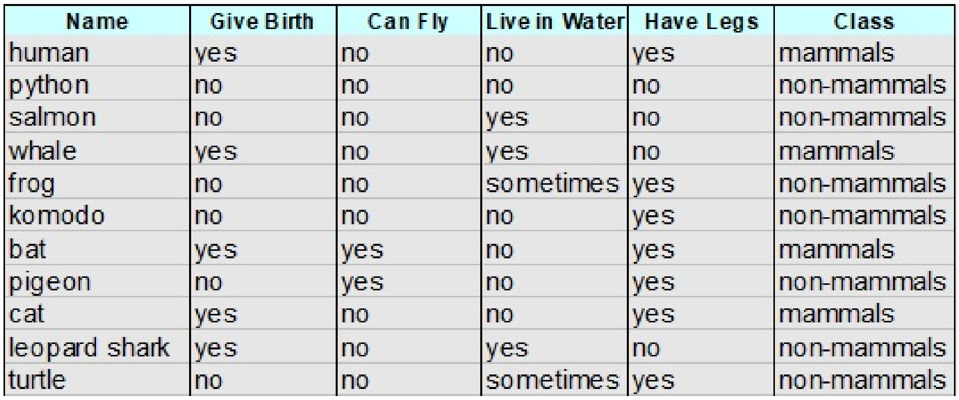
● YES [Mammals: 3/7 , Non-Mammals: 4/7] AND NO [Mammals: 1/4 , Non-Mammals: 3/4]

○ YES [Mammals: 2/4 , Non-Mammals: 2/4] AND NO [Mammals: 4/7 , Non-Mammals: 3/7]

Question 2

3 / 3 pts

You will be using the following dataset to answer this question and other questions below.



Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals

Using paper, build a decision tree for this data.

Use the variable "HaveLegs" as the root node. Split the node using "yes" and "no".

From there, you can see that neither of the resulting nodes are pure.

Calculate the GINI contribution for both of the resulting nodes. Select the correct answer.

○ The GINI for the YES node is .89 and the GINI for the NO node is .29

○ The GINI for the YES node is .41 and the GINI for the NO node is .30

○ The GINI for the YES node is .57 and the GINI for the NO node is .39

● The GINI for the YES node is .49 and the GINI for the NO node is .375

GINI for YES: 1 – (sum(p(YES|mammals)^2 + p(YES|Non-Mammals)^2) = 1 – ((3/7)^2 + (4/7)^2) = .49

GINI for NO: 1 – (sum(p(NO|mammals)^2 + p(NO|Non-Mammals)^2) = 1 – ((1/4)^2 + (3/4)^2) = .375

Question 3

2 / 2 pts

Based on the GINI calculations from Question 2, which node (the YES or the NO) is more pure and why?

○ The YES node is more pure because the GINI is not 0

● The NO node is more pure because the GINI is closer to 0

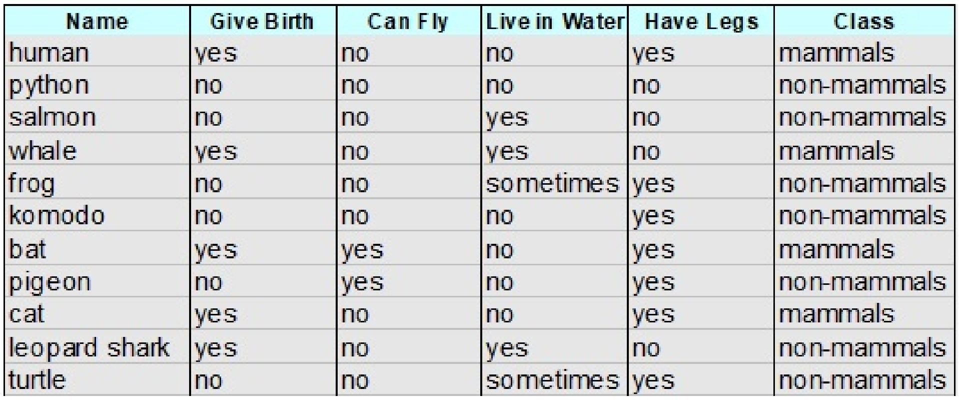
○ The NO node is more pure because the GINI is closer to 1

○ The YES node is more pure because the GINI is closer to .5

Question 4

2 / 2 pts

You will be using the following dataset to answer this question and other questions below.



Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals

Using paper, build a decision tree for this data. Use the variable "HaveLegs" as the root node. Split the node using "yes" and "no". From there, you can see that neither of the resulting nodes are pure.

Calculate the Entropy for both of the resulting nodes. Select the correct answer

Note: Use log_2(x) for the calculation, also the entropy of the YES node is computed using the sum of the entropy p(YES|mammals) and p(YES|Non-Mammals)

○ The Entropy for the YES node is .134 and the Entropy of the NO node is .454

○ The Entropy for the YES node is .721 and the Entropy of the NO node is .533

○ The Entropy for the YES node is .387 and the Entropy of the NO node is .211

● The Entropy for the YES node is .985 and the Entropy of the NO node is .811

Entropy for YES: -(3/7)log(3/7) - (4/7)log(4/7) = -(3/7)*(-1.22) - (4/7)*(-.81) =.5228 + .4628 = .985

Entropy for NO: -(1/4)log(1/4) - (3/4)log(3/4) = -(1/4)*(-2) - (3/4)*(-.415) = .5 + .311 = .811

The Entropy for the YES node is .985 and the Entropy of the NO node is .811

Question 5

2 / 2 pts

At this point, we have a Decision tree with HaveLegs as the root node. We have split that node using YES and NO. We have calculated the GINI and the Entropy for both of the YES and NO nodes that resulted from this split.

Neither the YES or the NO are pure. Therefore, we will need to split both the YES and the NO again.

Let’s focus on splitting the YES. Split the YES using GiveBirth. It is best to draw this out so you can see it.

Notice that GiveBirth has two options: yes and no.

Calculate the **Information GAIN** from splitting the HaveLegs YES into GiveBirth (yes or no). Use Entropy as the measure. You already know the Entropy for the HaveLegs YES. Now you need the Entropy for the GiveBirth (yes) and the GiveBirth (No).

From there, you can calculate the GAIN.

What is the information GAIN?

○ .784

○ .381

○ .562

● .985

Entropy for Parent =

Entropy for YES from parent node of HaveLegs:

-(3/7)log(3/7) - (4/7)log(4/7) = -(3/7)*(-1.22) - (4/7)*(-.81) =.5228 + .4628 = .985

Entropy for GiveBirth YES node

P(Mammals| HaveLegs=Yes and GiveBirth=Yes) is 3/3

P(Non-Mammals| HaveLegs=Yes and GiveBirth=Yes) is 0/3

Entropy: -(3/3)log(3/3) - (0/3)log(0/3) =0

P(Mammals| HaveLegs=Yes and GiveBirth=No) is 0/4

P(Non-Mammals| HaveLegs=Yes and GiveBirth=No) is 4/4

Entropy: -(0/4)log(0/4) - (4/4)log(4/4) =0

GAIN:

Information GAIN:

I(Parent) - sum over all children N(v)/N * I(v) =

.985 - (3/7) * 0 - (4/7)*0 = .985

Question 6

2 / 2 pts

Suppose you have a dataset and one of the variables is GPA.

Choose all options that could make sense - there is more than one answer.

☐ Split the variable for as many GPAs as there are. This tree cant overfit because the number of leaves will be N.

☑ Discretize the GPA into 5 groups: A, B, C, D, F. Then build a tree to split the data into these groups

☑ Split the node with GPA > 3 to one side and GPA <= 3 to the other side.

☑ Split the GPA into >= 3.0, between 3.0 and 2.0, and less than 2.0

Question 7

2 / 2 pts

Which of the following statements are true. Choose ALL that are True.

☑ Decision Trees can be used to model both qualitative and quantitative data

☑ Decision Trees are a supervised learning method

☑ Building a Decision Tree Model requires labeled data

☐ Classification Trees use Euclidean distance to measure node similarity

☐ Decision trees can be used to model only qualitative data

Question 8

2 / 2 pts

Which of the following statements are true. Choose ALL that are True.

☑ A pure node has only points (data values) that belong to one class or group.

☑ If the GINI of a node resulting from a split is 0, this means that the node is pure.

□ If the entropy of a node is 1, the node is pure.

☑ When you split a node, it is possible to measure the GINI for all resulting nodes.

☑ If a split results in one or more impure nodes, it is possible to split the impure nodes again

Question 9

2 / 2 pts

Suppose a node in a decision tree is not pure. This means that it contains points (data values) from more than one class or group. Suppose you want to split that node so that the resulting nodes are pure. One method is to choose a variable to split with and then to measure the information GAIN between the parent node (the node you split) and the children nodes (the nodes that resulted from the split. Suppose the split was binary - meaning the split resulted in two children.

Suppose the Entropy for the parent was .781

Suppose the information GAIN is .781.

What does this mean?

○ Because the GAIN is the same as the entropy for the parent, the split did not help with adding purity.

○ It is not possible for a GAIN to be the same as the entropy of the parent

● Because the information GAIN is the same as the Entropy of the parent, this is the maximum possible difference and so both children nodes are pure.

Question 10

1 / 1 pts

In random forest, many trees are created and a "vote" is used to choose the final label.

● True

○ False

Question 11

2 / 2 pts

Select all that are true about decision trees

☐ DTs can only be used for classification, i.e regression isn't possible with a DT

☑ For a given training set, there are MANY possible decision trees, this makes finding the correct tree a difficult search process

☑ Different trajectories along the tree lead to different prediction outcomes

☑ DTs predict the value of a target by learning simple decisions inferred from the training data

☐ Decision trees are evaluated via a non-sequential trajectory of rule based decisions

☐ The process starts at the leaves of the tree and works its way up to the root node

☑ The outcomes are stored in the contents of the tree's leaves

Question 12

1 / 1 pts

Algorithms can find “good” decision trees using a so called "Greedy" approach. Where they make a series of locally optimal decisions.

● True

○ False

Question 13

2 / 2 pts

Select all of the following common advantages of Decision trees.

- ☑ Easy to understand and interpret.
- ☑ Computationally Cheap
- ☐ Highly stable, small variations in the data result in small changes to the predictions
- ☑ Require minimal data preparation
- ☐ Very difficult to overfit
- ☐ DTs make predictions on a global level, as opposed to some algorithms that can only make "local" predictions
- ☐ It is easy to find globally optimal trees due to the narrow scope of the search space
- ☑ Easily handle multi-output problems
- ☑ The trees can handle both numerical and categorical data.

Question 14

1 / 1 pts

Random forests are an ensemble learning method (collection of "weak" learners) which operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by the majority of the trees.

● True

○ False

Question 15

2 / 2 pts

Select the best matching pair regarding decision tree

hyper-parameter that controls the number of layers in the tree	max_depth ⌄
hyper-parameter that controls the number of samples required to split an internal node	min_samples_split ⌄
The minimum number of samples required to be at a leaf node	min_samples_leaf ⌄
The method used to measure the quality of a split (“gini”, “entropy”, “log_loss”)	criterion ⌄

Question 16

2 / 2 pts

Select the best matching pairs for decision trees

number between 0-0.5, which indicates the likelihood of new, random data being misclassified	Gini Impurity
reduces the size of decision trees by removing sections of the tree that are non-critical or redundant	Pruning
-p log(p)	Entropy
Node that cant be split any further (all counts fall into one class)	Pure
A node that could be split futher	Im-pure

Incorrect

Question 17

0 / 2 pts

Choose all that are TRUE regarding SVM

☐ To model data with an SVM, the output data must be labeled

☑ SVMs are linear separators - they separate data into two groups.

☐ SVMs inputs can be qualitative or quantitative data.

☑ ONE SVM can classify data into two or more groups

Question 18

1 / 1 pts

Support-vector machines (SVM) are supervised learning models that analyze data for classification and regression analysis.

The goal is to separate the points described by input features in an N dimensional space using an N-1 dimensional hyperplane (i.e. decision boundary).

● True

○ False

Question 19

1 / 1 pts

SVMs are inherently multi-class classifiers

○ True

● False

Partial

Question 20

0.8 / 2 pts

Select the best matching pairs regarding SVM

at least one plane exists with all of one class on one side and all of the other class on the other side.	maximum-margin hyp
special subset of the training data, which is close to the hyperplane and strongly effects the fitting	support vectors.
shortest distance between nearest observations and the hyperplane	Hinge loss
No training predictions are allowed within the margin	Hard margin classifier
Allows mis-classifications and predictions within the margin	Soft margin classifier
Extends binary classifiers to multi-class classifiers by fitting one classifier per class.	One-vs-One
Extends binary classifiers to multi-class classifiers by fitting one classifier per pair of classes.	One-vs-Rest

hyperplane that lies halfway between data cloud edges	margin
Function that relaxes the hard margin constraint and allows classifications "inside" the margin	Linear separability
"Trick" the linear SCV into acting as a non-linear classifier	Nonlinear Kernels

Quiz Score: **32.8** out of 36