Remaining Time: 36 minutes, 53 seconds.

Question Completion Status:

QUESTION 1

Arrange the following datasets according to ascending values of entropy:

4.

Instance	a_1	a_2	a_3	ass
1	0	0	0	1
2	0	0	1	0
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	0
7	1	1	0	1
8	1	1	1	0

1.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1
4	0	1	1	1
5	1	0	0	1
6	1	0	1	1
7	1	1	0	1
8	1	1	1	1

1.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Blackboard

Remaining Time: 45 minutes, 01 second.

Question Completion Status:

Arrange the following datasets according to ascending values of entropy:

4.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	0
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	0
7	1	1	0	1
8	1	1	1	0

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1
4	0	1	1	1
5	1	0	0	1
6	1	0	1	1
7	1	1	0	1
8	1	1	1	1

1.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Blackboard

Remaining Time: 36 minutes, 46 seconds.

Question Completion Status:

6	1	0	1	1
7	1	1	0	1
8	1	1	1	1

1.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	1
7	1	1	0	1
8	1	1	1	0

2.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	0
3	0	1	0	1
4	0	1	1	0
5	1	0	0	0
6	1	0	1	0
7	1	1	0	1
8	1	1	1	0

3.

QUESTION 2

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Blackboard

Remaining Time: 44 minutes, 54 seconds.

Question Completion Status:

1.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	1
7	1	1	0	1
8	1	1	1	0

2.

Instance	a_1	a_2	a_3	Class
1	0	0	0	1
2	0	0	1	0
3	0	1	0	1
4	0	1	1	0
5	1	0	0	0
6	1	0	1	0
7	1	1	0	1
8	1	1	1	0

3.

QUESTION 2

Imagine you have 20 Boolean variables {X1, X2, ..., X20}. You would like to construct a full Boolean t
correct?

- The number of leaves in such a tree would be 2^{20} and number of ways of labeling them would

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Question Completion Status:

QUESTION 2

Imagine you have 20 Boolean variables {X1, X2, ..., X20}. You would like to construct a full Boolean tree i.e. a tree built using all 20 of these Booleans. Which of the following is correct?

- (●) The number of leaves in such a tree would be 2^{20} and number of ways of labeling them would be $2^{2^{20}}$
- ( ) The number of leaves in such a tree would be 20 and number of ways of labeling them would be 2^{20}
- ( ) The number of leaves in such a tree would be 2^{20} and number of ways of labeling them would also be 2^{20}
- ( ) The number of leaves in such a tree would be 2^{10} and number of ways of labeling them would be $2^{2^{10}}$

QUESTION 3

In a d-dimensional Euclidean space, what is the shortest distance from a point x_0 to a hyperplane $H : w^T x + b = 0$? (Notation: $\|w\|_2 = \sqrt{\sum_i w_i^2}$

- ( ) $|w^T x_0 + b|$
- (●) $|w^T x_0 + b| / \|w\|_2$
- ( ) $|(w/\|w\|_2)^T x_0 + b|$
- ( ) None of the above

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers

Question Completion Status:

QUESTION 4

2 points

Suppose there is a SVM classifier with two classes C1 and C2, each having n1 and n2 number of data points that are all linearly separable. Assume you are able to construct a SVM with 0 error. What is the minimum and maximum number of support vectors possible?

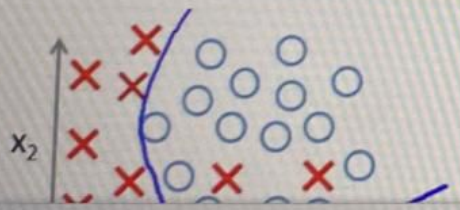
- (●) Minimum: 2
  Maximum: n1+n2
- ( ) Minimum: 2
  Maximum: 2
- ( ) Minimum: 2
  Maximum: (n1+n2)/2
- ( ) Minimum: 1
  Maximum: n1+n2

QUESTION 5

2 points

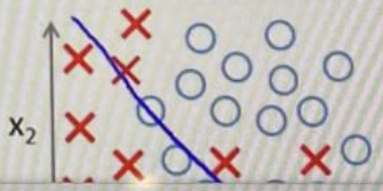
In the images below the points represent data point of two different classes and the blue line represents the classification model. Find their correct order in terms of increasing model complexity.

Model 1



Model 2

Model 3



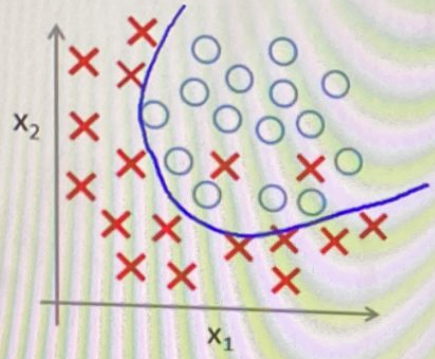
Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window | Save a

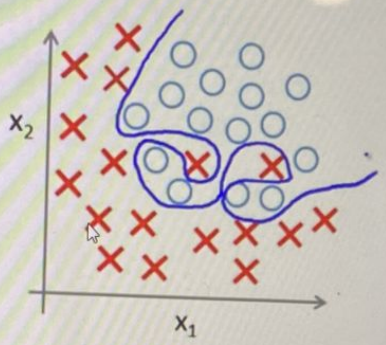
In the images below the points represent data point of two different classes and the blue line represents the classification model. Find their correct order in terms of increasing model complexity:

2 points

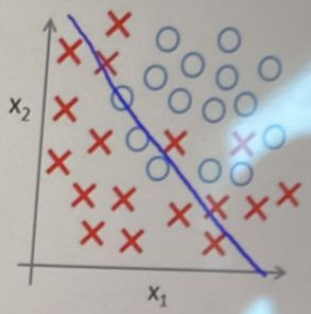
Model 1



Model 2



Model 3



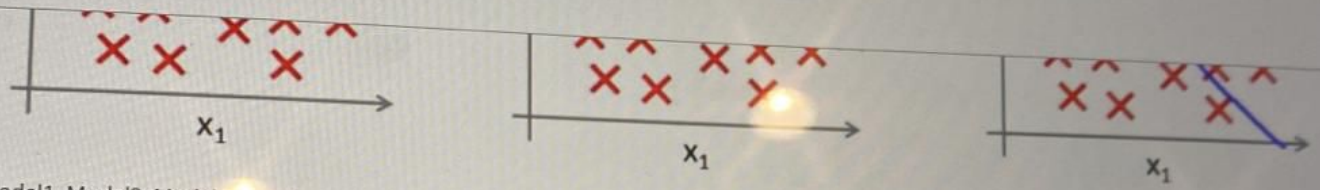
- ○ Model1, Model2, Model 3
- ○ Model1, Model 3, Model2
- ○ Model2, Model3, Model1
- ◉ Model3, Model1, Model2

QUESTION 6

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Remaining Time: 36 minutes, 19 seconds.

Question Completion Status:



- Model1, Model2, Model 3
- Model1, Model 3, Model2
- Model2, Model3, Model1
- Model3, Model1, Model2

QUESTION 6

2 point

Suppose you have a biased coin whose probability of getting heads is 0.6. You toss this coin 10 times, and observe the number of heads that occur, called B(X). What is the mean and variance of B(X)?

- Mean = 6
  Variance = 1
- Mean = 6
  Variance = 2.4
- Mean = 5
  Variance = 5
- Mean = 5
  Variance = 2.4

Remaining Time: 36 minutes, 08 seconds.

Question Completion Status:

QUESTION 7

For a classification problem, you propose a hypothesis as:
$h(x) = \text{sign}(w^T x + b)$
where the symbols have their usual meaning
and the real class label is $y \in [-1, 1]$

Which one of the following represents a misclassified data point:

- (●) $y_i h_i < 0$
- ( ) $h_i < 0$ and $y_i < 0$
- ( ) $y_i h_i > 0$
- ( ) $h_i > 0$ and $y_i > 0$

QUESTION 8

What is the difference between hard-margin and soft-margin SVM classification?

- (●) Hard margin SVM doesn't allow slack variables for misclassified points, while soft margin uses slack variables in the optimization condition.
- ( ) Hard margin gives better results
- ( ) Soft margin classification has less parameters than hard margin SVM
- ( ) Hard margin considers maximization of margin, as well as length of slack variables.

QUESTION 9

QUESTION 9

What is the potential problem with using a Naive Bayes classifier that was learned without Laplace smoo[illegible]?

2 point

- It can lead to over-emphasis on certain attributes
- It can lead to ignorance of certain attributes
- It can lead to zero probability values if one class has 0 instance having a particular value of attribute
- It can lead to very high probability values

QUESTION 10

5 poin

We wish to learn a decision tree to help students pick restaurants using three aspects – the price, the location of the restaurant and the speed of service. The data for training the tree is given below, where the target concept is the column labeled "Like?"

#	Price	Fast?	On Campus?	Like?
1	$	No	No	No
2	$	Yes	Yes	No
3	$$	No	No	No
4	$$	Yes	Yes	Yes
5	$$	Yes	No	Yes
6	$$	No	Yes	Yes
7	$$$	No	No	No
8	$$$	Yes	Yes	Yes

Let E(m, n) denote the entropy of a dataset having m instances of class1 and n instances of class 2 (where m > n).

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

QUESTION 10

We wish to learn a decision tree to help students pick restaurants using three aspects – the price, the location of the restaurant and the speed of service. The data for training the tree is given below, where the target concept is the column labeled "Like?"

#	Price	Fast?	On Campus?	Like?
1	$	No	No	No
2	$	Yes	Yes	No
3	$$	No	No	No
4	$$	Yes	Yes	Yes
5	$$	Yes	No	Yes
6	$$	No	Yes	Yes
7	$$$	No	No	No
8	$$$	Yes	Yes	Yes

Let E(m, n) denote the entropy of a dataset having m instances of class1 and n instances of class 2 (where m > n).
Which of the following denotes the Information Gain when the split is performed on attribute "Price"

- ○ $1 - \frac{4}{8}E(2, 2) - \frac{2}{8}E(3, 1) - \frac{2}{8}E(1, 0)$
- ○ $1 - \frac{4}{8}E(2, 0) - \frac{2}{8}E(3, 1) - \frac{2}{8}E(1, 1)$
- ○ $1 - \frac{2}{8}E(2, 0) - \frac{2}{8}E(3, 1) - \frac{4}{8}E(1, 1)$
- ◉ $1 - \frac{2}{8}E(2, 0) - \frac{1}{2}E(3, 1) - \frac{2}{8}E(1, 1)$

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

QUESTION 11

2 points

Suppose you have 100 data points belonging to two classes - 0 and 1 - for a continuous 2-dimensional dataset. You have to learn a classification function f defined as follows: f: X -> Y where X = (x1, x2) where x1 and x2 are continuous variables and Y is a Boolean with values 0 and 1. You can assume that the dataset is consistent and has no noise points.
You decide to use a decision tree classifier. What is the depth of the decision tree that can guarantee a zero training error for any distribution and labeling of the data points.

- 7
- 8
- (selected) A decision tree cannot guarantee zero training error for all distribution and labelings.
- 5

QUESTION 12

2 points

Which one(s) of the following represent the sigmoid activation function?

- $f(x) = \frac{1}{1+e^{x/2}}$
- $f(x) = \frac{e^{-x}}{1+e^{-x}}$
- (checked) $f(x) = \frac{1}{1+e^{-x}}$
- $f(x) = \frac{1}{1+e^{x}}$

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window | Save

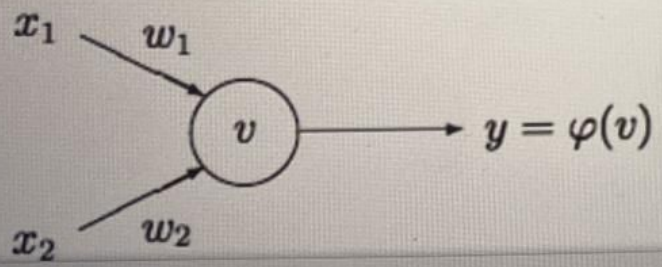
QUESTION 13

You are given a set of n coins. All of them, except the last one, are unbiased with equal probability of heads and tails. The last one is a fake coin with heads on both sides. You select a coin at random and toss it to get heads. What is the probability that the coin chosen was fake. That is, you would like to calculate P(fake|heads).

- ○ $\frac{1}{n}$
- ◉ $\frac{2}{n+1}$
- ○ $\frac{1}{2n+1}$
- ○ $\frac{2}{n}$

QUESTION 14

You would like to represent the AND function using the ANN below:

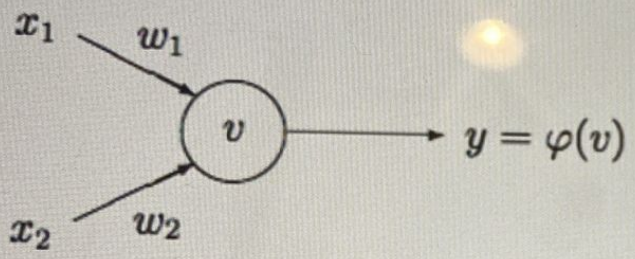


Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Question Completion Status:

QUESTION 14

You would like to represent the AND function using the ANN below:



Using the weights vector w = (w1, w2) = (1, 1) and v = $w^T x$, which one of the following functions φ can represent the AND function:

- $\varphi(v) = \begin{cases} 1 \text{ if } v \geq 1 \\ 0 \text{ otherwise} \end{cases}$
- $\varphi(v) = \begin{cases} 1 \text{ if } v > 0 \\ 0 \text{ otherwise} \end{cases}$
- $\varphi(v) = \begin{cases} 1 \text{ if } v > 0 \\ 0 \text{ otherwise} \end{cases}$
- $\varphi(v) = \begin{cases} 1 \text{ if } v > 1 \\ 0 \text{ otherwise} \end{cases}$

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers

QUESTION 15

Consider the data in the table below relating gender, and hours worked to probability of being rich or poor

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Which of the following are true? Round your calculations to 3 decimal places

- [x] P(Female | Rich) = 0.151
- [ ] P(Female | Rich) = 0.214
- [ ] P(Male | Poor) = 0.388
- [x] P(Male | Poor) = 0.612

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Question Completion Status:

rain	mild	high	true	N

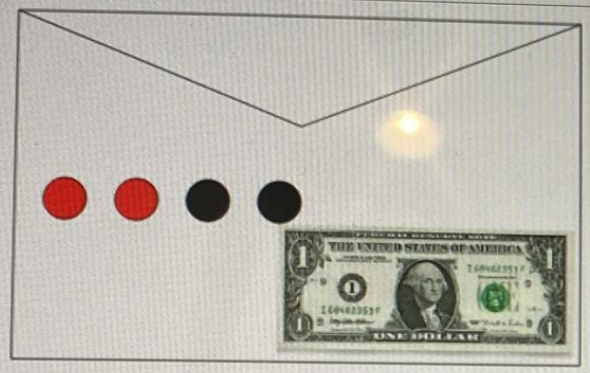
You have to train a Naive Bayes classifier using the above attributes and select correct options from the choices below
Note:
- The answers below are rounded up to 3 decimal places
- There is no need to use Laplace smoothing for this question.

- [ ] Prior of playing tennis is 0.643
- [x] Prior of not playing tennis is 0.643
- [ ] Prior of not playing tennis is 0.357
- [ ] Prior of not playing tennis is 0.5
- [ ] The conditional probability of outlook being rain given class P i.e. P(outlook=rain | Class=P) is 0.333
- [x] The probability of outlook being rain i.e P(outlook=rain) is 0.222
- [ ] The conditional probability of humidity being high given class N i.e. P(humidity=high | Class=N) is 0.800
- [ ] The conditional probability of humidity being high given class P i.e. P(humidity=high | Class=P) is 0.200
- [ ] Given the following test data X = {Outlook = rain, Temperature = cool, Humidity = high, Windy = true}. Using the naïve Bayes assumption, the product of likelihood and prior for data X given class P would be 0.008
- [ ] Given the following test data X = {Outlook = rain, Temperature = cool, Humidity = high, Windy = true}. Using the naïve Bayes assumption, the product of likelihood and prior for data X for class P would be 0.014
- [x] Given the following test data X = {Outlook = rain, Temperature = cool, Humidity = high, Windy = true}. Using the naïve Bayes assumption, the product of likelihood and prior for data X for class N would be 0.014
- [ ] Using the MAP approach, the most likely class for the test data X = {Outlook = rain, Temperature = cool, Humidity = high, Windy = true} would be P

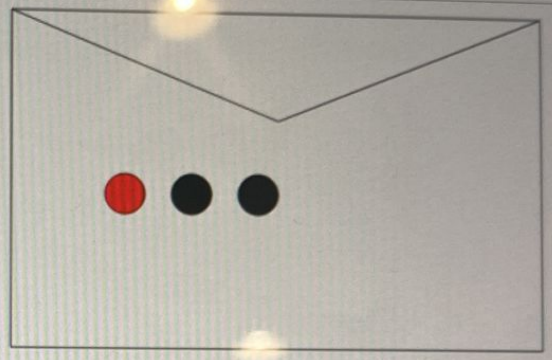
Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window

Question Completion Status:



The "Win" envelope has a dollar and four beads in it



The "Lose" envelope has three beads and no money

You are given a chance of drawing one bead at random from the envelopes. You can see its color, but not where it came from. Given that the bead drawn is black, what is the probability of win.
Note:
You can assume the priors for both envelopes to be the same i.e. P(win) = 0.5 P(lose)=0.5

- (●) 0.572
- ( ) 0.25
- ( ) 0.5
- ( ) 0.428

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers

QUESTION 18

Suppose you toss a fair coin three times. What is the expected value of the number of heads obtained?

- (●) 1.5
- ( ) Not enough information available
- ( ) 2.0
- ( ) 0.5

QUESTION 19

Let X1, . . . , Xk be k Boolean variables. Let f_k be a target concept which is a disjunction consisting of k literals.Example of f_2: X1 V X2. Example of f_3: X1 V X2 V X3
You would like to find the size of the best consistent decision tree that represents this function f_k for different values of k. Remember the best consistent tree is the shortest (or most compact) tree that is able to represent the concept correctly.
The size of the decision tree will be calculated in terms of number of nodes in the tree.
Choose all the correct options from the choices below.

- [ ] For k = 2, the best tree contains 3 internal nodes and 4 leaf nodes
- [x] For k = 2, the total size of the best consistent tree is 5 nodes
- [ ] For k = 3, the total size of the best consistent tree is 9 nodes
- [ ] For k = 3, the best tree contains 4 internal nodes and 5 leaf nodes
- [x] For k = 2, the best tree contains 2 internal nodes and 3 leaf nodes
- [x] For k = 3, the total size of the best consistent tree is 7 nodes
- [x] For k = 3, the best tree contains 3 internal nodes and 4 leaf nodes

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Remaining Time: 34 minutes, 46 seconds.

Question Completion Status:

QUESTION 19

Let X1, . . . , Xk be k Boolean variables. Let f_k be a target concept which is a disjunction consisting of k literals.Example of f_2: X1 V X2. Example of f_3: X1 V X2 V X3
You would like to find the size of the best consistent decision tree that represents this function f_k for different values of k. Remember the best consistent tree is the shortest (or most compact) tree that is able to represent the concept correctly.
The size of the decision tree will be calculated in terms of number of nodes in the tree.
Choose all the correct options from the choices below.

- [ ] For k = 2, the best tree contains 3 internal nodes and 4 leaf nodes
- [x] For k = 2, the total size of the best consistent tree is 5 nodes
- [ ] For k = 3, the total size of the best consistent tree is 9 nodes
- [ ] For k = 3, the best tree contains 4 internal nodes and 5 leaf nodes
- [x] For k = 2, the best tree contains 2 internal nodes and 3 leaf nodes
- [x] For k = 3, the total size of the best consistent tree is 7 nodes
- [x] For k = 3, the best tree contains 3 internal nodes and 4 leaf nodes
- [ ] For k = 2, the total size of the best consistent tree is 7 nodes
- [x] For any value k, the total size of the best consistent tree is 2k+1 nodes
- [ ] For any value k, the total size of the best consistent tree is 2^k+1 nodes

QUESTION 20

perceptron algorithm on this data

Question Completion Status:

☐ For any value k, the total size of the best consistent tree is 2^k+1 nodes

QUESTION 20

10 po

Assume you are given the following training data, which consists of three dimensions (x1, x2, x3) and and a class label. You have to train a perceptron algorithm on this data.

Instance	x1	x2	x3	Class
1	-1	1	-1	1
2	1	1	1	-1

You are given initial weights vector (w) as (0.5, -1.0, -0.5). Note: You can assume 0 bias for this problem. Recall, that a perceptron algorithm classifies points on the basis of following function:

$$f(x) = \text{sign}(w^T x)$$

You are to run the perceptron training algorithm for all instances in the above dataset. You can assume a learning rate of 1.
Select the correct options from the choices below:

☑ After applying the perceptron training rule to the first instance, the weights would be (-1.5, 1.0, -2.5)

☐ After applying the perceptron training rule to the first instance, the weights would be (-1.5, -1.0, -2.5)

☐ After applying the perceptron training rule to the second instance, the weights would be (-2.5, 1.0, -1.5)

☐ The perceptron training rule will correctly predict both the above data points after the first iteration

☐ When the perceptron training rule is applied to the first instance, the weights would not change

☐ After applying the perceptron training rule to the second instance, the weights would be (-1.5, 1.0, -2.5)

QUESTION 21

2 points

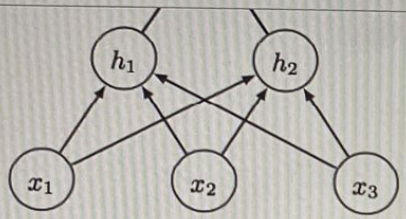
Consider the simple 1-hidden layer neural network shown below. The input layer consists of three dimensions (x1, x2, x3), the hidden layer consists of two units (h1, h2), and

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window | Save a

Remaining Time: 34 minutes, 30 seconds.

Question Completion Status:



The initial weights are given as below:

Weights connecting input layer to hidden layer: $W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$

Note that in the above matrix the element w_{11} i.e. the first row and first column represents the weight from x_1 to h_1, similarly the element w_{12} i.e. the first row and second column represents the weight from x_2 to h_1, and so on.

Weights connecting hidden layer to output layer: $V = [0 \ 1]$

where the elements have the same meaning as explained previously.

For the training data, you are given an input vector X = (1, 2, 1) and target output t = 1.
You can assume that the hidden layer and output layer both use the sigmoid activation function, and the error function is defined as below:

$E = \frac{1}{2}\sum_i (t_i - o_i)^2$ where the sum is over all training examples.

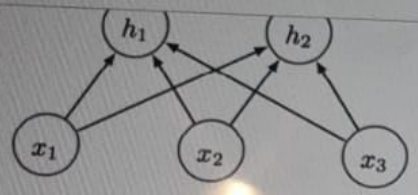
For the given training data and parameters, what will be the error at the output layer after the first forward pass?
Note: You can round up your calculation to 3 decimals at every intermediate step

- ○ 0.087
- ○ 0.534
- ◉ 0.094
- ○ 0.567

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window | Save

Question Completion Status:



The initial weights are given as below:

Weights connecting input layer to hidden layer: $W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$

Note that in the above matrix the element w_{11} i.e. the first row and first column represents the weight from x_1 to h_1, similarly the element w_{12} i.e. the first row and second column represents the weight from x_2 to h_1, and so on.

Weights connecting hidden layer to output layer: $V = [0 \ 1]$

where the elements have the same meaning as explained previously.
For the training data, you are given an input vector X = (1, 2, 1) and target output t = 1.
You can assume that the hidden layer and output layer both use the sigmoid activation function, and the error function is defined as below:
$E = \frac{1}{2}\sum_i (t_i - o_i)^2$ where the sum is over all training examples.

For the given training data and parameters, what will be the value of δ_y i.e. delta at the output neuron after the first forward pass. Remember that calculating this value is the first step during the backpropagation algorithm
Note: You can round up your calculation to 3 decimals at every intermediate step

- ○ 0.121
- ○ 0.678
- ◉ 0.106
- ○ 0.594

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Question Completion Status:

The initial weights are given as below:

Weights connecting input layer to hidden layer: $W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$

Note that in the above matrix the element w_{11} i.e. the first row and first column represents the weight from x_1 to h_1, similarly the element w_{12} i.e. the first row and second column represents the weight from x_2 to h_1, and so on.

Weights connecting hidden layer to output layer: $V = [0 \ 1]$

where the elements have the same meaning as explained previously.

For the training data, you are given an input vector X = (1, 2, 1) and target output t = 1.

Now, assume that each of the hidden layer and output layer neurons are Rectified Linear Units (ReLu) and are using the rectified activation function, which is defined as below:

$f_{ReLu}(x) = max(0, x)$

and the error function is defined as below:

$E = \frac{1}{2}\sum_i (t_i - o_i)^2$ where the sum is over all training examples.

For the given training data and parameters, what will be the value of the output produced by unit y after the first forward pass through the neural network?

Note: You can round up your calculation to 3 decimals at every intermediate step

- ○ 0.567
- ○ 2.0
- ◉ 0.0
- ○ 1.0

QUESTION 24

Question Completion Status:

- (●) 0.0
- ( ) 1.0

QUESTION 24

6 points

Consider the problem of designing a neural network for predicting a scalar output in a dataset having 3 attributes. You design a neural network having 2 layers of size 4 and 2 respectively. Which of the following would be true?

- [x] The number of connections between the first hidden layer and second hidden layer would be 10. Remember that we are including the bias unit connections also.
- [x] The number of neurons in the input layer would be 3.
- [ ] The number of connections between the input layer and first hidden layer would be 12. Remember that we are including the bias unit connections also.
- [x] The number of connections between the input layer and first hidden layer would be 16. Remember that we are including the bias unit connections also.

QUESTION 25

6 points

Which of the following is/are true about the learning rate?

- [ ] It is used to obtain the next point in a gradient descent algorithm
- [x] If too low a value is selected, the algorithm will take a very large time (i.e. large number of iterations) to converge
- [x] If too large of a value is selected, the algorithm will oscillate about the minimum point
- [x] Learning rate is a hyperparameter for the gradient descent algorithm

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All Answers | Close Window | Save