## Homework 3

Started: Oct 20 at 4:37pm

## **Quiz Instructions**

This homework consists of a collection of multiple choice questions.

More than one answer may be correct. You should select all the correct answers to get the points.

| , [ | Question 1   | 3 pts   |
|-----|--|---------|
|     | Suppose we have a binary classification problem with $m{n}$ features. Each feature in $m{c}$               |         |
|     | problem can take one of three values A, B or C. How many binary classifiers are poover this feature space? | ossible |
|     | $lacksquare$ $2^n$ $3^{2^n}$   |         |
|     | $\bigcirc$ $3^n$   |         |
|     | $\bigcirc$ $2^{3^n}$   |         |
|     |  |         |

| > | Question 2   | 3 pts |
|---|--|-------|
|   | How many disjunctions are possible with n Boolean features if we do not allow any negations? |       |
|   | $\bigcap n^2$  |       |
|   | $\bigcirc$ $2^{2^n}$   |       |
|   | $\bigcap n$  |       |
|   |  |       |

| $\bigcirc$ $2^n$ |  |  |  |
|------------------|--|--|--|
|                  |  |  |  |

An m-of-n function is defined as follows: Select a *fixed* subset of Boolean variables of size n. The function returns true for inputs where m of these chosen variables are true.

Which of the following statements are correct about m-of-n functions?

Every conjunction without negations can be represented as a m-of-n function

Every disjunction without negations can be represented as a m-of-n function

Every Boolean function can be represented as a m-of-n function

m-of-n functions can be represented by linear classifiers

Which of the following statements about decision trees are correct?

Every Boolean function can be represented as a decision tree

Every Boolean function can be represented by a unique decision tree

Real valued features have to be discretized to use them with decision trees

Decision trees represent only linearly separable functions

Question 5

Suppose we know that

$$P(X=A)=\tfrac{1}{16},$$

$$P(X=B)=\tfrac{1}{16},$$

$$P(X=C)=\tfrac{1}{8},$$

$$P(X=D)=\frac{1}{4},$$

$$P(X=E)=rac{1}{2}$$

Select all statements that are correct.

- These probabilities give the maximum possible value of the entropy of X
- Entropy(X) = 1.875
- Entropy(X) = 1.0
- These probabilities give the minimum possible value of the entropy of X
- Question 6 3 pts

Which of the following statements about the ID3 algorithm are correct?

- It is an online algorithm.
- It assumes that the training set is chosen uniformly at random from the instance space.
- It will always find a decision tree that will fit any training set.
- It is a batch algorithm.
- Question 7 3 pts

Suppose we have three features (x1, x2 and x3) and a label y that can be either A or B. We have the following training set:

0 0 0 B

0 1 1 A

1 1 0 B

| 1 0 1 B   |
|---|
| This data is not linearly separable.  |
| $igcap$ The entropy of the label is $2+rac{3}{4}{ m log}_2~3$                    |
| $igspace{ igspace{1.5mm} }$ The entropy of the label is $2-rac{3}{4}{ m log}_23$ |
| This data is linearly separable   |

Question 8
If your training data has a missing feature value, which of the following approaches can be used to handle it in the ID3 algorithm?
✓ Use fractional feature values representing the proportion of training examples that take each value.
✓ Discard the training example because we can't use it for training.
✓ Use the most common value of that feature among the other examples that share the same label.
✓ Use the most common value of that feature among the other examples.

| Question 9  | 3 pts |
|---|-------|
| A learning algorithm is said to overfit its training data if: |       |
| Its hypothesis space is too small to express the data.        |       |
| Its hypothesis space contains the true concept function       |       |
| Its training error is more than its generalization error.     |       |
| Its training error is less than its generalization error      |       |
|   |       |

| Question 10  | 3 pts           |
|--|-----------------|
| Use the following data with features x1, x2 and labels y and select all state correct. | ements that are |
| x1 x2 y  |                 |
| 0 0 1  |                 |
| 0 1 -1   |                 |
| 1 1 1  |                 |
| This function can not be represented by a decision tree.                               |                 |
| This function can be represented by a decision tree.                                   |                 |
| This function can be represented by a linear threshold unit.                           |                 |
| This function can not be represented by a linear threshold unit.                       |                 |
|  |                 |

| Que            | stion 11   | 3 pt |
|----------------|--|------|
| Which<br>separ | n of the following Boolean functions with variables $x_1, x_2, x_3, x_4$ are linearly able?      |      |
| $\checkmark$   | Label is true if any two out of $oldsymbol{x_1}$ , $oldsymbol{x_2}$ or $oldsymbol{x_4}$ are true |      |
| $\checkmark$   | $x_1 ee  eg x_2$   |      |
|                | Label is true when an even number of x's are true.   |      |

| You have a dataset on which you ran the Perceptron algorithm. You find that the algoridoesn't stop making mistakes. Which of the following may help? | lgorithm |
|--|----------|

|              | Delete examples where the algorithm makes mistakes and try again. |
|--------------|---|
| $\checkmark$ | Transform the data using a non-linear feature transformation.     |
| ✓            | Run multiple epochs over shuffled versions of the data.           |

| Question 13  | 3 p |
|--|-----|
| Which of the following linear threshold units is equivalent to the following Boolean function: $x_1 \lor x_2 \lor x_3$ ? |     |
| ${\color{red} {m igstar} \hspace{05in} Sgn(x_1+x_2+x_3-1)}$  |     |
| $ \boxed{ \qquad sgn(x_1+x_2+x_3+1) }$   |     |
|  |     |
| $ \boxed{ \qquad sgn(-x_1-x_2-x_3+1) }$  |     |

| Question 14  | 3 pts |
|--|-------|
| You have been hired as a machine learning consultant by a local company build a classifier whether a customer who received an email promotion will purchase or not. What can you say about this problem? |       |
| It is a regression problem   |       |
| It is a binary classification problem  |       |
| It is a multi-class classification problem   |       |
|  |       |

Question 15

Which of the following statements are true about the least mean square regression?

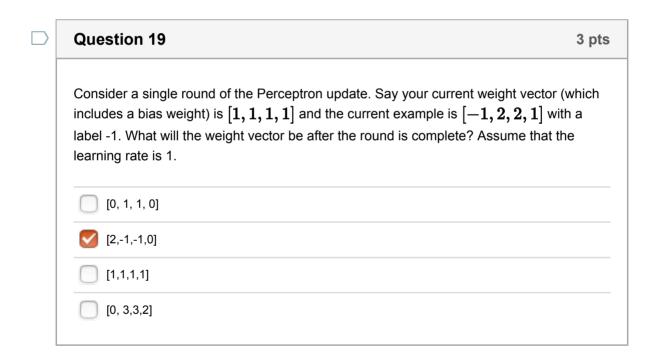
|              | Stochastic gradient descent will converge to a better optimum than gradient descent                |
|--------------|--|
| $\checkmark$ | The weights can be obtained analytically without requiring an optimization algorithm.              |
|              | Gradient descent will never converge to the optimum weights  |
|              | Gradient descent can eventually converge to the optimum weights if the algorithm runs long enough. |

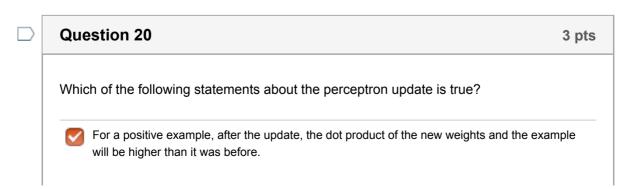
| Question 16  | 3 pt |
|--|------|
|  |      |
| How many mistakes will the Halving algorithm make on disjunctions with n Boolean |      |
| features where every variable has to be negated?                                 |      |
| $\bigcirc$ $O(2^n)$  |      |
|  |      |
| $\bigcirc O(\log n)$   |      |
| $\bigcirc$ $O(n^2)$  |      |
| $\bigcirc$ $O(n)$  |      |

| Question 17   | 3 pts   |
|---|---|
| How many k-disjunctions are possible with n Boolean variables if we do not allow negations? |   |
| $\binom{n}{k}$  |   |
| $\bigcirc 2^k \binom{n}{k}$   |   |
| $\bigcirc$ $3^k$  |   |
|   |   |
|   | How many k-disjunctions are possible with n Boolean variables if we do not allow negations? |

| $\bigcirc$ 2 <sup>k</sup> |  |  |  |
|---------------------------|--|--|--|
|                           |  |  |  |
|                           |  |  |  |

| Question 18  | 3 pts |
|--|-------|
| Which of the following statements is true about the original Perceptron algorithm? |       |
| It is an online algorithm  |       |
| ✓ It learns a linear classifier  |       |
| It is a batch algorithm  |       |
| It is a mistake bound algorithm  |       |





|                | For a negative example, after the update, the dot product of the new weights and the example will be higher than it was before. |
|----------------|---|
|                | For a positive example, after the update, the dot product of the new weights and the example will be lower than it was before.  |
| $ \checkmark $ | For a negative example, after the update, the dot product of the new weights and the example will be lower than it was before.  |

| Question 21  | 3 pts |
|--|-------|
| Consider the Boolean conjunction with two input features that is represented by th following data set: | е     |
| x1 x2 y  |       |
| 0 0 0  |       |
| 0 1 0  |       |
| 1 0 0  |       |
| 1 1 1  |       |
| What is the margin of this data set?   |       |
| $\frac{1}{2}$  |       |
| $\bigcirc$ $\sqrt{2}$  |       |
| <u> </u>   |       |
| $\frac{1}{2\sqrt{2}}$  |       |
|  |       |

| Question 22  | 3 pts |
|--|-------|
| According to the Perceptron mistake bound, what is the maximum number mistake that the Perceptron algorithm make on a disjunction in n dimensions? | S     |

|             | $\bigcirc$ $O(n^2)$   |         |
|-------------|---|---------|
|             | $lacksquare O(2^n)$   |         |
|             | $\bigcap O(\log(n))$  |         |
|             | 1   |         |
| )<br>       | Question 23   | 3 pts   |
|             | Which of the following assumptions are used to formalize the PAC model of learnal       | oility? |
|             | Examples are presented in a sequence to the learning algorithm                          |         |
|             | ✓ Training examples are drawn independently of each other                               |         |
|             | Training examples could be generated by an adversary                                    |         |
|             | Future examples will be drawn from the same fixed distribution as the training examples |         |
| <b>&gt;</b> | Question 24   | 3 pts   |
|             | Which of the following statements are true?   |         |
|             | The Halving algorithm can be used to learn a linear classifier.                         |         |
|             | The Halving algorithm gives the best possible mistake bound for all Boolean functions.  |         |
|             | No Boolean function can be learned under the mistake bound model.                       |         |
|             | The mistake bound model is only applicable for linear classifiers                       |         |
| $\supset$   | Question 25   | 3 pts   |
|             |   |         |

| Consider the fo   | ollowing dataset with four features (x1, x2, x3, x4) and a label y: |
|-------------------|---|
| x1 x2 x3 x4 y     |   |
| 1 0 1 1 0         |   |
| 1 1 0 0 1         |   |
| 0 0 0 1 0         |   |
| 1 1 1 1 1         |   |
|                   | has the highest information gain?                                   |
|                   | has the highest information gain?                                   |
| Which feature     | has the highest information gain?                                   |
| Which feature  x2 | has the highest information gain?                                   |

Quiz saved at 5:26pm Submit Quiz