Homework 3

Due Oct 3 at 11:59pm **Points** 75 **Questions** 25
Available Sep 26 at 11:59pm - Oct 4 at 11:59pm 8 days **Time Limit** None **Allowed Attempts** 2

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.

Attempt History

	Attempt	Time	Score
KEPT	Attempt 2	5 minutes	75 out of 75
LATEST	Attempt 2	5 minutes	75 out of 75
	Attempt 1	16 minutes	71.5 out of 75

Score for this attempt: **75** out of 75
Submitted Oct 2 at 6:08pm
This attempt took 5 minutes.

Question 1

3 / 3 pts

Suppose we have a binary classification problem with n features. Each feature in our problem can take one of three values A, B or C. How many binary classifiers are possible over this feature space?

- 2^n
- 3^n
- 3^{2^n}
- Correct! 2^{3^n}

Question 2

3 / 3 pts

How many disjunctions are possible with n Boolean features if we do not allow any negations?

- [ ] n
- [x] 2^n — Correct!
- [ ] n^2
- [ ] 2^{2^n}

Question 3

3 / 3 pts

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 Boolean function can be represented as a m-of-n function
- [x] Every disjunction without negations can be represented as a m-of-n function — Correct!
- [x] m-of-n functions can be represented by linear classifiers — Correct!
- [x] Every conjunction without negations can be represented as a m-of-n function — Correct!

Question 4

3 / 3 pts

Which of the following statements about decision trees are correct?

- [ ] Decision trees represent only linearly separable functions
- [ ] Every Boolean function can be represented by a unique decision tree
- [x] Every Boolean function can be represented as a decision tree

Correct!

- [x] Real valued features have to be discretized to use them with decision trees

Correct!

Question 5

3 / 3 pts

Suppose we know that

$$P(X = A) = \frac{1}{16},$$
$$P(X = B) = \frac{1}{16},$$
$$P(X = C) = \frac{1}{8},$$
$$P(X = D) = \frac{1}{4},$$
$$P(X = E) = \frac{1}{2}$$

Select all statements that are correct.

- [ ] These probabilities give the maximum possible value of the entropy of X
- [x] Entropy(X) = 1.875

Correct!

- [ ] These probabilities give the minimum possible value of the entropy of X
- [ ] Entropy(X) = 1.0

Question 6

3 / 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.

Correct! ☑ It will always find a decision tree that will fit any training set.

Correct! ☑ It is a batch algorithm.

Question 7

3 / 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:

x1	x2	x3	y
0	0	0	B
0	1	1	A
1	1	0	B
1	0	1	B

Correct! ☑ The entropy of the label is $2 - \frac{3}{4}\log_2 3$

☐ This data is not linearly separable.

☐ The entropy of the label is $2 + \frac{3}{4}\log_2 3$

Correct! ☑ This data is linearly separable

Question 8

3 / 3 pts

If your training data has a missing feature value, which of the following approaches can be used to handle it in the ID3 algorithm?

☐ Discard the training example because we can't use it for training.

Correct! ☑ Use fractional feature values representing the proportion of training examples that take each value.

Correct!

Use the most common value of that feature among the other examples that share the same label.

Correct!

Use the most common value of that feature among the other examples.

Question 9

3 / 3 pts

A learning algorithm is said to overfit its training data if:

Its hypothesis space contains the true concept function

Correct!

Its training error is less than its generalization error

Its hypothesis space is too small to express the data.

Its training error is more than its generalization error.

Question 10

3 / 3 pts

Use the following data with features x1, x2 and labels y and select all statements that are correct.

x1	x2	y
0	0	1
0	1	-1
1	0	-1
1	1	1

Correct!

This function can not be represented by a linear threshold unit.

This function can not be represented by a decision tree.

Correct!

This function can be represented by a decision tree.

This function can be represented by a linear threshold unit.

Question 11

3 / 3 pts

Which of the following Boolean functions with variables x_1, x_2, x_3, x_4 are linearly separable?

Correct!

- [x] $x_1 \vee \neg x_2$

Correct!

- [x] $x_1 \wedge \neg x_2$

Correct!

- [x] Label is true if any two out of x_1, x_2 or x_4 are true
- [ ] Label is true when an even number of x's are true.

Question 12

3 / 3 pts

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

- [ ] Delete examples where the algorithm makes mistakes and try again.
- [ ] Run multiple epochs over shuffled versions of the data.
- [ ] Nothing will help.

Correct!

- [x] Transform the data using a non-linear feature transformation.

Question 13

3 / 3 pts

Which of the following linear threshold units is equivalent to the following Boolean function: $x_1 \vee x_2 \vee x_3$?

- [ ] $sgn(-x_1 - x_2 - x_3 + 1)$

- [ ] $sgn(x_1 + x_2 + x_3 + 1)$

Correct!

- [x] $sgn(x_1 + x_2 + x_3 - 1)$
- [ ] $sgn(-x_1 - x_2 - x_3 - 1)$

Question 14

3 / 3 pts

You have been hired as a machine learning consultant by a local company. You have to build a classifier whether a customer who received an email promotion will make a purchase or not. What can you say about this problem?

- [ ] It is a regression problem

Correct!

- [x] It is a binary classification problem
- [ ] There is not enough information yet.
- [ ] It is a multi-class classification problem

Question 15

3 / 3 pts

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

Correct!

- [x] Gradient descent can eventually converge to the optimum weights if the algorithm runs long enough.
- [ ] Gradient descent will never converge to the optimum weights

Correct!

- [x] The weights can be obtained analytically without requiring an optimization algorithm.
- [ ] Stochastic gradient descent will converge to a better optimum than gradient descent

Question 16

3 / 3 pts

How many mistakes will the Halving algorithm make on disjunctions with n Boolean features where *every* variable has to be negated?

- $O(2^n)$
- $O(n^2)$
- $O(\log n)$
- Correct! $O(n)$

Question 17

3 / 3 pts

How many k-disjunctions are possible with n Boolean variables if we do not allow negations?

- Correct! $\binom{n}{k}$
- 3^k
- $2^k \binom{n}{k}$
- 2^k

Question 18

3 / 3 pts

Which of the following statements is true about the original Perceptron algorithm?

Correct! ☑ It is a mistake bound algorithm

Correct! ☑ It is an online algorithm

☐ It is a batch algorithm

Correct! ☑ It learns a linear classifier

Question 19

3 / 3 pts

Consider a single round of the Perceptron update. Say your current weight vector (which includes a bias weight) is $[1, 1, 1, 1]$ and the current example is $[-1, 2, 2, 1]$ with a label -1. What will the weight vector be after the round is complete? Assume that the learning rate is 1.

☐ [1,1,1,1]

☐ [0, 3,3,2]

☐ [0, 1, 1, 0]

Correct! ☑ [2,-1,-1,0]

Question 20

3 / 3 pts

Which of the following statements about the perceptron update is true?

Correct! ☑ For a positive 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.

Correct!

For a negative example, after the update, the dot product of the new weights and the example will be lower than it was before.

For a negative example, after the update, the dot product of the new weights and the example will be higher than it was before.

Question 21

3 / 3 pts

Consider the Boolean disjunction with two input features that is represented by the following data set:

x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

What is the margin of this data set?

1

Correct!

$\frac{1}{2\sqrt{2}}$

$\sqrt{2}$

$\frac{1}{2}$

Question 22

3 / 3 pts

According to the Perceptron mistake bound, what is the maximum number mistakes that the Perceptron algorithm make on a disjunction in n dimensions?

Correct!

- [x] $O(n^2)$
- [ ] $O(2^n)$
- [ ] $O(\log(n))$
- [ ] 1

Question 23

3 / 3 pts

Which of the following assumptions are used to formalize the PAC model of learnability?

- [ ] Training examples could be generated by an adversary

Correct!

- [x] Future examples will be drawn from the same fixed distribution as the training examples
- [ ] Examples are presented in a sequence to the learning algorithm

Correct!

- [x] Training examples are drawn independently of each other

Question 24

3 / 3 pts

Which of the following statements are true?

Correct!

- [x] The Halving algorithm can be used to learn a linear classifier.

Correct!

- [x] 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

Question 25

3 / 3 pts

Consider the following 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

Which feature has the highest information gain?

Correct!

- [x] x2
- [ ] x3
- [ ] x4
- [ ] x1

Quiz Score: **75** out of 75