Started on	Friday, 18 May 2018, 1:06 PM
State	Finished
Completed on	Friday, 18 May 2018, 1:34 PM
Time taken	28 mins 15 secs
Grade	**11.0** out of 11.0 (**100**%)
Feedback	Well done!

Question 1

Correct

Mark 1.0 out of 1.0

Sampling is called *with replacement* when a unit selected at random from the population(dataset) is not returned to the population, and then, the selected unit cannot be selected again in a further sampling process.

Select one:

- True
- False ✓

The correct answer is 'False'.

Correct

Marks for this submission: 1.0/1.0.

Question 2

Correct

Mark 1.0 out of 1.0

Which of the following statements are **correct**?

Select one:

- a. In AdaBoost, each tuple in training dataset D_i is sampled uniformly from the original dataset D with replacement.
- b. In Bagging, each tuple in training dataset D_i is sampled uniformly from the original dataset D with replacement. ✓
- c. Bagging assigns a different weight to each training tuple.
- d. AdaBoost takes a majority voting approach to predict a class of unknown tuple

Your answer is correct.

The correct answer is: In Bagging, each tuple in training dataset D_i is sampled uniformly from the original dataset D with replacement.

Correct

Marks for this submission: 1.0/1.0.

Question 3

Correct

Mark 1.0 out of 1.0

The following table shows the weights of five items in dataset D after $i-1$ iterations of AdaBoost.

	x_1	x_2	x_3	x_4	x_5
weight (w_j)	0.2	0.1	0.15	0.15	0.4

(Note that the sum of weights is 1)

To train the i-th classifier, we sampled 5 tuples with replacement from the dataset D and obtained the following training dataset D_i:

$D_i = (x_1, x_1, x_2, x_2, x_5)$

We trained some classification model M_i using D_i and the classifier correctly classified x_1 and x_2 and misclassified x_5.

What will be the weight of x_5 after adjusting weights and normalisation based on the classification result?

(Note that $err(M_i) = 0.4$ in this case)

Write your answer to two decimal places.

Answer: 0.43

Weights of x_1 and x_2 before normalisation are

$w_1 = 0.2 \times (2/3)$

$w_1 = 0.1 \times (2/3)$

The correct answer is: 0.44

Correct

Marks for this submission: 1.0/1.0.

Question 4

Correct

Mark 1.0 out of 1.0

Choose answers that correctly state the characteristics of data stream.

Select one or more:

- [x] a. Random access to a data tuple is expensive. Linear or sublinear computational techniques are required to analyse the stream. ✓
- [x] b. Some level of approximation is acceptable in data stream analysis. ✓
- [x] c. In general, it is not feasible to store every data tuple. To overcome the storage problem, we often store summaries of data seen so far. ✓
- [ ] d. Every data item needs to be processed in data stream analysis.

Your answer is correct.

The correct answers are: Random access to a data tuple is expensive. Linear or sublinear computational techniques are required to analyse the stream., In general, it is not feasible to store every data tuple. To overcome the storage problem, we often store summaries of data seen so far., Some level of approximation is acceptable in data stream analysis.

Correct

Marks for this submission: 1.0/1.0.

Question 5

Correct

Mark 1.0 out of 1.0

Match the correct statement which describes a given method for a data stream.

Histogram ✓	can be used to approximate the frequency distribution of element values
Sliding window ✓	makes a decision on some recent data
Random sampling ✓	refers to the process of probabilistic choice of a data item to be processedor not

Your answer is correct.

The correct answer is: Histogram → can be used to approximate the frequency distribution of element values, Sliding window → makes a decision on some recent data, Random sampling → refers to the process of probabilistic choice of a data item to be processedor not

Correct

Marks for this submission: 1.0/1.0.

Question 6

Correct

Mark 1.0 out of 1.0

Multidimensional OLAP analysis is still needed in stream data analysis. However, due to the limited memory, disk space, and processing power, it is impossible to store the detailed level of data and compute a fully materialised cube. Several methods have been proposed to obtain an efficient OLAP with data stream.

tilted time frame ✓ method compress the time dimension of the data using different granularity. This approach focuses on fact that 1) the most recent time is registered at the finest granularity, and 2) most distance time is registered at a coarser granularity. For example, natural tilted time frame model ✓ structures time frames in multiple granularities based on the natural time scale, whereas logarithmic tilted time frame model ✓ structures time frames in multiple granularities according to a logarithmic scale.

The above approach can only be applied to a time dimension of the data, and materialisation of all possible cubes are still too costly.

critical layers ✓ approach tries to mitigate by dynamically computing and storing two critical cuboids. The upper layer, called the observation layer ✓, is the layer at which an analyst (or an automated system) would like to continuously study the data. The bottom layer, called the minimal interest layer ✓, is the minimally interesting layer that an analyst would like to study.

Your answer is correct.

The correct answer is:
Multidimensional OLAP analysis is still needed in stream data analysis. However, due to the limited memory, disk space, and processing power, it is impossible to store the detailed level of data and compute a fully materialised cube. Several methods have been proposed to obtain an efficient OLAP with data stream.

[tilted time frame] method compress the time dimension of the data using different granularity. This approach focuses on fact that 1) the most recent time is registered at the finest granularity, and 2) most distance time is registered at a coarser granularity. For example, [natural tilted time frame model] structures time frames in multiple granularities based on the natural time scale, whereas [logarithmic tilted time frame model] structures time frames in multiple granularities according to a logarithmic scale.

The above approach can only be applied to a time dimension of the data, and materialisation of all possible cubes are still too costly. [critical layers] approach tries to mitigate by dynamically computing and storing two critical cuboids. The upper layer,

called the [observation layer], is the layer at which an analyst (or an automated system) would like to continuously study the data. The bottom layer, called the [minimal interest layer], is the minimally interesting layer that an analyst would like to study.

Correct
Marks for this submission: 1.0/1.0.

Question 7

Correct

Mark 1.0 out of 1.0

Lossy Counting Algorithm

Let's say we'd like to estimate frequencies of items in a stream using lossy counting algorithm.

We set our margin of error, ϵ to be 0.1 and started to count frequencies.

After a while, we decide to estimate the frequency of item e, so we search the corresponding item our data structure D and find an entry whose value is $(e, 3, 9)$.

What will be the maximum frequency of item e?

Answer: 12

this entry is inserted at 10th bucket and after the insertion the item appears two more times in the stream.

The correct answer is: 12

Correct
Marks for this submission: 1.0/1.0.

Question 8

Correct

Mark 1.0 out of 1.0

What will be the minimum frequency of item e?

Answer: 3

The correct answer is: 3

Correct
Marks for this submission: 1.0/1.0.

Question 9

Correct

Mark 1.0 out of 1.0

Choose the correct description of time series characteristic

Cyclic movements or cyclic variations ✓	Long-term oscillations about a trend line, which may or may not be periodic
Trend or long-term movements ✓	a general direction in which a time-series is moving over a long interval of time
Irregular or random movements ✓	sporadic motion of time series due to random or chance events, such as labor dis
Seasonal movements or seasonal variations ✓	systematic or calendar related movements

Your answer is correct.

The correct answer is: Cyclic movements or cyclic variations → Long-term oscillations about a trend line, which may or may not be periodic, Trend or long-term movements → a general direction in which a time-series is moving over a long interval of time, Irregular or random movements → sporadic motion of time series due to random or chance events, such as labor disputes or floods, Seasonal movements or seasonal variations → systematic or calendar related movements

Correct

Marks for this submission: 1.0/1.0.

Question 10

Correct

Mark 1.0 out of 1.0

Given sequence

2, 4, 6, 6, 4, 2, 0, 4

What is the sequence of the moving average of order 4

Answer: 4.5,5,4.5,3,2.5 ✓

The correct answer is: 4.5 5 4.5 3 2.5

Correct

Marks for this submission: 1.0/1.0.

Question 11

Correct

Mark 1.0 out of 1.0

Given the same sequence, what is the sequence of the weighted moving average of order 2 with weights (2, 3)?

Answer: 3.2, 5.2, 6, 4.8, 2.8, 0.8, 2.4

The correct answer is: 3.2 5.2 6 4.8 2.8 0.8 2.4

Correct

Marks for this submission: 1.0/1.0.