SM Quiz 1

Due No due date **Points** 14 **Questions** 7
Available Nov 2 at 9:25am - Nov 2 at 10am 35 minutes **Time Limit** 20 Minutes

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	16 minutes	9 out of 14

ⓘ Correct answers are hidden.

Score for this quiz: **9** out of 14
Submitted Nov 2 at 9:42am
This attempt took 16 minutes.

Question 1

2 / 2 pts

SM-SI.009. **Decision trees**…

☐ … split the nodes along attributes that has the lowest impurity

☐ … are in every case hard to construct manually, and always hard to interpret

☐ … create a tree structure, that is always balanced, that is each leaves are in the same distance from the root nodes.

☑ … use measurements of impurity, such as Gini and entropy

☑ … split the nodes along attributes that provides the biggest decrease of impurity

Partial

Question 2

1 / 2 pts

SM-SI.012. We want to monitor, that users of a search engine how many times issue the same query more than once (2 times for simplicity). We have a storage space limited to approximately 1/10 of the expected number of queries. Which **sampling method** can help us achieving this?

☐ For each incoming query we generate a uniformly random number between 0 and 1 and if it is less then 0.1, then we store the query. At the end, to get a good approximation, enough to calculate how many, issued by the same user occurs twice.

☑ We store all the queries of 1/10 of the users. At the end, to get a good approximation, enough to calculate how many, issued by the same user occurs twice.

☐ We hash the user with a hash function, that has 10 buckets, and store the query and the user if the user is mapped into a predefined bucket. At the end, to get a good approximation, enough to calculate how many, issued by the same user occurs twice.

☐ For each incoming query we generate a uniformly random number between 0 and 1 and if it is greater then 0.9, then we store the query. At the end, to get a good approximation, enough to calculate how many, issued by the same user occurs twice.

Incorrect

Question 3

0 / 2 pts

SM-SI.004. Mark the most important **challenges** in stream mining:

☑ It might only be possible to perform a single pass during stream analysis

☐ Streams are complete

☐ Streams adhere to a common statistical deviation

☐ It might not be possible to store data stream history

☑ Streams are unbounded in nature

Question 4

2 / 2 pts

SM-SI.002. The **Bloom filter**…

☑ …relies on the use of multiple hash functions

□ …never produces true positives

□ …is always 100% precise

□ …is not used in the context of stream mining

□ …never makes false positives

Question 5

2 / 2 pts

SM-SI.001**. Sampling** in the context of stream mining…

☑ …is often based on fixed fraction sampling

□ …relies on the use of data stream shards

□ …always relies on the use of data stream partitions

☑ …is often based on the creation of a representative sample

Question 6

2 / 2 pts

SM-SI.006**. Cluster features**…

☑ … are used in stream clustering

□ … can be relevant in the context of anomaly detection

□ … consist of data points, linear sums and big data

☑ … consist of data points, linear sums and sums of squares

□ … are used in stream classification

Incorrect

Question 7

0 / 2 pts

SM-SI.007. The **Count Min Sketch** (a table that contains multiple arrayS and uses multiple hash functions)

☑ … is used to count the number of occurrences of the different elements (aaabbbc ->a:3, b:3,c:1)

□ … is used to count how many different elements is in the stream (aaabbbc ->3)

□ … gives us either the correct answer or an underestimation, but never overestimate.

□ … gives us an overestimation of the right answer

☑ … is mapping each incoming elements to each array using all hash functions.

Quiz Score: **9** out of 14