DS-200^Q&As

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.geekcert.com/ds-200.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Cloudera Official Exam Center

Instant Download After Purchase

100% Money Back Guarantee

365 Days Free Update

800,000+ Satisfied Customers

QUESTION 1

How can the naiveté of the naive Bayes classifier be advantageous?

A. It does not require you to make strong assumptions about the data because it is a non- parametric

B. It significantly reduces the size of the parameter space, thus reducing the risk of over fitting

C. It allows you to reduce bias with no tradeoff in variance

D. It guarantees convergence of the estimator

Correct Answer: A

QUESTION 2

You have user profile records in an OLTP database that you want to join with web server logs which you have already ingested into HDFS. What is the best way to acquire the user profile for use in HDFS?

A. Ingest with Hadoop streaming

B. Ingest with Apache Flume

C. Ingest using Hive\\'s LOAD DATA command

D. Ingest using Sqoop

E. Ingest using Pig\\'s LOAD command

Correct Answer: BD

Reference: https://thinkbiganalytics.com/leading_big_data_technologies/ingestion-and- streaming-withstorm-kafka-flume/

QUESTION 3

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two

functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M)

generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

For line in file: If random_uniform ()

Method B

i = 0

for line in file:

if i % 10 = = 0;

print line

i += 1

Method C

idxs = random_permotation (N) [: (N/10)]

i = 0

for line in file:

if i in idxs:

print line

i +=1

Which method might introduce unexpected correlations?

A. Method A

B. Method B

C. Method C

Correct Answer: C

QUESTION 4

From historical data, you know that 50% of students who take Cloudera\'s Introduction to Data Science: Building Recommenders Systems training course pass this exam, while only 25% of students who did not take the training course pass this exam. You also know that 50% of this exam\'s candidates also take Cloudera\'s Introduction to Data Science: Building Recommendations Systems training course.

If we know that a person has passed this exam, what is the probability that they took cloudera\'s introduction to Data Science: Building Recommender Systems training course?

A. 2/3

B. 1/2

C. 3/4

D. 3/5

Correct Answer: B

QUESTION 5

Which two techniques should you use to avoid overfitting a classification model to a data set?

A. Include a small number "noise" features that are not through to be correlated with the dependent variable.

B. Replicate features that are through to be significant predicators of the dependent variable multiple time for each observation.

C. Separate your input data into a training set that is used for fitting and a test set that is used for evaluating the model\\'s performance

D. Include a regularization term in the model\\'s objective function to control how precisely the model fits the data

E. Preprocess the data to exclude a typical observation from the model input

Correct Answer: AE