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.passapply.com/ds-200.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

There are 20 patients with acute lymphoblastic leukemia (ALL) and 32 patients with acute myeloid leukemia (AML), both variants of a blood cancer.

The makeup of the groups as follows:

ALL GROUP			
	Male	Female	
Caucasian	14	1	15
Asian-American	5	0	5
	19	1	20

AML GROUP			
	Male	Female	
Caucasian	9	4	13
Asian-American	7	12	19
	16	16	32

Each individual has an expression value for each of 10000 different genes. The expression value for each gene is a continuous value between -1 and 1.

You\\'ve built your model for discriminating between AML and ALL patients and you find that it works quite well on your current data. One month later, a collaboration tells you she has fresh data from 100 new AML/ ALL patients. You run the samples through your model, and turns out your model has very poor predictive accuracy on the new samples; specifically, your model predicts that all males have ALL. What is the most reliable way to fix this problem?

A. Change the distance metric

B. Reduce the number of dimensions

C. Use a Gibbs sampler on a Bayesian network

D. Perform matched sampling across other provided variables

Correct Answer: D

QUESTION 2

Which best describes the primary function of Flume?

A. Flume is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with an infrastructure consisting of sources and sinks for importing and evaluating large data sets

B. Flume acts as a Hadoop filesystem for log files

C. Flume Imports data from SQL/relational database into your Hadoop cluster

D. Flume provides a query languages for Hadoop similar to SQL

E. Flume is a distributed server for collecting and moving large amount of data into HDFS as it\\'s produced from streaming data flows

Correct Answer: D

QUESTION 3

What are three benefits of running feature selection analysis before filtering a classification model?

A. Eliminates the need to include a regularization term

B. Reduces the number of subjective decisions required to construct the model

C. Guarantee the optimally of the final model

D. Speeds up the model fitting process

E. Develops an understanding of the importance of different features

F. Improves the predictive performance of the model

Correct Answer: DEF

QUESTION 4

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 will have the best runtime performance?

A. Method A

B. Method B

C. Method C

Correct Answer: A

QUESTION 5

What is the most common reason for a k-means clustering algorithm to returns a sub-optimal clustering of its input?

A. Non-negative values for the distance function

B. Input data set is too large

C. Non-normal distribution of the input data

D. Poor selection of the initial controls

Correct Answer: C