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.leads4pass.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

You\\'ve built a model that has ten different variables with complicated independence relationships between them, and both continuous and discrete variables that have complicated, multi-parameter distributions. Computing the joint probability distribution is complex, but it turns out that computing the conditional probabilities for the variables is easy. What is the most computationally efficient for computing the expected value?

A. Method of moments

B. Markov Chain Monte Carlo

C. Gibbs sampling

D. Numerical quadrature

Correct Answer: B

QUESTION 2

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.

With which type of plot can you encode the most amount of the data visually?

Rather than use all 10,000 features to separate AML from ALL, you pick a small subnet of features to

separate them optimally. You feature vectors have 10,000 dimensions while you only have 52 data points. You use

cross-validation to test your chosen set of features. What three methods will choose the features in an optimal way?

A. Singular value Decomposition

B. Bootstrapping

C. Markov chain Monte Carlo

D. Hidden Markov

E. Bayesian Information Criterion

F. Mutual Information

Correct Answer: CDF

QUESTION 3

Which recommender system technique is domain specific?

A. Content-based collaboration filtering

B. Item-based collaborative filtering

C. User-based collaborative filtering

D. Naïve Bayes classifier

Correct Answer: C

Reference: http://www.cs.cmu.edu/~srosenth/papers/Rosenthal_RecSys09.pdf

QUESTION 4

You want to understand more about how users browse your public website. For example, you war know which pages they visit prior to placing an order. You have a server farm of 200 web server hosting your website. Which is the most efficient process to gather these web servers access logs into your Hadoop cluster for analysis?

A. Sample the web server logs web servers and copy them into HDFS using curl

B. Channel these click streams into Hadoop using Hadoop Streaming

C. Write a MapReduce job with the web servers for mappers and the Hadoop cluster nodes for reducers

D. Import all user clicks from your OLTP databases Into Hadoop using Sqoop

E. Ingest the server web logs into HDFS using Flume

Correct Answer: C

QUESTION 5

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 is least likely to give you exactly 10% of your data?

A. Method A

B. Method B

C. Method C

Correct Answer: B