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.pass4itsure.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

A company has 20 software engineers working to fix on a project. Over the past week, the team has fixed 100 bugs. Although the average number of bugs. Although the average number of bugs fixed per engineer id five. None of the engineer fixed exactly five bugs last week.

You want to understand how productive each engineer is at fixing bugs. What is the best way to visualize the distribution of bug fixes per engineer?

A. A bar chart of engineers vs. number of bugs fixed

B. A scatter plot of engineers vs. number of bugs fixed

C. A normal distribution of the mean and standard deviation of bug fixes per engineer D. A histogram that groups engineers to together based on the number of bugs they fixed

Correct Answer: A

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

You have a large m x n data matrix M. You decide you want to perform dimension reduction/clustering on your data and have decide to use the singular value decomposition (SVD; also called principal components analysis PCA)

Refer to the passage above.

What represents the SVD of the Matrix standard M given the following information:

U is m x m unitary V is n x n unitary S is m x n diagonal Q is n x n invertible D is n x n diagonal L is m x m lower triangular U is m x m upper triangular

A. M = U S V

B. M = U P

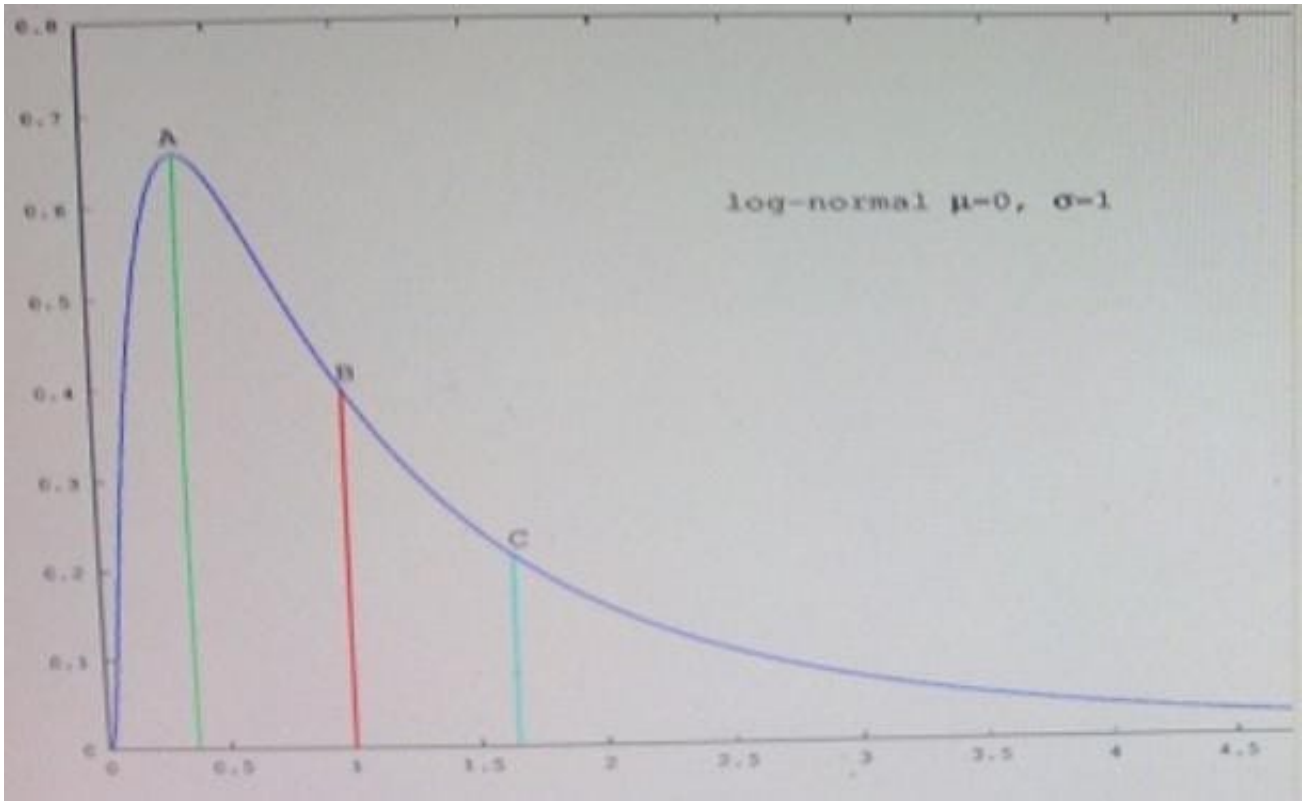
C. M = Q D Q-1

D. M = L U

Correct Answer: A

QUESTION 4

Refer to the exhibit.



Which point in the figure is the mean?

A. A

B. B

C. C

Correct Answer: B

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 might introduce unexpected correlations?

A. Method A

B. Method B

C. Method C

Correct Answer: C

QUESTION 6

You are building a k-nearest neighbor classifier (k-NN) on a labeled set of points in a high- dimensional space. You determine that the classifier has a large error on the training data. What is the most likely problem?

A. High-dimensional spaces effectively make local neighborhoods global

B. k-NN compotation does not coverage in high dimensions

C. k was too small

D. The VC-dimension of a k-NN classifier is too high

Correct Answer: B

QUESTION 7

You are building a system to perform outlier detection for a large online retailer. You need to build a system to detect if the total dollar value of sales are outside the norm for each U.S. state, as determined from the physical location of the buyer for each purchase. The retailer\\'s data sources are scattered across multiple systems and databases and are unorganized with little coordination or shared data or keys between the various data sources.

Below are the sources of data available to you. Determine which three will give you the smallest set of data sources but still allow you to implement the outlier detector by state.

A. Database of employees that Includes only the employee ID, start date, and department

B. Database of users that contains only their user ID, name, and a list of every Item the user has viewed

C. Transaction log that contains only basket ID, basket amount, time of sale completion, and a session ID

D. Database of user sessions that includes only session ID, corresponding user ID, and the corresponding IP address

E. External database mapping IP addresses to geographic locations

F. Database of items that includes only the item name, item ID, and warehouse location

G. Database of shipments that includes only the basket ID, shipment address, shipment date, and shipment method

Correct Answer: ADF

QUESTION 8

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

QUESTION 9

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

QUESTION 10

You need to analyze 60,000,000 images stored in JPEG format, each of which is approximately 25 KB. Because your Hadoop cluster isn\\'t optimized for storing and processing many small files you decide to do the following actions:

1.

Group the individual images into a set of larger files

2.

Use the set of larger files as input for a MapReduce job that processes them directly with Python using Hadoop streaming

Which data serialization system gives you the flexibility to do this?

A. CSV

B. XML

C. HTML

D. Avro

E. Sequence Files

F. JSON

Correct Answer: BF