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

In what format are web server log files usually generated and how must you transform them in order to make them usable for analysis in Hadoop?

A. XML files that you need to convert to JSON

B. Text files that require parsing into useful fields

C. CSV files that require parsing into useful fields

D. HTML files that you need to convert to plain text or CSV

E. Binary files that may require decompression and conversion using AVRO

Correct Answer: AB

QUESTION 2

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)

You performed singular value decomposition (SVD; also called principal components analysis or PCA) on you data matrix but you did not center your data first. What does your first singular component describe?

A. The mean of the data set

B. The variance of the data set

C. The standard deviation of the data set

D. The maximum of the data set

E. The median of the data set

Correct Answer: C

QUESTION 3

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

A. Method A

B. Method B

C. Method C

Correct Answer: B

QUESTION 5

Given the following sample of numbers from a distribution:

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89

How do high-level languages like Apache Hive and Apache Pig efficiently calculate approximately percentiles for a distribution?

A. They sort all of the input samples and the lookup the samples for each percentile

B. They maintain index of input data as it is loaded into HDFS and load them into memory

C. They use pivots to assign each observations to the reducer that calculate each percentile

D. They assign sample observations to buckets and then aggregate the buckets to compute the approximations

Correct Answer: C

QUESTION 6

You have a directory containing a number of comma-separated files. Each file has three columns and each filename has a .csv extension. You want to have a single tab-separated file (all .tsv) that contains all the rows from all the files.

Which command is guaranteed to produce the desired output if you have more than 20,000 files to process?

A. Find . name `*, CSV\\' print0 | sargs -0 cat | tr `,\\' `\t\\' > all.tsv

B. Find . name `name * .CSV\\' | cat | awk `BEGIN {FS = "," OFS = "\t"} {print $1, $2, $3}\\' > all.tsv

C. Find . name `*.CSV\\' | tr `,\\' `\t\\' | cat > all.tsv

D. Find . name `*.CSV\\' | cat > all.tsv

E. Cat *.CSV > all.tsv

Correct Answer: B

QUESTION 7

Which three metrics are useful in measuring the accuracy and quality of a recommender system?

A. Mutual Information

B. RMSF

C. Tanimoto coefficient

D. Pearson correlation

E. Precision

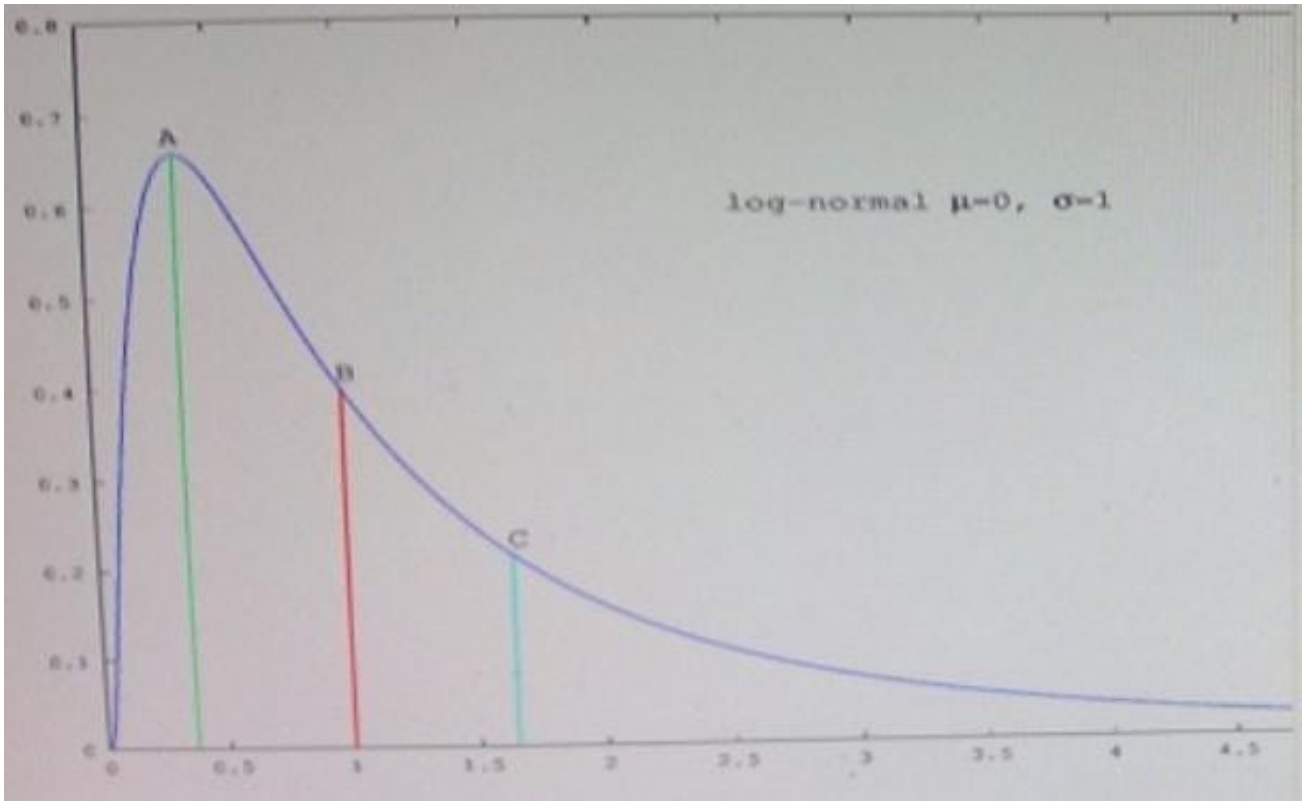
F. Recall

Correct Answer: CDE

Reference: https://lirias.kuleuven.be/bitstream/123456789/289803/3/datasets-cameraready.pdf

QUESTION 8

Refer to the exhibit.



Which point in the figure is the mean?

A. A

B. B

C. C

Correct Answer: B

QUESTION 9

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

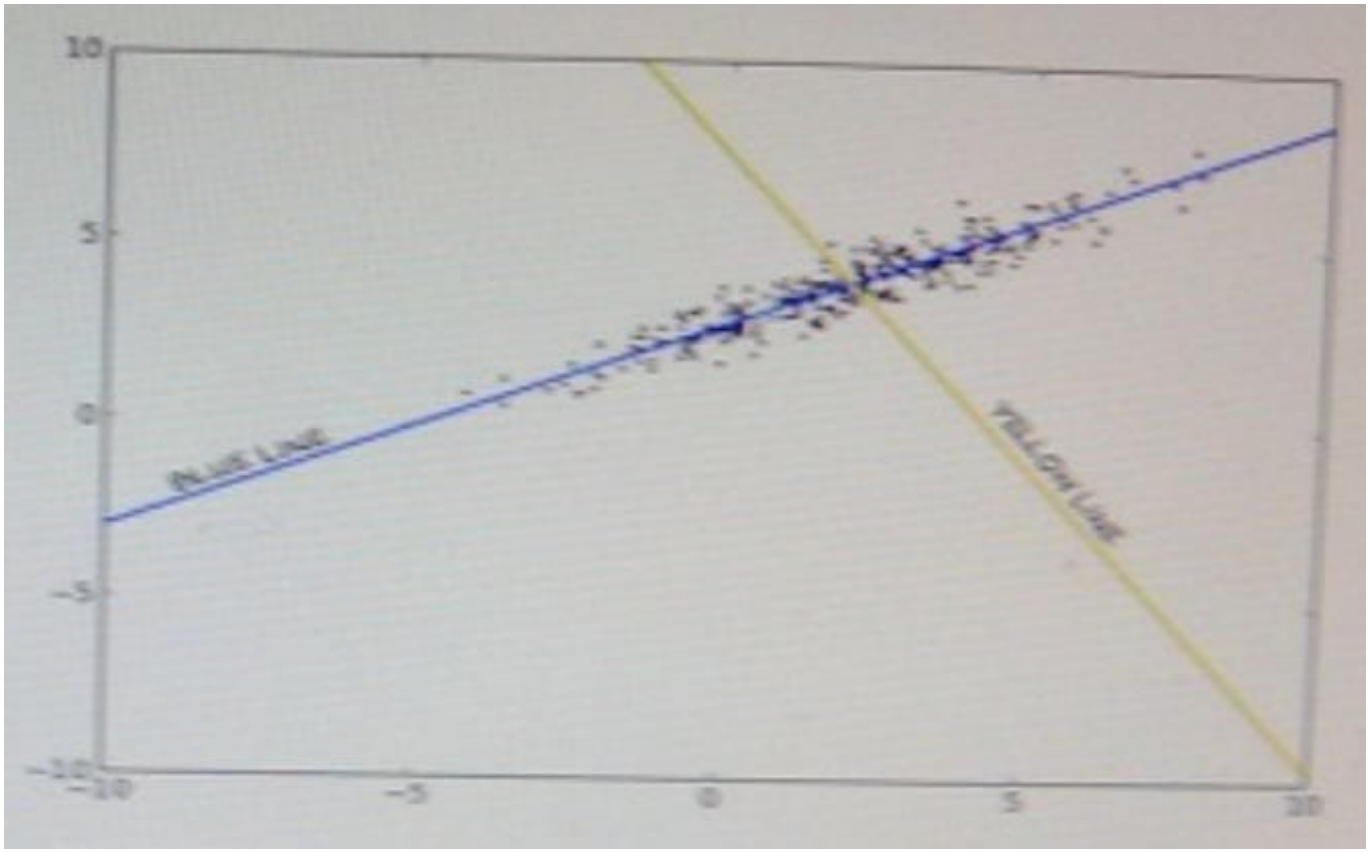
Correct Answer: CDF

QUESTION 10

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)

For the moment, assume that your data matrix M is 500 x 2. The figure below shows a plot of the data.



Which line represents the second principal component?

A. Blue

B. Yellow

Correct Answer: A