

Cloudera

Exam DS-200

Data Science Essentials

Version: 6.0

[Total Questions: 60]

Question No : 1

Why should stop an interactive machinelearningalgorithm assoon as the performanceof the model on a test set stops improving?

- A. To avoid the need for cross-validating the model
- B. To prevent overfitting
- C. To increase the VC (VAPNIK-Chervonenkis) dimension for the model
- D. To keep the number of terms in the model as possible
- E. To maintain the highest VC (Vapnik-Chervonenkis) dimension for the model

Answer: B

Question No : 2

What is default delimiterfor Hive tables?

- A. ^A (Control-A)
- B. , (comma)
- C. \t (tab)
- D. : (colon)

Answer: A

Reference:<http://blog.spryinc.com/2013/10/four-useful-tricks-for-working-with-hive.html>(change the delimiter when exporting hive table)

Question No : 3

Certain individuals are moresusceptibleto autismif they have particularcombinationsofgenesexpressed in their DNA. Givena sample of DNAfrom personswho have autismand a sample of DNAfrom persons who do not haveautism,determine the best technique forpredictingwhetheror nota given individualis susceptibleto developing autism?

- A. Native Bayes
- B. Linear Regression
- C. Survival analysis

D. Sequencealignment

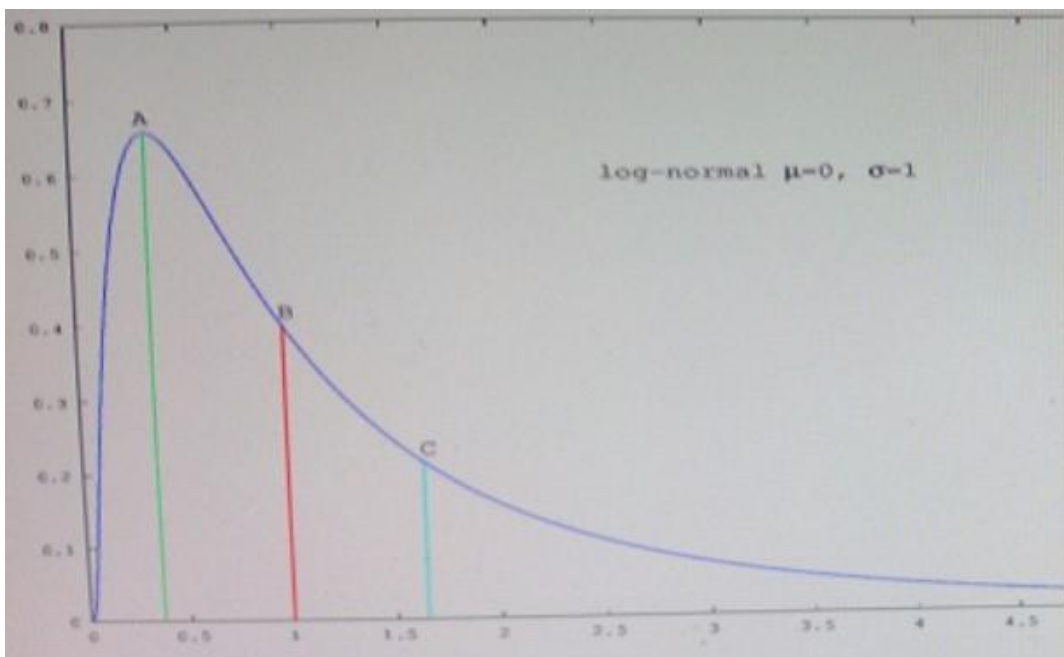
Answer: B**Question No : 4**

You are working with a logistic regression model to predict the probability that a user will click on an ad. Your model has hundreds of features, and you're not sure if all of those features are helping your prediction. Which regularization technique should you use to prune features that aren't contributing to the model?

- A. Convex
- B. Uniform
- C. L2
- D. L1

Answer: A**Question No : 5**

Refer to the exhibit.



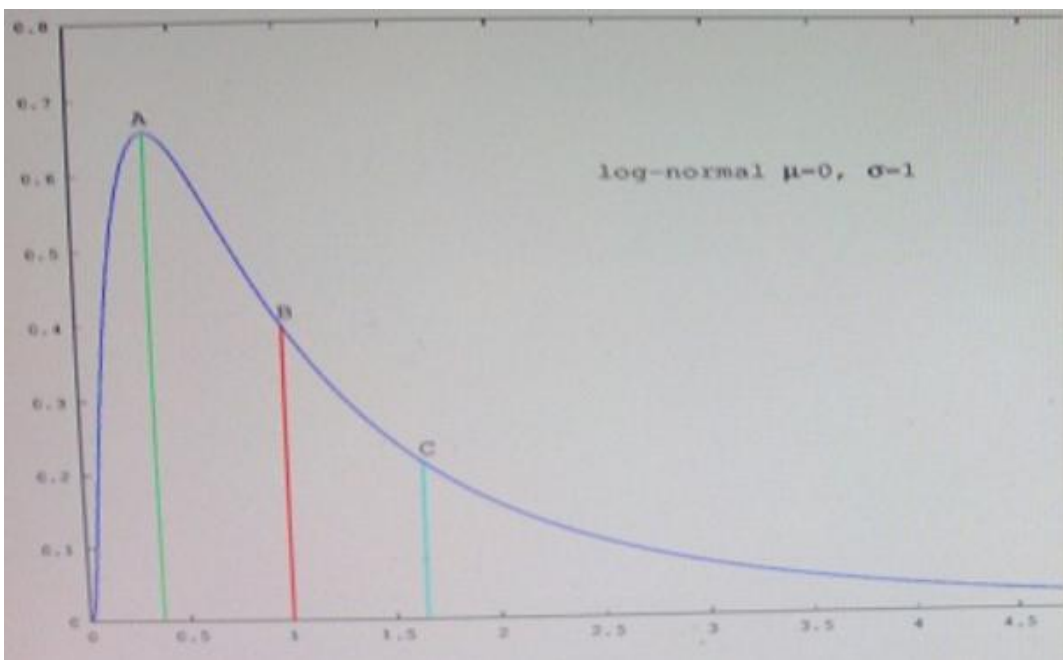
Which point in the figure is the median?

- A. A
- B. B
- C. C

Answer: A

Question No : 6

Refer to the exhibit.



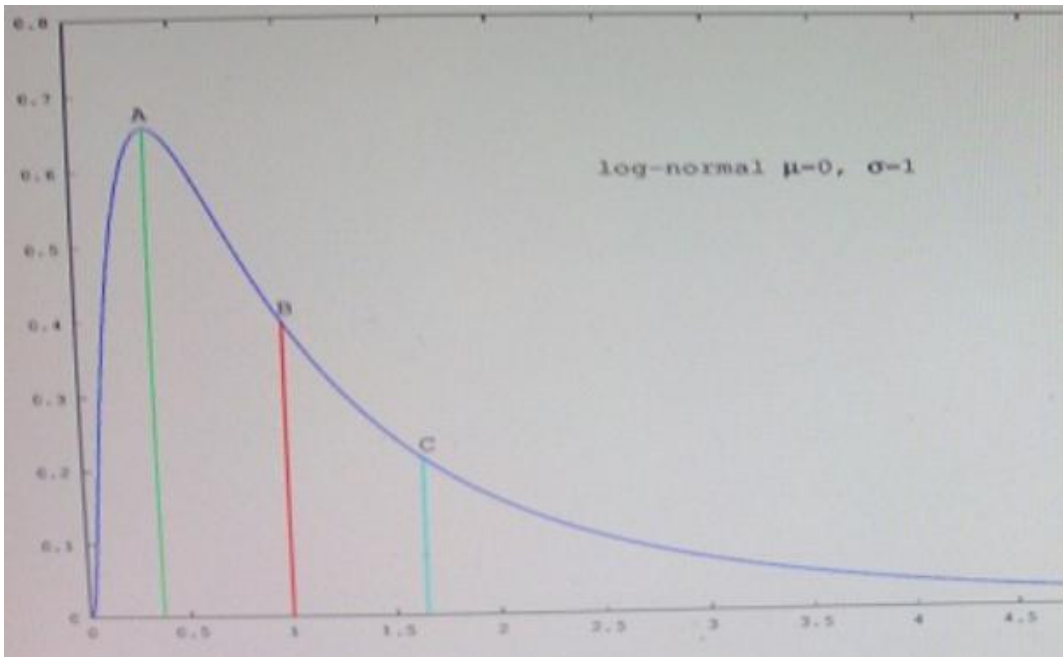
Which point in the figure is the mode?

- A. A
- B. B
- C. C

Answer: C

Question No : 7

Refer to the exhibit.



Which point in the figure is the mean?

- A. A
- B. B
- C. C

Answer: B

Question No : 8

Under what two conditions does stochastic gradient descent outperform 2nd-order optimization techniques such as iteratively reweighted least squares?

- A. When the volume of input data is so large and diverse that a 2nd-order optimization technique can be fit to a sample of the data
- B. When the model's estimates must be updated in real-time in order to account for new observations.
- C. When the input data can easily fit into memory on a single machine, but we want to calculate confidence intervals for all of the parameters in the model.
- D. When we are required to find the parameters that return the optimal value of the objective function.

Answer: A,B

Question No : 9

What is the result of the following command (the database username is foo and password is bar)?

```
$ sqoop list-tables - -connect jdbc:mysql://localhost/databasename - -table - -  
usernamefoo - -password bar
```

- A. sqoop lists only those tables in the specified MySQL database that have not already been imported into FDFS
- B. sqoop returns an error
- C. sqoop lists the available tables from the database
- D. sqoop imports all the tables from SQLHDFS

Answer: C

Reference: <https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-15/getting-sqoop>

Question No : 10

What is the most common reason for a k-means clustering algorithm to return 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

Answer: C

Question No : 11

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

Answer: D

Question No : 12

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