# IT认证电子书

质　量　更　高　　服　务　更　好

半年免费升级服务

http://www.itrenzheng.com

**Exam** : **DS-200**

**Title** : Data Science Essentials

**Version** : Demo

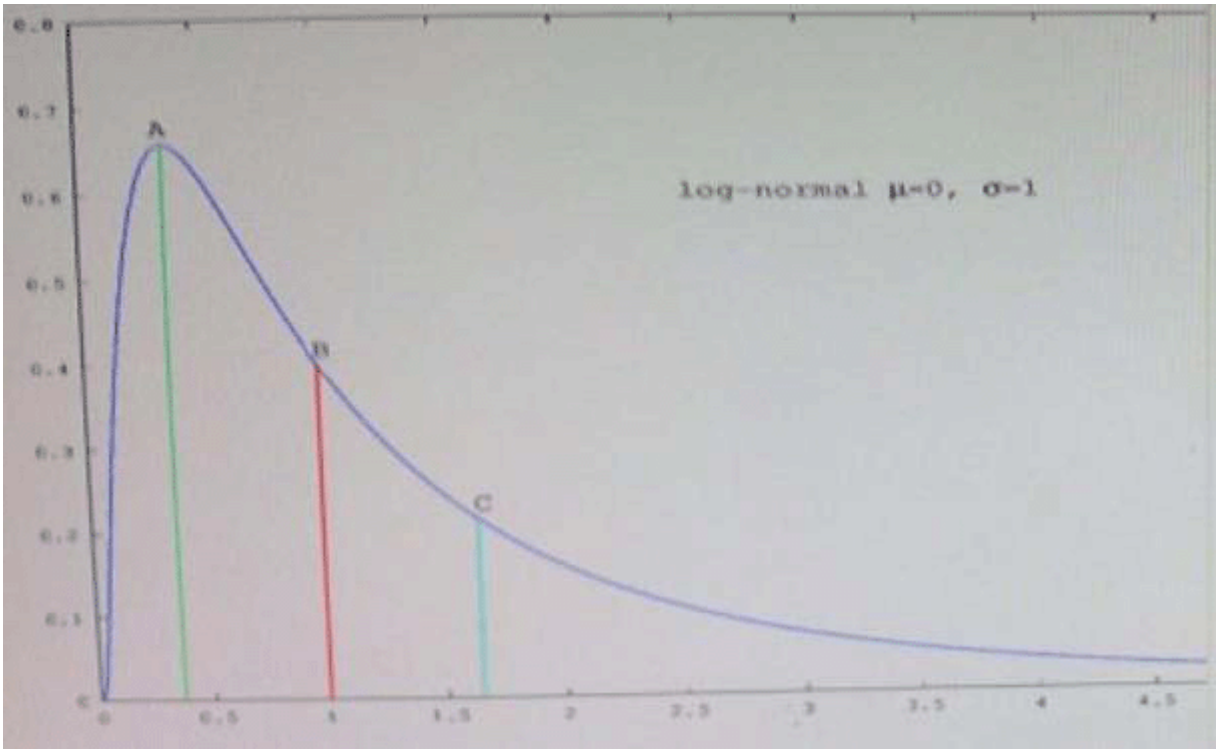
1.Why should stop an interactive machine learning algorithm as soon as the performance of the model on a test set stops improving?
A. To avoid the need for cross-validating the model
B. To prevent overfitting
C. To increase the VC (VAPNIK-Chervonenkis) dimension for the model
D. To keep the number of terms in the model as possible
E. To maintain the highest VC (Vapnik-Chervonenkis) dimension for the model
**Answer:** B

2.What is default delimiter for Hive tables?
A. ^A (Control-A)
B. , (comma)
C. \t (tab)
D. : (colon)
**Answer:** A
**Explanation:**
Reference:http://blog.spryinc.com/2013/10/four-useful-tricks-for-working-with-hive.html(change the delimiter when exporting hive table)

3.Certain individuals are more susceptible to autism if they have particular combinations of genes expressed in their DNA.
Given a sample of DNA from persons who have autism and a sample of DNA from persons who do not have autism, determine the best technique for predicting whether or not a given individual is susceptible to developing autism?
A. Native Bayes
B. Linear Regression
C. Survival analysis
D. Sequencealignment
**Answer:** B

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

5.Refer to the exhibit.




Which point in the figure is the median?

A. A

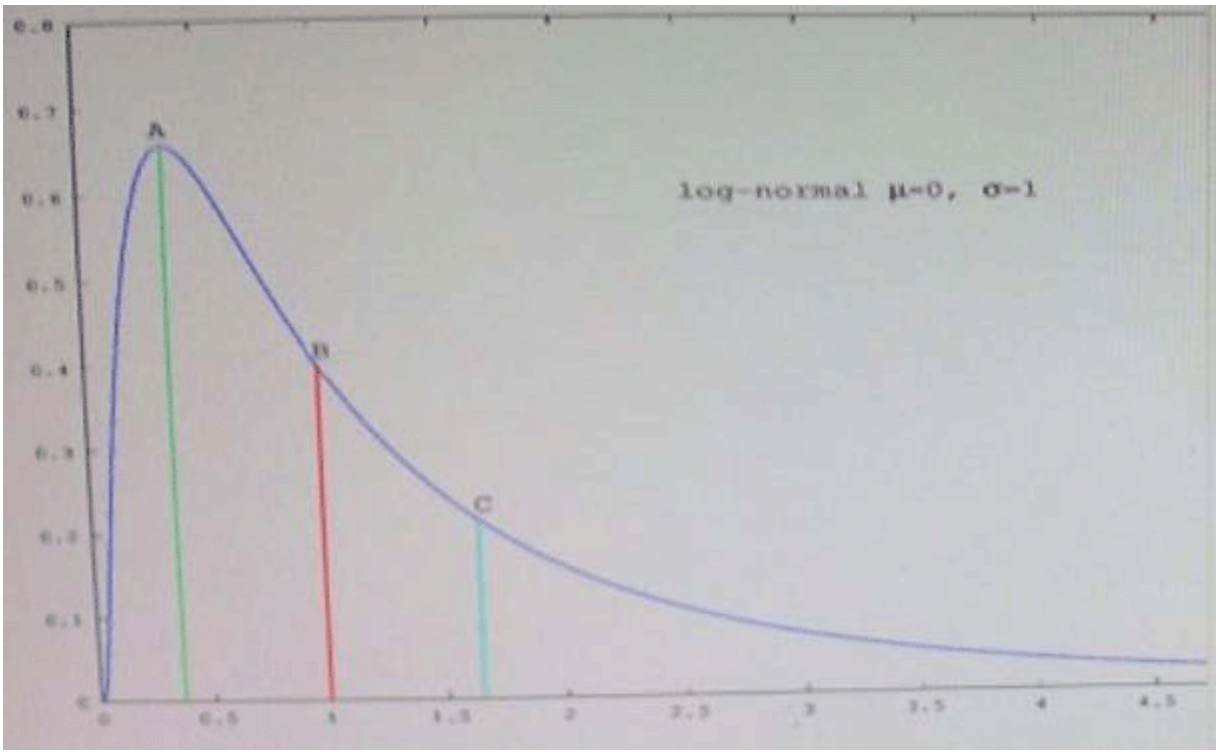
B. B

C. C

**Answer:** A

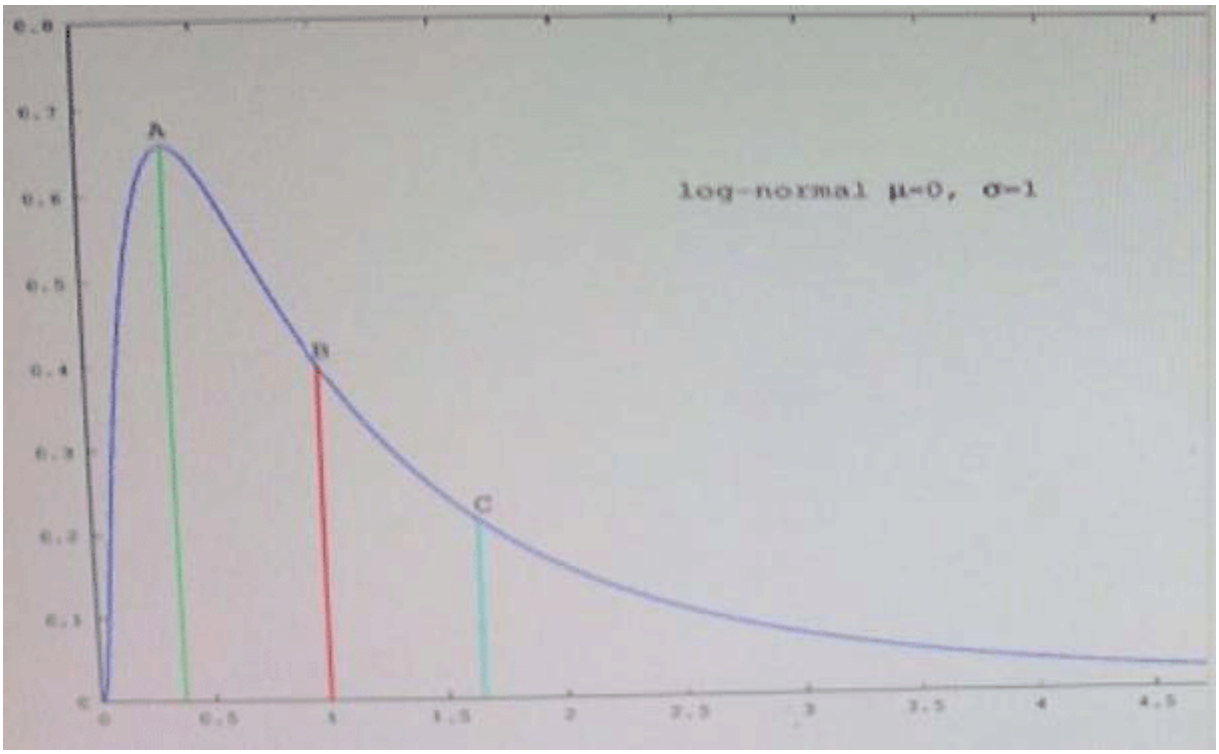
6

Refer to the exhibit.



Which point in the figure is the mode?
A. A
B. B
C. C
**Answer:** C

7.Refer to the exhibit.




Which point in the figure is the mean?
A. A
B. B
C. C
**Answer:** B

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 newobservations.
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

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 - - username foo -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. sqoopimports all the tables from SQLHDFS
**Answer:** C
Explanation:
Reference:https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-15/gettingsqoop

10.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
**Answer:** C