Vendor: Cloudera

Exam Code: DS-200

Exam Name: Data Science Essentials

Version: DEMO

QUESTION 1
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:
https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-15/getting-sqoop

QUESTION 2
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

QUESTION 3
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 4
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 want to use the data from the 52 patients in the scenario to improve the ability of doctors being able to distinguish between ALL and AML. What type of data science problem is this?

A. Classification
B. Regression
C. Clustering
D. Filtering

Answer: D

QUESTION 5
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?

A. A heat map sorting the individuals by group
B. A histogram of the expression values
C. A scatter plot of two largest principal components

Answer: C

QUESTION 6
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

Answer: CDF

QUESTION 7
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?
You choose to perform agglomerative hierarchical clustering on the 10,000 features. How much RAM do you need to hold the distance Matrix, assuming each distance value is 64-bit double?

A. ~ 800 MB
B. ~ 400 MB
C. ~ 160 KB
D. ~ 4 MB

Answer: B

QUESTION 8
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

Answer: C

QUESTION 9
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

Answer: A

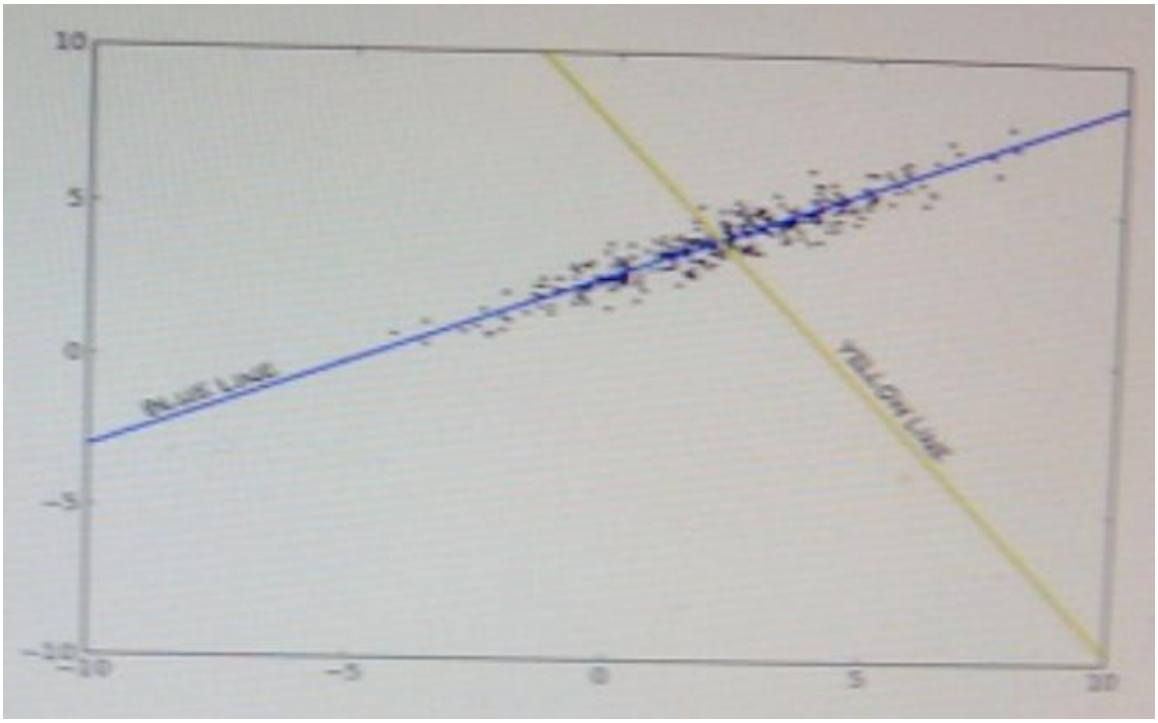
QUESTION 10
Many machine learning algorithm involve finding the Global minimum of a convex loss function, primarily because:

A. The additive inverse of a convex function is concave
B. The derivative of convex function is always defined
C. The second derivative of a convex function is a constant
D. Any local minimum of a convex is also a global minimum

Answer: B

QUESTION 11
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

Answer: A

QUESTION 12
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 forevaluating 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

Answer: AE

Thank You for Trying Our Product

PassLeader Certification Exam Features:

★ More than 99,900 Satisfied Customers Worldwide.

★ Average 99.9% Success Rate.

★ Free Update to match latest and real exam scenarios.

★ Instant Download Access! No Setup required.

★ Questions & Answers are downloadable in PDF format and VCE test engine format.

★ Multi-Platform capabilities - Windows, Laptop, Mac, Android, iPhone, iPod, iPad.

★ 100% Guaranteed Success or 100% Money Back Guarantee.

★ Fast, helpful support 24x7.

View list of all certification exams: http://www.passleader.com/all-products.html

10% Discount Coupon Code: **STNAR2014**