DS-200$^{Q\&As}$

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

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

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

Correct Answer: B

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)

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

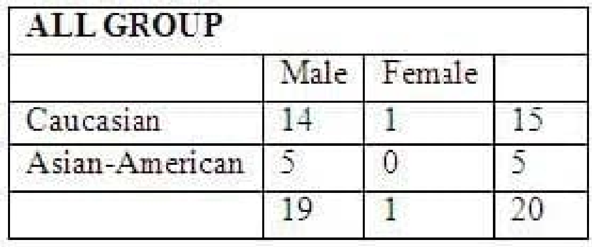
E. The median of the data set

Correct Answer: C

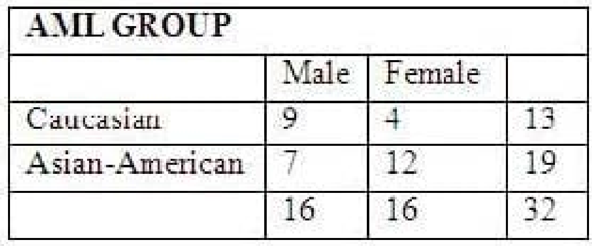
QUESTION 4

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: 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.



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

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

Correct Answer: C

QUESTION 5

You want to build a classification model to identify spam comments on a blog. You decide to use the words in the comment text as inputs to your model. Which criteria should you use when deciding which words to use as features in order to contribute to making the correct classification decision?

A. Choose words for your sample that are most correlated with the Spam label

B. Choose words for your sample that occur most frequently in the text

C. Choose words, for your sample that have the largest mutual information with the spam label

D. Choose words for your sample that are least correlated with the spam label

Correct Answer: A