DS-200^Q&As

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

QUESTION 1

You have acquired a new data source of millions of customer records, and you\\'ve this data into HDFS. Prior to analysis, you want to change all customer registration to the same date format, make all addresses uppercase, and remove all customer names (for anonymization). Which process will accomplish all three objectives?

A. Adapt the data cleansing module in Mahout to your data, and invoke the Mahout library when you run your analysis

B. Pull this data into an RDBMS using sqoop and scrub records using stored procedures

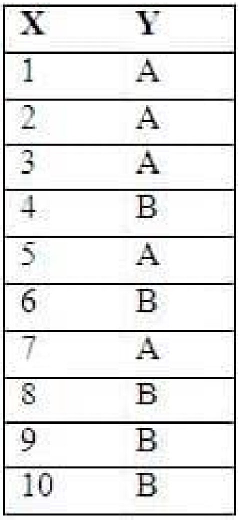
C. Write a script that receives records on stdin, corrects them, and then writes them to stdout. Then, invoke this script in a map-only Hadoop Streaming Job

D. Write a MapReduce job with a mapper to change words to uppercase and to reduce different forms of dates to a single form

Correct Answer: C

QUESTION 2

Consider the following sample from a distribution that contains a continuous X and label Y that is either A or B:



X	Y
1	A
2	A
3	A
4	B
5	A
6	B
7	A
8	B
9	B
10	B

Which is the best cut point for X if you want to discretize these values into two buckets in a way that minimizes the sum of chi-square values?

A. X 8

B. X 6

C. X 5

D. X 4

E. X 2

Correct Answer: D

QUESTION 3

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

Correct Answer: B

QUESTION 4

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.

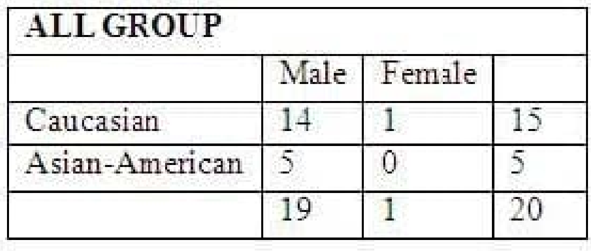
D. When we are required to find the parameters that return the optimal value of the objective function.

Correct Answer: AB

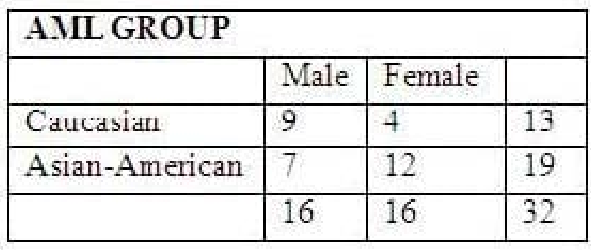
QUESTION 5

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.

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

Correct Answer: D