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

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

QUESTION 1

You are building a k-nearest neighbor classifier (k-NN) on a labeled set of points in a high- dimensional space. You determine that the classifier has a large error on the training data. What is the most likely problem?

A. High-dimensional spaces effectively make local neighborhoods global

B. k-NN compotation does not coverage in high dimensions

C. k was too small

D. The VC-dimension of a k-NN classifier is too high

Correct Answer: B

QUESTION 2

Given the following sample of numbers from a distribution:

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89

How do high-level languages like Apache Hive and Apache Pig efficiently calculate approximately percentiles for a distribution?

A. They sort all of the input samples and the lookup the samples for each percentile

B. They maintain index of input data as it is loaded into HDFS and load them into memory

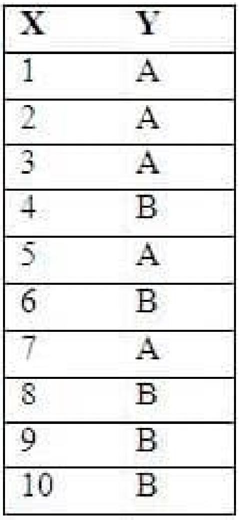
C. They use pivots to assign each observations to the reducer that calculate each percentile

D. They assign sample observations to buckets and then aggregate the buckets to compute the approximations

Correct Answer: C

QUESTION 3

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 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

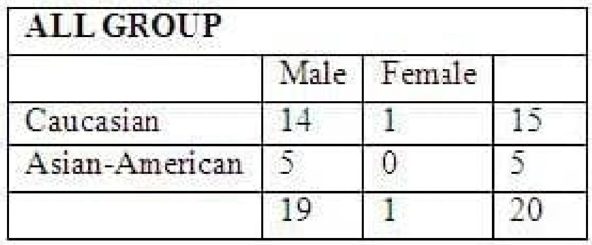
D. L1

Correct Answer: A

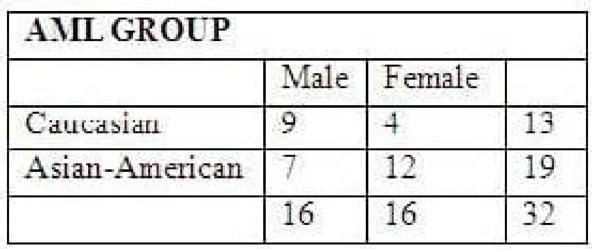
QUESTION 5

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?

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

Correct Answer: CDF