DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Q&As

Databricks Certified Professional Data Scientist Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.geekcert.com/databricks-certified-professional-data-scientist.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

In which of the scenario you can use the regression to predict the values?

A. Samsung can use it for mobile sales forecast

B. Mobile companies can use it to forecast manufacturing defects

C. Probability of the celebrity divorce

D. Only 1 and 2

E. All 1 ,2 and 3

Correct Answer: E

Explanation: Regression is a tool which Companies may use this for things such as sales forecasts or forecasting manufacturing defects. Another creative example is predicting the probability of celebrity divorce.

QUESTION 2

You are working in a classification model for a book, written by HadoopExam Learning Resources and decided to use building a text classification model for determining whether this book is for Hadoop or Cloud computing. You have to select the proper features (feature selection) hence, to cut down on the size of the feature space, you will use the mutual information of each word with the label of hadoop or cloud to select the 1000 best features to use as input to a Naive Bayes model. When you compare the performance of a model built with the 250 best features to a model built with the 1000 best features, you notice that the model with only 250 features performs slightly better on our test data.

What would help you choose better features for your model?

A. Include least mutual information with other selected features as a feature selection criterion

B. Include the number of times each of the words appears in the book in your model

C. Decrease the size of our training data

D. Evaluate a model that only includes the top 100 words

Correct Answer: A

Explanation: Correlation measures the linear relationship (Pearson\\'s correlation) or monotonic relationship (Spearman\\'s correlation) between two variables, X and Y. Mutual information is more general and measures the reduction of uncertainty in Y after observing X. It is the KL distance between the joint density and the product of the individual densities. So Ml can measure non-monotonic relationships and other more complicated relationships Mutual information is a quantification of the dependency between random variables. It is sometimes contrasted with linear correlation since mutual information captures nonlinear dependence. Features with high mutual information with the predicted value are good. However a feature may have high mutual information because it is highly correlated with another feature that has already been selected. Choosing another feature with somewhat less mutual information with the predicted value, but low mutual information with other selected features, may be more beneficial. Hence it may help to also prefer features that are less redundant with other selected features.

QUESTION 3

Which of the following is a correct example of the target variable in regression (supervised learning)?

A. Nominal values like true, false

B. Reptile, fish, mammal, amphibian, plant, fungi

C. Infinite number of numeric values, such as 0.100, 42.001, 1000.743..

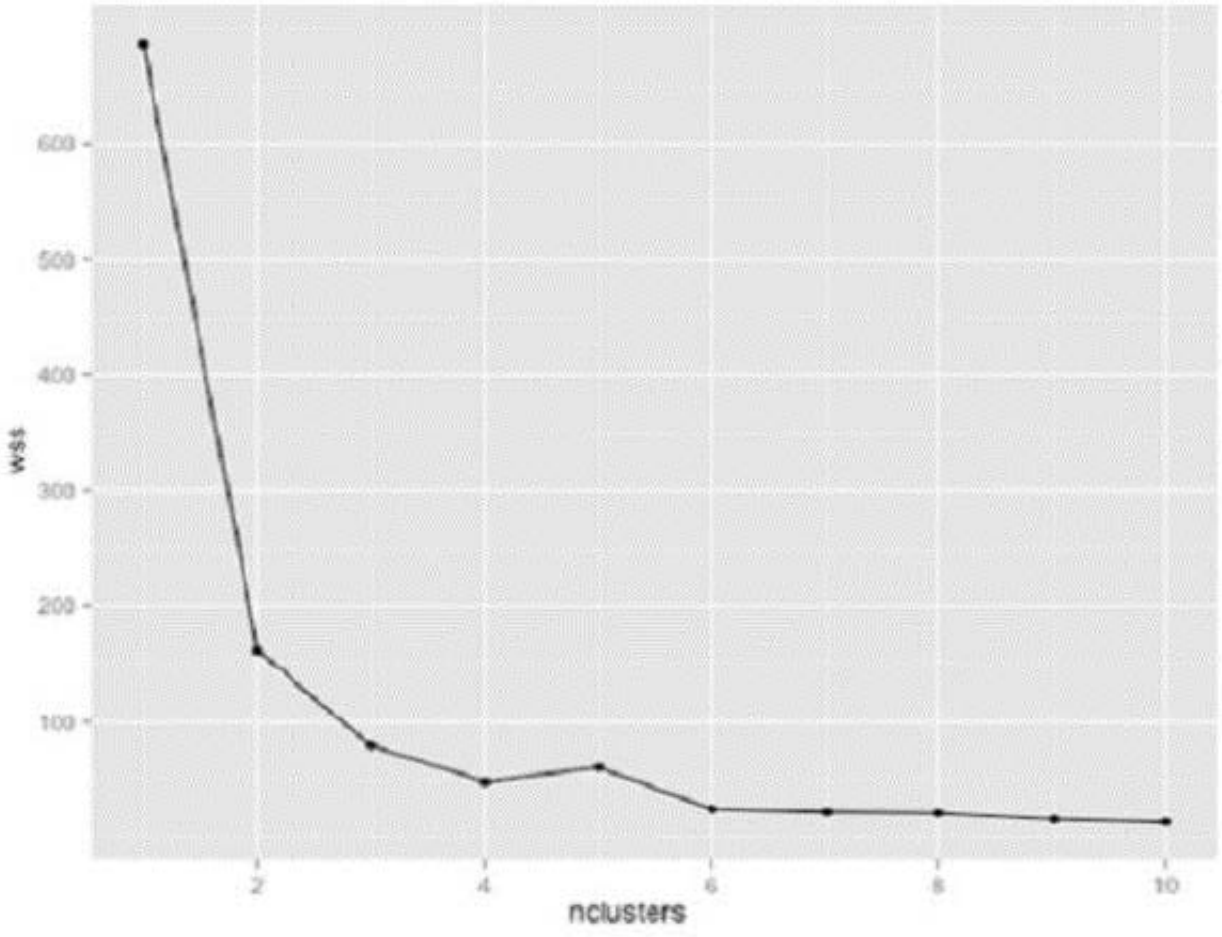
D. All of the above

Correct Answer: D

Explanation: We address two cases of the target variable. The first case occurs when the target variable can take only nominal values: true or false; reptile, fish: mammal, amphibian, plant, fungi. The second case of classification occurs when the target variable can take an infinite number of numeric values, such as 0.100, 42.001, 1000.743, ... This case is called regression.

QUESTION 4

Refer to the exhibit.



You are using K-means clustering to classify customer behavior for a large retailer. You need to determine the optimum

number of customer groups. You plot the within-sum-of- squares (wss) data as shown in the exhibit. How many customer groups should you specify?

A. 2

B. 3

C. 4

D. 8

Correct Answer: C

QUESTION 5

Of all the smokers in a particular district, 40% prefer brand A and 60% prefer brand B. Of those smokers who prefer brand A. 30% are females, and of those who prefer brand B. 40% are female. What is the probability that a randomly selected smoker prefers brand A, given that the person selected is a female?

Which of the following is a best way to solve this problem?

A. Bays Theorem

B. Poisson Distribution

C. Binomial Distribution

D. None of the above

Correct Answer: A