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.leads4pass.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

Refer to exhibit

Independent Variable	Coefficient	P-Value
A	0.45	0
B	3.67	0
C	1.23	0

$R^2 = 0.10$

You are asked to write a report on how specific variables impact your client\\'s sales using a data set provided to you by the client. The data includes 15 variables that the client views as directly related to sales, and you are restricted to these variables only. After a preliminary analysis of the data, the following findings were made: 1. Multicollinearity is not an issue among the variables 2. Only three variables-A, B, and C-have significant correlation with sales You build a linear regression model on the dependent variable of sales with the independent variables of A, B, and C. The results of the regression are seen in the exhibit. You cannot request additional data. what is a way that you could try to increase the R2 of the model without artificially inflating it?

A. Create clusters based on the data and use them as model inputs

B. Force all 15 variables into the model as independent variables

C. Create interaction variables based only on variables A, B, and C

D. Break variables A, B, and C into their own univariate models

Correct Answer: A

Explanation: In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variable) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. (This term should be distinguished from multivariate linear regression^ where multiple correlated dependent variables are predicted, rather than a single scalar variable.) In linear regression data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly: linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X: which is the domain of multivariate analysis.

QUESTION 2

A bio-scientist is working on the analysis of the cancer cells. To identify whether the cell is cancerous or not, there has been hundreds of tests are done with small variations to say yes to the problem. Given the test result for a sample of healthy and cancerous cells, which of the following technique you will use to determine whether a cell is healthy?

A. Linear regression

B. Collaborative filtering

C. Naive Bayes

D. Identification Test

Correct Answer: C

Explanation: In this problem you have been given high-dimensional independent variables like yes, no: test results etc. and you have to predict either valid or not valid (One of two). So all of the below technique can be applied to this problem. Support vector machines Naive Bayes Logistic regression Random decision forests

QUESTION 3

A fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the:

A. Presence of the other features.

B. Absence of the other features.

C. Presence or absence of the other features

D. None of the above

Correct Answer: C

Explanation: In simple terms, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

QUESTION 4

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 MI 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 5

Which of the following problem you can solve using binomial distribution

A. A manufacturer of metal pistons finds that on the average: 12% of his pistons are rejected because they are either oversize or undersize. What is the probability that a batch of 10 pistons will contain no more than 2 rejects?

B. A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson\\'s law to calculate the probability that in a given week he will sell Some policies

C. Vehicles pass through a junction on a busy road at an average rate of 300 per hour Find the probability that none passes in a given minute.

D. It was found that the mean length of 100 parts produced by a lathe was 20.05 mm with a standard deviation of 0.02 mm. Find the probability that a part selected at random would have a length between 20.03 mm and 20.08 mm

Correct Answer: A

Explanation: The entire problem can be solved using below method Binomial: A manufacturer of metal pistons finds that on the average, 12% of his pistons are rejected because they are either oversize or undersize. What is the probability that a batch of 10 pistons will contain no more than 2 rejects? Poisson: A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson\\'s law to calculate the probability that in a given week he will sell Some policies Poisson: Vehicles pass through a junction on a busy road at an average rate of 300 per hour Find the probability that none passes in a given minute. Normal: It was found that the mean length of 100 parts produced by a lathe was

20.05 mm with a standard deviation of 0.02 mm. Find the probability that a part selected at random would have a length between 20 03 mm and 20.08 mm