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.passapply.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

Which of the below best describe the Principal component analysis

A. Dimensionality reduction

B. Collaborative filtering

C. Classification

D. Regression

E. Clustering

Correct Answer: A

QUESTION 3

Spam filtering of the emails is an example of

A. Supervised learning

B. Unsupervised learning

C. Clustering

D. 1 and 3 are correct

E. 2 and 3 are correct

Correct Answer: A

Explanation: Clustering is an example of unsupervised learning. The clustering algorithm finds groups within the data without being told what to look for upfront. This contrasts with classification, an example of supervised machine learning, which is the process of determining to which class an observation belongs. A common application of classification is spam filtering. With spam filtering we use labeled data to train the classifier: e-mails marked as spam or ham.

QUESTION 4

RMSE is a useful metric for evaluating which types of models?

A. Logistic regression

B. Naive Bayes classifier

C. Linear regression

D. All of the above

Correct Answer: C

Explanation: Error calculation allows you to see how well a machine learning method is performing.

One way of determining this performance is to calculate a numerical error This number is sometimes a percent,

however it can also be a score or distance. The goal is usually to minimize an error percent or distance:

however th goal may be to minimize or maximize a score. Encog supports the following error calculation methods.

Sum of Squares Error (ESS)

Root Mean Square Error (RMS)

Mean Square Error (MSE) (default)

SOM Error (Euclidean Distance Error)

RMSE measures error of a predicted numeric value, and so applies to contexts like regression and some recommender system techniques, which rely on predicting a numeric value. It is not relevant to classification techniques

like logistic regression and Naive Bayes, which predict categorical values. It also is not relevant to unsupervied techniques like clustering. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used

measure of the

differences between values predicted by a model or an estimator and the values actually observed. Basically,

the RMSD represents the sample standard deviation of the differences between predicted values and observed values.

These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the

magnitudes

of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy,

but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent.

QUESTION 5

In which of the following scenario we can use naTve Bayes theorem for classification

A. Classify whether a given person is a male or a female based on the measured features. The features include height, weight and foot size.

B. To classify whether an email is spam or not spam

C. To identify whether a fruit is an orange or not based on features like diameter, color and shape

Correct Answer: ABC

Explanation: naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They requires a small amount of training data to estimate the necessary parameters