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.pass4itsure.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

Scenario: Suppose that Bob can decide to go to work by one of three modes of transportation, car, bus, or commuter train. Because of high traffic, if he decides to go by car. there is a 50% chance he will be late. If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20%. The commuter train is almost never late, with a probability of only 1 %, but is more expensive than the bus.

Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car. Since he does not know Which mode of transportation Bob usually uses, he gives a prior probability of 1 3 to each of the three possibilities. Which of the following method the boss will use to estimate of the probability that Bob drove to work?

A. Naive Bayes

B. Linear regression

C. Random decision forests

D. None of the above

Correct Answer: A

Explanation: Bayes\\' theorem (also known as Bayes\\' rule) is a useful tool for calculating conditional probabilities.

QUESTION 2

You are doing advanced analytics for the one of the medical application using the regression and you have two variables which are weight and height and they are very important input variables, which cannot be ignored and they are also

highly co-related.

What is the best solution for that?

A. You will take cube root of height

B. You will take square root of weight

C. You will take square of the height.

D. You would consider using BMI (Body Mass Index)

Correct Answer: D

Explanation: If multiple variables are highly co-related then it is better you consider using the either of the variable which correlates more (which is not in the given option) or go for the new variable which is a function of the both the variable in this case it could be BMI (Body Mass Index). Because it is a function of both weight and height as per the below formula. BMI = Weight/(Height * Height)

QUESTION 3

If E1 and E2 are two events, how do you represent the conditional probability given that E2 occurs given that E1 has

occurred?

A. P(E1)/P(E2)

B. P(E1+E2)/P(E1)

C. P(E2)/P(E1)

D. P(E2)/(P(E1+E2)

Correct Answer: C

QUESTION 4

The method based on principal component analysis (PCA) evaluates the features according to:

A. The projection of the largest eigenvector of the correlation matrix on the initial dimensions

B. According to the magnitude of the components of the discriminate vector

C. The projection of the smallest eigenvector of the correlation matrix on the initial dimensions

D. None of the above

Correct Answer: A

Explanation: Feature Selection:

The method based on principal component analysis (PCA) evaluates the features according to the projection of the largest eigenvector of the correlation matrix on the initial dimensions, the method based on Fisher\\'s linear discriminate

analysis evaluates. Them according to the magnitude of the components of the discriminate vector.

QUESTION 5

Which of the following true with regards to the K-Means clustering algorithm?

A. Labels are not pre-assigned to each objects in the cluster.

B. Labels are pre-assigned to each objects in the cluster.

C. It classify the data based on the labels.

D. It discovers the center of each cluster.

E. It find each objects fall in which particular cluster

Correct Answer: ADE

Explanation: Clustering does not require any predefined labels on the object, rather it consider the attributes on the object. Hence, option-B is out. Clustering is different than classification technique.

Hence you can discard the option-C as well. It does not use the pre-defined labels, hence it is called unsupervised

learning and option-Ais correct. Main purpose of the Clustering technique is to determine the center of each Cluster and then

find the distance from that center. If object is near the center than it would fall in that particular cluster. Hence, finally you will have group or clusters created and get to know that objects fall in which particular cluster.