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

May have a trend component that is quadratic in nature. Which pattern of data will indicate that the trend in the time series data is quadratic in nature?

A. Naive Bayesian classifier

B. Decision tree

C. Linear regression

D. K-means clustering

Correct Answer: D

Explanation: kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to kmeans, including ones for the initial values of the cluster centroids, and for the maximum number of iterations. Clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes. Some specific applications of k-means are image processing, medical and customer segmentation. Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics. Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns.

QUESTION 2

Suppose you have been given two Random Variables X and Y, whose joint distribution is already known, the marginal distribution of X is simply the probability distribution of X averaging over information about Y. It is the probability distribution of X when the value of Y is not known. So how do you calculate the marginal distribution of X

A. This is typically calculated by summing the joint probability distribution over Y.

B. This is typically calculated by integrating the joint probability distribution over Y

C. This is typically calculated by summing (In case of discrete variable) the joint probability distribution over Y

D. This is typically calculated by integrating(In case of continuous variable) the joint probability distribution over Y.

Correct Answer: ABCD

Explanation: Given two random variables X and Y whose joint distribution is known, the marginal distribution of X is simply the probability distribution of X averaging over information about Y. It is the probability distribution of X when the value of Y is not known. This is typically calculated by summing or integrating the joint probability distribution over Y. \\' For discrete random variables, the marginal probability mass function can be written as Pr(X = x). This is

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y) = \sum_y \Pr(X = x|Y = y) \Pr(Y = y),$$

Text

Description automatically generated with low confidence where Pr(X = x,Y = y) is the joint distribution of X and Y, while Pr(X = x|Y = y) is the conditional distribution of X given Y In this case, the variable Y has been marginalized out. Bivariate marginal and joint probabilities for discrete random variables are often displayed as two-way tables. Similarly for continuous random variables, the marginal probability density function can be written as pX(x). This is

$$p_X(x) = \int_y p_{X,Y}(x, y)\ \mathrm{d}y = \int_y p_{X|Y}(x|y)\, p_Y(y)\ \mathrm{d}y,$$

Diagram Description automatically generated with medium confidence where pX.Y(x.y) gives the joint distribution of X and Y while pX|Y(x|y) gives the conditional distribution for X given Y Again: the variable Y has been marginalized out.

Note that a marginal probability can always be written as an expected value:

$$p_X(x) = \int_y p_{X|Y}(x|y)\, p_Y(y)\ \mathrm{d}y = \mathbb{E}_Y[p_{X|Y}(x|y)]$$

Text, letter Intuitively, the marginal probability of X is computed by examining the conditional probability of X given a particular value of Y, and then averaging this conditional probability over the distribution of all values of Y This follows from the definition of expected value, i.e. in general

$$\mathbb{E}_Y[f(Y)] = \int_y f(y) p_Y(y)\ \mathrm{d}y$$

A picture containing diagram

QUESTION 3

In which of the scenario you can use the linear regression model?

A. Predicting Home Price based on the location and house area

B. Predicting demand of the goods and services based on the weather

C. Predicting tumor size reduction based on input as number of radiation treatment

D. Predicting sales of the text book based on the number of students in state

Correct Answer: ABCD

Explanation: : You can use the linear regression model for predicting the continuous output variable based on the input variables. In all the cases mentioned in the question option, you can see that output can be predicted based on the input variable. Option-A: Input: Location, House Area and Output: House Price Option-B : Input: Weather condition, Output: Demand for the goods and services Option-C : Input: Number of Radiation Session Output: Tumor Size Reduction Option-D : Input: Number of students and Output: Sale quantity of text book

QUESTION 4

Projecting a multi-dimensional dataset onto which vector has the greatest variance?

A. first principal component

B. first eigenvector

C. not enough information given to answer

D. second eigenvector

E. second principal component

Correct Answer: A

Explanation: 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 discriminant analysis evaluates. Them according to the magnitude of the components of the discriminant vector. The first principal component corresponds to the greatest variance in the data, by definition. If we project the data onto the first principal component line, the data is more spread out (higher variance) than if projected onto any other line, including other principal components.

QUESTION 5

A website is opened 3 times by a user. What is the probability of he clicks 2 times the advertisement, is best calculated by"

A. Binomial

B. Poisson

C. Normal

D. Any of the above

Correct Answer: A

Explanation: In a binomial distribution, only 2 parameters, namely n and p, are needed to determine the probability. Where p is the probability of success and q is the probability of failure in a binomial trial, then the expected number of successes in n trials. This is a binomial distribution because there are only 2 possible outcomes (we get a 5 or we don\\'t).

SCIENTIST VCE Dumps SCIENTIST Study Guide SCIENTIST Exam Questions