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.certbus.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 statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model\\'s parameters and the normalizing constant usually ignored in MLEs because:

A. The normalizing constant is always very close to 1

B. The normalizing constant only has a small impact on the maximum likelihood

C. The normalizing constant is often zero and can cause division by zero

D. The normalizing constant doesn\\'t impact the maximizing value

Correct Answer: D

Explanation: (Change the explanation even it is correct)A normalizing constant is positive, and multiplying or dividing a series of values by a positive number does not affect which of them is the largest. Maximum likelihood estimation is concerned only with finding a maximum value, so normalizing constants can be ignored.

QUESTION 2

You are working as a data science consultant for a gaming company. You have three member team and all other stake holders are from the company itself like project managers and project sponsored, data team etc. During the discussion project managed asked you that when can you tell me that the model you are using is robust enough, after which step you can consider answer for this question?

A. Data Preparation

B. Discovery

C. Operationalize

D. Model planning

E. Model building

Correct Answer: E

To answer whether the model you are building is robust enough or not you need to have answer below questions at least

-Model is performing as expected with the test data or not?

-Whatever hypothesis defined in the initial phase is being tested or not?

-Do we need more data?

- Domain experts are convinced or not with the model? And all these can be answered when you have built the model and tested with the test data sets. Hence, correct option will be Model Building.

QUESTION 3

In which lifecycle stage are appropriate analytical techniques determined?

A. Model planning

B. Model building

C. Data preparation

D. Discovery

Correct Answer: A

Explanation: In Phase 3, the data science team identifies candidate models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project, It is during this phase that the team refers to the hypotheses developed in Phase 1, when they first became acquainted with the data and understanding the business problems or domain area. These hypotheses help the team frame the analytics to execute in Phase 4 and select the right methods to achieve its objectives. Some of the activities to consider in this phase include the following: Assess the structure of the datasets. The structure of the datasets is one factor that dictates the tools and analytical techniques for the next phase. Depending on whether the team plans to analyze textual data or transactional data, for example, different tools and approaches are required. Ensure that the analytical techniques enable the team to meet the business objectives and accept or reject the working hypotheses. Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow. A few example models include association rules and logistic regression Other tools, such as Alpine Miner, enable users to set up a series of steps and analyses and can serve as a front-end user interface (Ul) for manipulating Big Data sources in PostgreSQL.

QUESTION 4

A. 2.4

B. 24 0

C. .24

D. .48

E. 4.8

Correct Answer: C

Explanation: Given no additional information, the MLE for the probability of an item in the test set is exactly its frequency in the training set. The method of maximum likelihood corresponds to many well-known estimation methods in statistics.

For example, one may be interested in the heights of adult female penguins, but be unable to measure the height of every single penguin in a population due to cost or time constraints. Assuming that the heights are normally (Gaussian)

distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE would accomplish this by taking the mean and variance as

parameters and finding particular parametric values that make the observed results the most probable (given the

model).

In general, for a fixed set of data and underlying statistical model the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the

selected model with the observed data, and for discrete random variables it indeed maximizes the probability of the observed data under the resulting distribution. Maximum-likelihood estimation gives a unified approach to estimation, which is

well-defined in the case of the normal distribution and many other problems. However in some complicated problems, difficulties do occur: in such problems, maximum-likelihood estimators are unsuitable or do not exist.

QUESTION 5

Which of the following question statement falls under data science category?

A. What happened in last six months?

B. How many products have been sold in a last month?

C. Where is a problem for sales?

D. Which is the optimal scenario for selling this product?

E. What happens, if these scenario continues?

Correct Answer: DE

Explanation: This question wants to check your understanding about BI and Data Science. BI was already existing and analytics team already using it. They need to improve and learn data science technique to solve some problems. If you check the option given in the question, it will confuse you. But if you have worked in BI or as a Data Scientist then it is easy to answer. First 3 option can be easily answered using reporting solution, what sales happened in last six month, what was the problem etc. But for the last two option you need to apply data science techniques like which all scenarios are optimal for product sales, you need to collect the data and applying various techniques for that. Hence, last two option can only be answered using Data Science technique And for this you need to apply techniques like Optimization, predictive modeling, statistical analysis on structured and un-structured data.

QUESTION 6

Suppose you have been given a relatively high-dimension set of independent variables and you are asked to come up with a model that predicts one of Two possible outcomes like "YES" or "NO", then which of the following technique best fit?

A. Support vector machines

B. Naive Bayes

C. Logistic regression

D. Random decision forests

E. All of the above

Correct Answer: E

Explanation: In this problem you have been given high-dimensional independent variables like yeS; nO; no English words , 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 7

You are studying the behavior of a population, and you are provided with multidimensional data at the individual level. You have identified four specific individuals who are valuable to your study, and would like to find all users who are most similar to each individual. Which algorithm is the most appropriate for this study?

A. Association rules

B. Decision trees

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 8

Suppose that the probability that a pedestrian will be tul by a car while crossing the toad at a pedestrian crossing without paying attention to the traffic light is lo be computed. Let H be a discrete random variable taking one value from (Hit. Not Hit). Let L be a discrete random variable taking one value from (Red. Yellow. Green).

Realistically, H will be dependent on L That is, P(H = Hit) and P(H = Not Hit) will take different values depending on whether L is red, yellow or green. A person is. for example, far more likely to be hit by a car when trying to cross while Hie lights for cross traffic are green than if they are red In other words, for any given possible pair of values for Hand L. one must consider the joint probability distribution of H and L to find the probability* of that pair of events occurring together if Hie pedestrian ignores the state of the light

Here is a table showing the conditional probabilities of being bit. defending on ibe stale of the lights (Note that the columns in this table must add up to 1 because the probability of being hit oi not hit is 1 regardless of the stale of the light.)

Conditional distribution: P(H|L)			
	L=Green	L=Yellow	L=Red
H=Not Hit	0.99	0.9	0.2
H=Hit	0.01	0.1	0.8

To find the joint probability distribution, we need more data. Let's say that P(L=green) = 0.2, P(L=yellow) = 0.1, and P(L=red) = 0.7. Multiplying each column in the conditional distribution by the probability of that column occurring, we find the joint probability distribution of H and L, given in the central 2×3 block of entries. (Note that the cells in this 2×3 block add up to 1).

Joint distribution: P(H,L)				
	L=Green	L=Yellow	L=Red	Marginal probability P(H)
H=Not Hit	0.198	0.09	0.14	0.428
H=Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Select the correct statement which applies to above example

A. The marginal probability P(H=Hit) is the sum along the H=Hit row of this joint distribution table, as this is the probability of being hit when the lights are red OR yellow OR green.

B. marginal probability that P(H=Not Hit) is the sum of the H=Not Hit row

C. marginal probability that P(H=Not Hit) is the sum of the H= Hit row

Correct Answer: AB

Explanation: The marginal probability P(H=Hit) is the sum along the H=Hit row of this joint distribution table, as this is the probability of being hit when the lights are red OR yellow OR green. Similarly, the marginal probability that P(H=Not Hit) is the sum of the H=Not Hit row

QUESTION 9

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 10

Select the correct statement which applies to K-Nearest Neighbors

A. No Assumption about the data

B. Computationally expensive

C. Require less memory

D. Works with Numeric Values

Correct Answer: ABD

Explanation: : k-Nearest Neighbors Pros: High accuracy insensitive to outliers, no assumptions about data Cons: Computationally expensive, requires a lot of memory Works with: Numeric values, nominal values

QUESTION 11

Support vector machines (SVMs) are a set of supervised learning methods used for:

A. Linear classification

B. Non-linear classification

C. Regression

Correct Answer: ABC

Explanation: In machine learning, support vector machines (SVMs). also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data and recognize patterns^ used for classification and regression analysis. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel tricky implicitly mapping their inputs into high- dimensional feature spaces.

QUESTION 12

Which technique you would be using to solve the below problem statement? "What is the probability that individual customer will not repay the loan amount?"

A. Classification

B. Clustering

C. Linear Regression

D. Logistic Regression

E. Hypothesis testing

Correct Answer: D

QUESTION 13

Which of the following statement is true for the R square value in the regression model?

A. When R square =1 , all the residuals are equal to 0

B. When R square =0, all the residual are equal to 1

C. R square can be increased by adding more variables to the model.

D. R-squared never decreases upon adding more independent variables.

Correct Answer: ACD

Explanation: R square can be made high, it means when we add more variables R-square will increase. And R-square will never decreases if you add more independent variables. Higher R square value can have lower the residuals.

QUESTION 14

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

QUESTION 15

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.

DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST PDF Dumps

DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Practice Test

DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Braindumps