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.pass2lead.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

If you are trying to predict or forecast a discrete target value, then which is the correct options?

A. Supervised Learning regression algorithms

B. Supervised Learning classification algorithms

C. Un supervised Learning

D. Density estimation algorithm

Correct Answer: B

Explanation: If you\\'re trying to predict or forecast a target value, then you need to look into supervised learning. If not, then unsupervised learning is the place you want to be. If you\\'ve chosen supervised learning, what\\'s your target value? Is it

a discrete value like Yes/No:

1/2/3, A/B/C: or Red/Yellow/Black? If so: then you want to look into classification. If the target value can take on a number of values, say any value from 0.00 to 100.00: or-999 to 999, or +_to -_, then you need to look into regression.

QUESTION 2

Select the correct statement regarding the naive Bayes classification:

A. it only requires a small amount of training data to estimate the parameters

B. Independent variables can be assumed

C. only the variances of the variables for each class need to be determined

D. for each class entire covariance matrix need to be determined

Correct Answer: ABC

Explanation: An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

QUESTION 3

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 4

A problem statement is given as below

Hospital records show that of patients suffering from a certain disease, 75% die of it. What is the probability that of 6 randomly selected patients, 4 will recover?

Which of the following model will you use to solve it.

A. Binomial

B. Poisson

C. Normal

D. Any of the above

Correct Answer: A

QUESTION 5

You have collected the 100\\'s of parameters about the 1000\\'s of websites e.g. daily hits, average time on the websites, number of unique visitors, number of returning visitors etc. Now you have find the most important parameters which can best describe a website, so which of the following technique you will use:

A. PCA (Principal component analysis)

B. Linear Regression

C. Logistic Regression

D. Clustering

Correct Answer: A

Explanation: Principal component analysis . or PCA, is a technique for taking a dataset that is in the form of a set of tuples representing points in a high-dimensional space and finding the dimensions along which the tuples line up best. The idea is to treat the set of tuples as a matrix M and find the eigenvectors for MMT or M T M . The matrix of these eigenvectors can be thought of as a rigid rotation in a high-dimensional space. When you apply this transformation to

the original data, the axis corresponding to the principal eigenvector is the one along which the points are most "spread out,11 More precisely this axis is the one along which the variance of the data is maximized. Put another way, the points can best be viewed as lying along this axis, with small deviations from this axis.

QUESTION 6

A denote the event \\'student is female\\' and let B denote the event \\'student is French\\'. In a class of 100 students suppose 60 are French, and suppose that 10 of the French students are females. Find the probability that if I pick a French student, it will be a girl, that is, find P(A|B).

A. 1/3

B. 2/3

C. 1/6

D. 2/6

Correct Answer: C

Explanation: Since 10 out of 100 students are both French and female, then P(AandB)=10100 Also. 60 out of the 100 students are French, so P(B)=60100 So the required probability is: P(A|B)=P(AandB)P(B)=10/10060/100=16

QUESTION 7

In which of the following scenario you should apply the Bay\\'s Theorem?

A. The sample space is partitioned into a set of mutually exclusive events {A1, A2, . .., An }.

B. Within the sample space, there exists an event B, for which P(B) > 0.

C. The analytical goal is to compute a conditional probability of the form: P(Ak | B).

D. In all above cases

Correct Answer: D

QUESTION 8

Google Adwords studies the number of men, and women, clicking the advertisement on search engine during the midnight for an hour each day.

Google find that the number of men that click can be modeled as a random variable with distribution Poisson(X), and likewise the number of women that click as Poisson(Y).

What is likely to be the best model of the total number of advertisement clicks during the midnight for an hour ?

A. Binomial(X+Y,X+Y)

B. Poisson(X/Y)

C. Normal(X+Y(M+Y)1/2) D. Poisson(X+Y)

Correct Answer: D

Explanation: The total number of clicks is the sum of the number of men and women. The sum of two Poisson random variables also follows a Poisson distribution with rate equal to the sum of their rates. The Normal and Binomial distribution can approximate the Poisson distribution in certain cases, but the expressions above do not approximate Poisson(X+Y).

QUESTION 9

Reducing the data from many features to a small number so that we can properly visualize it in two or three dimensions. It is done in_____

A. supervised learning

B. un-supervised learning

C. k-Nearest Neighbors

D. Support vector machines

Correct Answer: B

Explanation: The opposite of supervised learning is a set of tasks known as unsupervised learning. In unsupervised learning, there\\'s no label or target value given for the data. A task where we group similar items together is known as clustering. In unsupervised learning, we may also want to find statistical values that describe the data. This is known as density estimation. Another task of unsupervised learning may be reducing the data from many features to a small number so that we can properly visualize it in two or three dimensions

QUESTION 10

Which activity is performed in the Operationalize phase of the Data Analytics Lifecycle?

A. Define the process to maintain the model

B. Try different analytical techniques

C. Try different variables

D. Transform existing variables

Correct Answer: A

Explanation: Operationalize In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users. In Phase 4. the team scored the model in the analytics sandbox.

QUESTION 11

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 12

You are creating a model for the recommending the book at Amazon.com, so which of the following recommender system you will use you don\\'t have cold start problem?

A. Naive Bayes classifier

B. Item-based collaborative filtering

C. User-based collaborative filtering

D. Content-based filtering

Correct Answer: D

Explanation: The cold start problem is most prevalent in recommender systems. Recommender systems form a specific type of information filtering (IF) technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the user. Typically, a recommender system compares the user\\'s profile to some reference characteristics. These characteristics may be from the information item (the content-based approach) or the user\\'s social environment (the collaborative filtering approach). In the content-based approach, the system must be capable of matching the characteristics of an item against relevant features in the user\\'s profile. In order to do this, it must first construct a sufficiently-detailed model of the user\\'s tastes and preferences through preference elicitation. This may be done either explicitly (by querying the user) or implicitly (by observing the user\\'s behaviour). In both cases, the cold start problem would imply that the user has to dedicate an amount of effort using the system in its \\'dumb\\' state - contributing to the construction of their user profile - before the system can start providing any intelligent recommendations. Content-based filtering recommender systems use information about items or users to make recommendations, rather than user preferences, so it will perform well with little user preference data. Item-based and user-based collaborative filtering makes predictions based on users\\' preferences for items, os they will typically perform poorly with little user preference data. Logistic regression is not recommender system technique.

QUESTION 13

You are using k-means clustering to classify heart patients for a hospital. You have chosen Patient Sex, Height, Weight, Age and Income as measures and have used 3 clusters. When you create a pair-wise plot of the clusters, you notice that there is significant overlap between the clusters. What should you do?

A. Identify additional measures to add to the analysis

B. Remove one of the measures

C. Decrease the number of clusters

D. Increase the number of clusters

Correct Answer: C

QUESTION 14

Which of the following are advantages of the Support Vector machines?

A. Effective in high dimensional spaces.

B. it is memory efficient

C. possible to specify custom kernels

D. Effective in cases where number of dimensions is greater than the number of samples

E. Number of features is much greater than the number of samples, the method still give good performances

F. SVMs directly provide probability estimates

Correct Answer: ABCD

Explanation: Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, the method is likely to give poor performances.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

QUESTION 15

Under which circumstance do you need to implement N-fold cross-validation after creating a regression model?

A. The data is unformatted.

B. There is not enough data to create a test set.

C. There are missing values in the data.

D. There are categorical variables in the model.

Correct Answer: B