Free Questions for Databricks-Certified-Professional-Data-Scientist by braindumpscollection

Shared by Mosley on 22-07-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A data scientist is asked to implement an article recommendation feature for an on-line magazine.

The magazine does not want to use client tracking technologies such as cookies or reading history. Therefore, only the style and subject matter of the current article is available for making recommendations. All of the magazine's articles are stored in a database in a format suitable for analytics.

Which method should the data scientist try first?

Options:

A- K Means Clustering

B- Naive Bayesian

C- Logistic Regression

D- Association Rules

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.

Answer:

A

Question 2

Question Type: MultipleChoice

What describes a true property of Logistic Regression method?

Options:

A- It handles missing values well.

B- It works well with discrete variables that have many distinct values.

C- It is robust with redundant variables and correlated variables.

D- It works well with variables that affect the outcome in a discontinuous way.

Answer:

C

Question 3

Question Type: **MultipleChoice**

Which of the following statement true with regards to Linear Regression Model?

Options:

A- Ordinary Least Square can be used to estimates the parameters in linear model

B- In Linear model, it tries to find multiple lines which can approximate the relationship between the outcome and input variables.

C- Ordinary Least Square is a sum of the individual distance between each point and the fitted line of regression model.

D- Ordinary Least Square is a sum of the squared individual distance between each point and the fitted line of regression model.

Linear regression model are represented using the below equation

$$Y=B(0) + B(1)X$$

Where B(0) is intercept and B(1) is a slope. As B(0) and B(1) changes then fitted line also shifts accordingly on the plot. The purpose of

the Ordinary Least Square method is to estimates these parameters B(0) and B(1). And similarly it is a sum of squared distance between the observed point and the fitted line. Ordinary least squares (OLS) regression minimizes the sum of the squared residuals. A model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

Answer:

A, D

Question 4

Question Type: MultipleChoice

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

Options:

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

Answer:

A, B, C, D

Explanation:

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 5

Question Type: **MultipleChoice**

What type of output generated in case of linear regression?

Options:

A- Continuous variable

B- Discrete Variable

C- Any of the Continuous and Discrete variable

D- Values between 0 and 1

Linear regression model generate continuous output variable.

Answer:

A

Question 6

Question Type: **MultipleChoice**

Select the sequence of the developing machine learning applications

A) Analyze the input data

B) Prepare the input data

C) Collect data

D) Train the algorithm

E) Test the algorithm

F) Use It

Options:

A- A, B, C, D, E, F

B- C, B, A, D, E, F

C- C, A, B, D, E, F

D- C, B, A, D, E, F

1 Collect data. You could collect the samples by scraping a website and extracting data: or you could get information from an RSS feed or an API. You could have a device collect wind speed measurements and send them to you, or blood glucose levels, or anything you can measure. The number of options is endless. To save some time and effort you could use publicly available data

2 Prepare the input data. Once you have this data, you need to make sure it's in a useable format. The format we'll be using in this book is the Python list. We'll talk about Python more in a little bit, and lists are reviewed in appendix A. The benefit of having this standard format is that you can mix and match algorithms and data sources. You may need to do some algorithm-specific formatting here. Some algorithms need features in a special format, some algorithms can deal with target variables and features as strings, and some need them to be integers. We'll get to this later but the algorithm-specific formatting is usually trivial compared to collecting data.

3 Analyze the input data. This is looking at the data from the previous task. This could be as simple as looking at the data you've parsed

in a text editor to make sure steps 1 and 2 are actually working and you don't have a bunch of empty values. You can also look at the data to see if you can recognize any patterns or if there's anything obvious^ such as a few data points that are vastly different from the rest of the set. Plotting data in one: two, or three dimensions can also help. But most of the time you'll have more than three features, and you can't easily plot the data across all features at one time. You could, however use some advanced methods we'll talk about later to distill multiple dimensions down to two or three so you can visualize the data.

4 If you're working with a production system and you know what the data should look like, or you trust its source: you can skip this step. This step takes human involvement, and for an automated system you don't want human involvement. The value of this step is that it makes you understand you don't have garbage coming in.

5 Train the algorithm. This is where the machine learning takes place. This step and the next step are where the 'core' algorithms lie, depending on the algorithm.You feed the algorithm good clean data from the first two steps andextract knowledge or information. This knowledge you often store in a formatthat's readily useable by a machine for the next two steps.In the case of unsupervised learning, there's no training step because youdon't have a target value. Everything is used in the next step.

6 Test the algorithm. This is where the information learned in the previous step isput to use. When you're evaluating an algorithm, you'll test it to see how well itdoes. In the case of supervised learning, you have some known values you can use to evaluate the algorithm. In unsupervised learning, you may have to use some other metrics to evaluate the success. In either case, if you're not satisfied, you can go back to step 4, change some things, and try testing again. Often thecollection or preparation of the data may have been the problem, and you'll have to go back to step 1.

7 Use it. Here you make a real program to do some task, and once again you see if all the previous steps worked as you expected. You might encounter some new data and have to revisit steps 1-5.

Answer:

D

Question 7

Question Type: MultipleChoice

Select the correct objectives of principal component analysis

Options:

A- To reduce the dimensionality of the data set

B- To identify new meaningful underlying variables

C- To discover the dimensionality of the data set

D- Only 1 and 2

E- All 1, 2 and 3

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible: and each succeeding component accounts for as much of the remaining variability as possible.
Objectives of principal component analysis
1. To discover or to reduce the dimensionality of the data set.
2. To identify new meaningful underlying variables.

Answer:

E

Question 8

Question Type: MultipleChoice

Select the correct statement which applies to Principal component analysis (PCA)

Options:

A- Is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables.

B- Is a mathematical procedure that transforms a number of (possibly) correlated variables into a (higher) number of uncorrelated variables

C- Increase the dimensionality of the data set.

D- 1 and 3 are correct

E- 1 and 2 are correct

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrected variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component

accounts for as much of the remaining variability as possible.

Answer:

A