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

Shared by Eaton on 20-10-2022

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Question-18. What is the best way to ensure that the k-means algorithm will find a good clustering of a collection of vectors?

Options:

A- Only consider values of k larger than log(N), where N is the number of observations in the data set
B- Run at least log(N) iterations of Lloyd's algorithm, where N is the number of observations in the data set
C- Choose the initial centroids so that they all He along different axes
D- Choose the initial centroids so that they are far away from each other
k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining, k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.
The problem is computationally difficult (NP-hard); however there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes
This Question-is about the properties that make k-means an effective clustering heuristic which primarily deal with ensuring that the initial centers are far away from each other. This is how modern k-means algorithms like k-means++ guarantee that with high probability Lloyd's algorithm will find a clustering within a constant factor of the optimal possible clustering for each k.

Answer:

D

Question 2

Question Type: MultipleChoice

Select the correct algorithm of unsupervised algorithm

Options:

A- K-Nearest Neighbors
B- K-Means
C- Support Vector Machines
D- Naive Bayes
Sup Supervised learning tasks
Classification Regression
k-Nearest Neighbors Linear
Naive Bayes Locally weighted linear
Support vector machines Ridge
Decision trees Lasso
Unsupervised learning tasks Clustering Density estimation k-Means Expectation maximization
DBSCAN Parzen window

Answer:

A

Question 3

Question Type: MultipleChoice

Which of the following is not a correct application for the Classification?

Options:

A- credit scoring
B- tumor detection
C- image recognition
D- drug discovery
Classification : Build models to classify data into different categories credit scoring, tumor detection, image recognition Regression: Build models to predict continuous data, electricity load forecasting, algorithmic trading, drug discovery

Answer:

D

Question 4

Question Type: MultipleChoice

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

Options:

A- Define the process to maintain the model
B- Try different analytical techniques
C- Try different variables
D- Transform existing variables

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.

Answer:

A

Question 5

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 6

Question Type: MultipleChoice

Digit recognition, is an example of.....

Options:

A- Classification
B- Clustering
C- Unsupervised learning
D- None of the above

Supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Digit recognition: once again, is a common example of classification learning. More generally, classification learning is appropriate for any problem where deducing a classification is useful and the classification is easy to determine. In some cases, it might not even be necessary to give pre-determined classifications to every instance of a problem if the agent can work out the classifications for itself. This would be an example of unsupervised learning in a classification context.

Answer:

A

Question 7

Question Type: MultipleChoice

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

Options:

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.
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.

Answer:

A, C, D

Question 8

Question Type: MultipleChoice

Which of the following is a Continuous Probability Distributions?

Options:

A- Binomial probability distribution
B- Negative binomial distribution
C- Poisson probability distribution
D- Normal probability distribution

Answer:

D