

complete your programming course

about resources, doubts and more!

MYEXAMPLES.NET

Comptia

(DY0-001)

CompTIA DataX

Total: **82 Questions**

Link:

Question: 1

Which of the following issues should a data scientist be most concerned about when generating a synthetic data set?

- A.The data set consuming too many resources
- B.The data set having insufficient features
- C.The data set having insufficient row observations
- D.The data set not being representative of the population

Answer: D

Question: 2

A data scientist is performing a linear regression and wants to construct a model that explains the most variation in the data. Which of the following should the data scientist maximize when evaluating the regression performance metrics?

- A.Accuracy
- B.R²
- C.p value
- D.AUC

Answer: B

Question: 3

A data scientist is building an inferential model with a single predictor variable. A scatter plot of the independent variable against the real-number dependent variable shows a strong relationship between them. The predictor variable is normally distributed with very few outliers. Which of the following algorithms is the best fit for this model, given the data scientist wants the model to be easily interpreted?

- A.A logistic regression
- B.An exponential regression
- C.A linear regression
- D.A probit regression

Answer: C

Question: 4

A data scientist wants to evaluate the performance of various nonlinear models. Which of the following is best suited for this task?

- A.AIC
- B.Chi-squared test
- C.MCC
- D.ANOVA

Answer: C

Question: 5

Which of the following is the layer that is responsible for the depth in deep learning?

- A.Convolution
- B.Dropout
- C.Pooling
- D.Hidden

Answer: D

Question: 6

Which of the following modeling tools is appropriate for solving a scheduling problem?

- A.One-armed bandit
- B.Constrained optimization
- C.Decision tree
- D.Gradient descent

Answer: B

Question: 7

Which of the following environmental changes is most likely to resolve a memory constraint error when running a complex model using distributed computing?

- A.Converting an on-premises deployment to a containerized deployment
- B.Migrating to a cloud deployment
- C.Moving model processing to an edge deployment
- D.Adding nodes to a cluster deployment

Answer: D

Question: 8

A data analyst wants to save a newly analyzed data set to a local storage option. The data set must meet the following requirements:

- Be minimal in size -
 - Have the ability to be ingested quickly
 - Have the associated schema, including data types, stored with it
- Which of the following file types is the best to use?

- A.JSON
- B.Parquet

C.XML
D.CSV

Answer: B

Question: 9

Which of the following is a key difference between KNN and k-means machine-learning techniques?

- A.KNN operates exclusively on continuous data, while k-means can work with both continuous and categorical data.
- B.KNN performs better with longitudinal data sets, while k-means performs better with survey data sets. C.KNN is used for finding centroids, while k-means is used for finding nearest neighbors.
- D.KNN is used for classification, while k-means is used for clustering.

Answer: D

Question: 10

A data scientist needs to:

Build a predictive model that gives the likelihood that a car will get a flat tire.

Provide a data set of cars that had flat tires and cars that did not.

All the cars in the data set had sensors taking weekly measurements of tire pressure similar to the sensors that will be installed in the cars consumers drive. Which of the following is the most immediate data concern?

- A.Granularity misalignment
- B.Multivariate outliers
- C.Insufficient domain expertise
- D.Lagged observations

Answer: A

Question: 11

The term "greedy algorithms" refers to machine-learning algorithms that:

- A.update priors as more data is seen.
- B.examine every node of a tree before making a decision.
- C.apply a theoretical model to the distribution of the data.
- D.make the locally optimal decision.

Answer: D

Question: 12

A data scientist is deploying a model that needs to be accessed by multiple departments with minimal development effort by the departments. Which of the following APIs would be best for the data scientist to use?

- A.SOAP
- B.RPC
- C.JSON
- D.REST

Answer: D

Question: 13

Which of the following compute delivery models allows packaging of only critical dependencies while developing a reusable asset?

- A.Thin clients
- B.Containers
- C.Virtual machines
- D.Edge devices

Answer: B

Question: 14

A data analyst is analyzing data and would like to build conceptual associations. Which of the following is the best way to accomplish this task?

- A.n-grams
- B.NER
- C.TF-IDF
- D.POS

Answer: A

Question: 15

Which of the following belong in a presentation to the senior management team and/or C-suite executives? (Choose two.)

- A.Full literature reviews
- B.Code snippets
- C.Final recommendations
- D.High-level results
- E.Detailed explanations of statistical tests
- F.Security keys and login information

Answer: C

Question: 16

During EDA, a data scientist wants to look for patterns, such as linearity, in the data. Which of the following plots should the data scientist use?

- A.Violin
- B.Box-and-whisker
- C.Scatter
- D.Q-Q

Answer: C

Question: 17

Which of the following distribution methods or models can most effectively represent the actual arrival times of a bus that runs on an hourly schedule?

- A.Binomial
- B.Exponential
- C.Normal
- D.Poisson

Answer: C

Question: 18

A data scientist has constructed a model that meets the minimum performance requirements specified in the proposal for a prediction project. The data scientist thinks the model's accuracy should be improved, but the proposed deadline is approaching. Which of the following actions should the data scientist take first?

- A.Continue collecting data.
- B.Request additional funding.
- C.Consult the key project stakeholder.
- D.Test additional model specifications.

Answer: C

Question: 19

Which of the following best describes the minimization of the residual term in a ridge linear regression?

- A. $|e|$
- B. e
- C. e^2
- D.0

Answer: C

Question: 20

A statistician notices gaps in data associated with age-related illnesses and wants to further aggregate these observations. Which of the following is the best technique to achieve this goal?

- A. Label encoding
- B. Linearization
- C. Binning
- D. Imputing

Answer: C

Question: 21

A data scientist needs to analyze a company's chemical businesses and is using the master database of the conglomerate company. Nothing in the data differentiates the data observations for the different businesses. Which of the following is the most efficient way to identify the chemical businesses' observations?

- A. Ingest the data from all of the hard drives and perform exploratory data analysis to identify which business is responsible for chemical operations.
- B. Perform analysis on all of the data and create a summary report on the results relevant to chemical operations.
- C. Consult with the business team to identify which sites are responsible for chemical operations and ingest only the relevant data for analysis.
- D. Ingest data from the hard drive containing the most data and present sample results on the chemical operations.

Answer: C

Question: 22

Which of the following distance metrics for KNN is best described as a straight line?

- A. Radial
- B. Euclidean
- C. Cosine
- D. Manhattan

Answer: B

Question: 23

A data scientist is building a forecasting model for the price of copper. The only input in this model is the daily price of copper for the last ten years. Which of the following forecasting techniques is the most appropriate for the data scientist to use?

- A. Autoregressive
- B. Moving average
- C. Dynamic time warping
- D. Relative strength

Answer: A

Question: 24

An analyst wants to show how the component pieces of a company's business units contribute to the company's overall revenue. Which of the following should the analyst use to best demonstrate this breakdown?

- A.Box-and-whisker chart
- B.Sankey diagram
- C.Scatter plot matrix
- D.Residual chart

Answer: B

Question: 25

Which of the following does k represent in the k-means model?

- A.Number of model tests
- B.Number of data splits
- C.Number of clusters
- D.Distance between features

Answer: C

MYEXAM.FX