DUMPSBOSS.

Databricks Certified Professional Data Scientist Exam

Databricks Databricks-Certified-Professional-Data-Scientist

Version Demo

Total Demo Questions: 10

Total Premium Questions: 138

Buy Premium PDF

https://dumpsboss.co

support@dumpsboss.co

support@dumpsboss.co
dumpsboss.co

QUESTION NO: 1

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

Above is an example of

A. Linear Regression

B. Logistic Regression

C. Recommendation system

D. Maximum likelihood estimation

E. Hierarchical linear models

ANSWER: B

Explanation:

: Logistic regression

Pros: Computationally inexpensive, easy to implement, knowledge representation easy to interpret

Cons: Prone to underfitting, may have low accuracy Works with: Numeric values, nominal values

QUESTION NO: 2

You are working in an ecommerce organization, where you are designing and evaluating a recommender system, you need to select which of the following metric wilt always have the largest value?

A. Root Mean Square Error

B. Sum of Errors

C. Mean Absolute Error

D. Both land 2

E. Information is not good enough.

ANSWER: E

QUESTION NO: 3

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

ANSWER: A B C

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 NO: 4

Which of the following true with regards to the K-Means clustering algorithm?

A. Labels are not pre-assigned to each objects in the cluster.

B. Labels are pre-assigned to each objects in the cluster.

C. It classify the data based on the labels.

D. It discovers the center of each cluster.

E. It find each objects fall in which particular cluster

ANSWER: A D E

Explanation:

: Clustering does not require any predefined labels on the object, rather it consider the attributes on the object. Hence, option-B is out. Clustering is different than classification technique.

Hence you can discard the option-C as well. It does not use the pre-defined labels, hence it is called unsupervised learning and option-Ais correct. Main purpose of the Clustering technique is to determine the center of each Cluster and then find the distance from that center. If object is near the center than it would fall in that particular cluster. Hence, finally you will have group or clusters created and get to know that objects fall in which particular cluster.

QUESTION NO: 5

Which of the following steps you will be using in the discovery phase?

A. What all are the data sources for the project?

B. Analyze the Raw data and its format and structure.

C. What all tools are required, in the project?

D. What is the network capacity required

E. What Unix server capacity required?

ANSWER: A B C D E

Explanation:

: During the discovery phase you need to find how much resources are required as early as possible and for that even you can involve various stakeholders like

Software engineering team, DBAs,

Network engineers, System administrators etc. for your requirement and these resources are already available or you need to procure them. Also, what would be source of the data?

What all tools and software's are required to execute the same?

QUESTION NO: 6

Select the correct statement which applies to Supervised learning

A. We asks the machine to learn from our data when we specify a target variable.

B. Lesser machine's task to only divining some pattern from the input data to get the target variable

C. Instead of telling the machine Predict Y for our data X, we're asking What can you tell me about X?

ANSWER: A B C

Explanation:

:

: Supervised learning asks the machine to learn from our data when we specify a target variable.

This reduces the machine's task to only divining some pattern from the input data to get the target variable.

In unsupervised learning we don't have a target variable as we did in classification and regression.

Instead of telling the machine Predict Y for our data X> we're asking What can you tell me about X?

Things we ask the machine to tell us about

X may be What are the six best groups we can make out of X? or What three features occur together most frequently in X?

QUESTION NO: 7

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has

rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. Which of the following will you use to calculate the probability whether it will rain on the

day of Marie's wedding?

A. Naive Bayes

B. Logistic Regression

C. Random Decision Forests

D. All of the above

ANSWER: A

Explanation:

: The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. You should consider Bayes' theorem when the following conditions exist.

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

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

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

QUESTION NO: 8

While working with Netflix the movie rating websites you have developed a recommender system that has produced ratings predictions for your data set that are consistently exactly 1 higher for the user-item pairs in your dataset than the ratings given in the dataset. There are n items in the dataset. What will be the calculated RMSE of your recommender system on the dataset?

A. 1

B. 2

C. 0

D. n/2

ANSWER: A

Explanation:

: The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Basically, the RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. RMSE is calculated as the square root of the mean of the squares of the errors. The error in every case in this example is 1. The square of 1 is 1 The average of n items with value 1 is 1 The square root of 1 is 1 The RMSE is therefore 1

QUESTION NO: 9

Which of the following are point estimation methods?

A. MAP

B. MLE

C. MMSE

ANSWER: A B C

Explanation:

: Point estimators

• minimum-variance mean-unbiased estimator (MVUE), minimizes the risk (expected loss)of the squared-error loss-function.
• best linear unbiased estimator (BLUE)

• minimum mean squared error (MMSE)

• median-unbiased estimator, minimizes the risk of the absolute-error loss function

• maximum likelihood (ML)

• method of moments, generalized method of moments

QUESTION NO: 10

If E1 and E2 are two events, how do you represent the conditional probability given that E2 occurs given that E1 has occurred?

A. P(E1)/P(E2)

B. P(E1+E2)/P(E1)

C. P(E2)/P(E1)

D. P(E2)/(P(E1+E2)

ANSWER: C