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.pass4itsure.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

RMSE is a useful metric for evaluating which types of models?

A. Logistic regression

B. Naive Bayes classifier

C. Linear regression

D. All of the above

Correct Answer: C

Explanation: Error calculation allows you to see how well a machine learning method is performing.

One way of determining this performance is to calculate a numerical error This number is sometimes a percent,

however it can also be a score or distance. The goal is usually to minimize an error percent or distance:

however th goal may be to minimize or maximize a score. Encog supports the following error calculation methods.

Sum of Squares Error (ESS)

Root Mean Square Error (RMS)

Mean Square Error (MSE) (default)

SOM Error (Euclidean Distance Error)

RMSE measures error of a predicted numeric value, and so applies to contexts like regression and some recommender system techniques, which rely on predicting a numeric value. It is not relevant to classification techniques

like logistic regression and Naive Bayes, which predict categorical values. It also is not relevant to unsupervied techniques like clustering. 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.

QUESTION 2

Let\\'s say you have two cases as below for the movie ratings

1.

You recommend to a user a movie with four stars and he really doesn\\'t like it and he\\'d rate it two stars

2.

You recommend a movie with three stars but the user loves it (he\\'d rate it five stars). So which statement correctly applies?

A. In both cases, the contribution to the RMSE is the same

B. In both cases, the contribution to the RMSE is the different

C. In both cases, the contribution to the RMSE, could varies

D. None of the above

Correct Answer: A

QUESTION 3

Suppose you have been given a relatively high-dimension set of independent variables and you are asked to come up with a model that predicts one of Two possible outcomes like "YES" or "NO", then which of the following technique best fit?

A. Support vector machines

B. Naive Bayes

C. Logistic regression

D. Random decision forests

E. All of the above

Correct Answer: E

Explanation: In this problem you have been given high-dimensional independent variables like yeS; nO; no English words , test results etc. and you have to predict either valid or not valid (One of two). So all of the below technique can be applied to this problem. Support vector machines Naive Bayes Logistic regression Random decision forests

QUESTION 4

A. It creates the smaller models

B. It requires the lesser memory to store the coefficients for the model

C. It reduces the non-significant features e.g. punctuations

D. Noisy features are removed

Correct Answer: B

Explanation: This hashed feature approach has the distinct advantage of requiring less memory and one less pass through the training data, but it can make it much harder to reverse engineer vectors to determine which original feature mapped to a vector location. This is because multiple features may hash to the same location. With large vectors or with multiple locations per feature, this isn\\'t a problem for accuracy but it can make it hard to understand what a classifier is doing. Models always have a coefficient per feature, which are stored in memory during model building. The hashing trick collapses a high number of features to a small number which reduces the number of coefficients and thus memory requirements. Noisy features are not removed; they are combined with other features and so still have an impact. The validity of this approach depends a lot on the nature of the features and problem domain; knowledge of the domain is important to understand whether it is applicable or will likely produce poor results. While hashing features may produce a smaller model, it will be one built from odd combinations of real-world features, and so will be harder to interpret. An additional benefit of feature hashing is that the unknown and unbounded vocabularies typical of word-like variables aren\\'t a problem.

QUESTION 5

Assume some output variable "y" is a linear combination of some independent input variables "A" plus some independent noise "e". The way the independent variables are combined is defined by a parameter vector B y=AB+e where X is an m x n matrix. B is a vector of n unknowns, and b is a vector of m values. Assuming that m is not equal to n and the columns of X are linearly independent, which expression correctly solves for B?

A. $b * (A^T * A)^{-1} * A^T$

B. $A^{-1} * b$

C. $(A^T * A)^{-1} * b$

D. $(A^T * A)^{-1} * A^T * b$

A. Option A

B. Option B

C. Option C

D. Option D

Correct Answer: D

Explanation: This is the standard solution of the normal equations for linear regression. Because A is not square, you cannot simply take its inverse.

QUESTION 6

Which is an example of supervised learning?

A. PCA

B. k-means clustering

C. SVD

D. EM

E. SVM

Correct Answer: E

Explanation: SVMs can be used to solve various real world problems:

SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings.

Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.

SVMs are also useful in medical science to classify proteins with up to 90% of the compounds classified correctly.

Hand-written characters can be recognized using SVM

QUESTION 7

You are working in a data analytics company as a data scientist, you have been given a set of various types of Pizzas available across various premium food centers in a country. This data is given as numeric values like Calorie. Size, and Sale per day etc. You need to group all the pizzas with the similar properties, which of the following technique you would be using for that?

A. Association Rules

B. Naive Bayes Classifier

C. K-means Clustering

D. Linear Regression

E. Grouping

Correct Answer: C

Explanation: Using K means clustering you can create group of objects based on their properties. Where K is number of the groups. In this case, in each group you determine the center of the group and then find the how far each object characteristics from the center. If it is near the center than it can be part of the group. Suppose we have 100 objects and we need to determine 4 groups. Hence, here K=4. Now we determine 4 center values and based on that center value we determine the distance of each object from the center.

QUESTION 8

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 9

What is the considerable difference between L1 and L2 regularization?

A. L1 regularization has more accuracy of the resulting model

B. Size of the model can be much smaller in L1 regularization than that produced by L2- regularization

C. L2-regularization can be of vital importance when the application is deployed in resource-tight environments such as cell-phones.

D. All of the above are correct

Correct Answer: B

Explanation: The two most common regularization methods are called L1 and L2 regularization. L1 regularization penalizes the weight vector for its L1-norm (i.e. the sum of the absolute values of the weights), whereas L2 regularization uses its L2-norm. There is usually not a considerable difference between the two methods in terms of the accuracy of the resulting model (Gao et al 2007), but L1 regularization has a significant advantage in practice. Because many of the weights of the features become zero as a result of L1- regularized training, the size of the model can be much smaller than that produced by L2- regularization. Compact models require less space on memory and storage, and enable the application to start up quickly. These merits can be of vital importance when the application is deployed in resource-tight environments such as cell-phones. Regularization works by adding the penalty associated with the coefficient values to the error of the hypothesis. This way, an accurate hypothesis with unlikely coefficients would be penalized whila a somewhat less accurate but more conservative hypothesis with low coefficients would not be penalized as much.

QUESTION 10

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

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

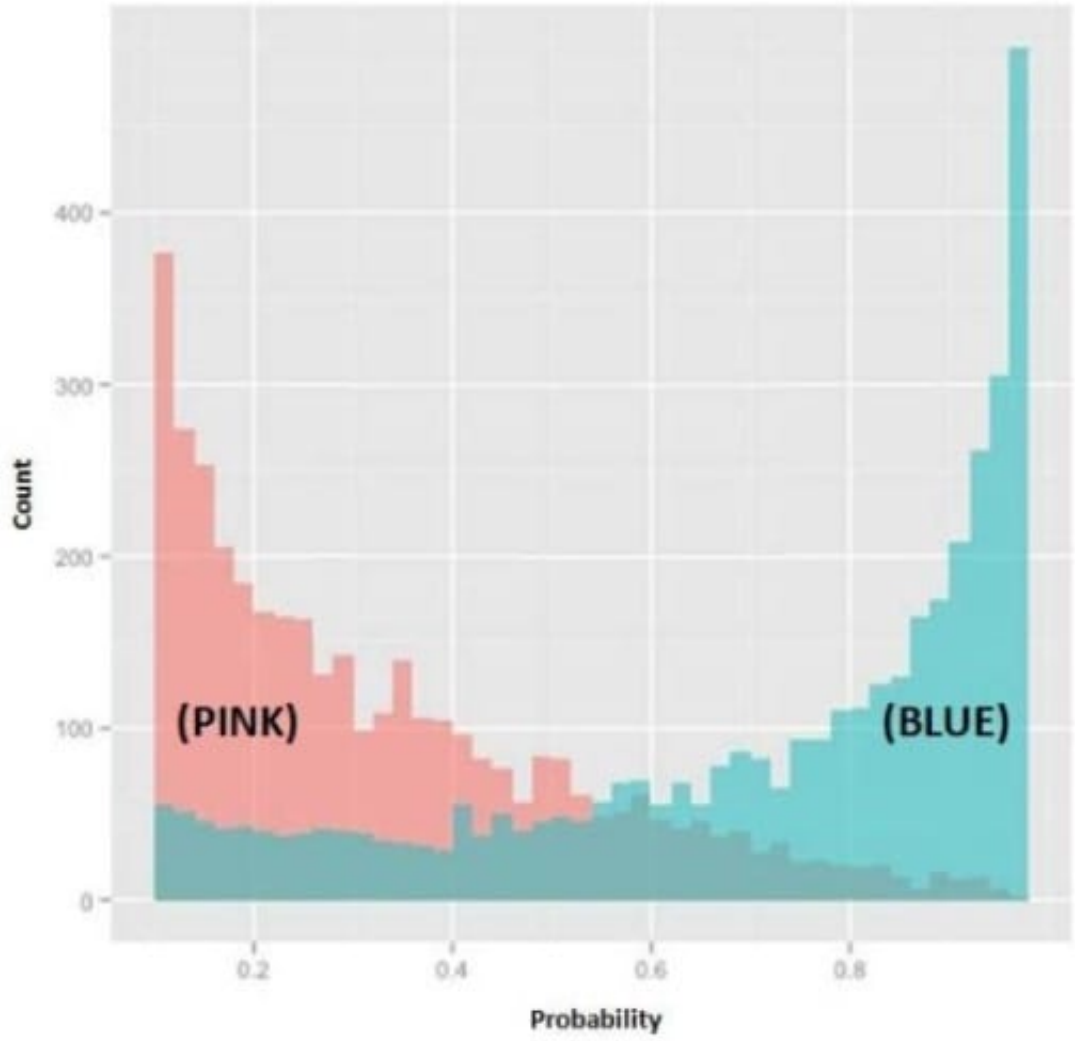
Correct Answer: D

Explanation: 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.

QUESTION 11

Refer to Exhibit



In the exhibit, the x-axis represents the derived probability of a borrower defaulting on a loan. Also in the exhibit, the pink represents borrowers that are known to have not defaulted on their loan, and the blue represents borrowers that are known to have defaulted on their loan. Which analytical method could produce the probabilities needed to build this exhibit?

A. Linear Regression

B. Logistic Regression

C. Discriminant Analysis

D. Association Rules

Correct Answer: B

QUESTION 12

You are designing a recommendation engine for a website where the ability to generate more personalized recommendations by analyzing information from the past activity of a specific user, or the history of other users deemed to be of

similar taste to a given user. These resources are used as user profiling and helps the site recommend content on a user-by-user basis. The more a given user makes use of the system, the better the recommendations become, as the

system gains data to improve its model of that user.

What kind of this recommendation engine is ?

A. Naive Bayes classifier

B. Collaborative filtering

C. Logistic Regression

D. Content-based filtering

Correct Answer: B

Another aspect of collaborative filtering systems is the ability to generate more personalized recommendations by analyzing information from the past activity of a specific user, or the history of other users deemed to be of similar taste to a given user. These resources are used as user profiling and help the site recommend content on a user-by- user basis. The more a given user makes use of the system, the better the recommendations become, as the system gains data to improve its model of that user

---

QUESTION 13

You are asked to create a model to predict the total number of monthly subscribers for a specific magazine. You are provided with 1 year\\'s worth of subscription and payment data, user demographic data, and 10 years worth of content of the magazine (articles and pictures). Which algorithm is the most appropriate for building a predictive model for subscribers?

A. Linear regression

B. Logistic regression

C. Decision trees

D. TF-IDF

Correct Answer: A

: A data model explicitly describes a relationship between predictor and response variables. Linear regression fits a data model that is linear in the model coefficients. The most common type of linear regression is a least-squares fit, which can fit both lines and polynomials, among other linear models. Before you model the relationship between pairs of quantities, it is a good idea to perform correlation analysis to establish if a linear relationship exists between these quantities. Be aware that variables can have nonlinear relationships, which correlation analysis cannot detect. For more information, see Linear Correlation. If you need to fit data with a nonlinear model, transform the variables to make the relationship linear. Alternatively try to fit a nonlinear function directly using either the Statistics and Machine Learning Toolbox nlinfit function, the Optimization Toolbox Isqcurvefit function, or by applying functions in the Curve Fitting Toolbox.

QUESTION 14

Select the correct objectives of principal component analysis:

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

Correct Answer: E

Explanation: 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.

QUESTION 15

In which phase of the analytic lifecycle would you expect to spend most of the project time?

A. Discovery

B. Data preparation

C. Communicate Results

D. Operationalize

Correct Answer: B

In the data preparation phase of the Data Analytics Lifecycle, the data range and distribution can be obtained. If the data is skewed, viewing the logarithm of the data (if it\\'s all positive) can help detect structures that might otherwise be overlooked in a graph with a regular, nonlogarithmic scale. When preparing the data, one should look for signs of dirty data, as explained in the previous section. Examining if the data is unimodal or multimodal will give an idea of how many distinct populations with different behavior patterns might be mixed into the overall population. Many modeling techniques assume that the data follows a normal distribution. Therefore, it is important to know if the available dataset can match that assumption before applying any of those modeling techniques.