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.leads4pass.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

You are working on a problem where you have to predict whether the claim is done valid or not. And you find that most of the claims which are having spelling errors as well as corrections in the manually filled claim forms compare to the honest claims. Which of the following technique is suitable to find out whether the claim is valid or not?

A. Naive Bayes

B. Logistic Regression

C. Random Decision Forests

D. Any one of the above

Correct Answer: D

Explanation: In this problem you have been given high-dimensional independent variables like texts, corrections, 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 2

Regularization is a very important technique in machine learning to prevent over fitting. And Optimizing with a L1 regularization term is harder than with an L2 regularization term because

A. The penalty term is not differentiate

B. The second derivative is not constant

C. The objective function is not convex

D. The constraints are quadratic

Correct Answer: A

Explanation: Regularization is a very important technique in machine learning to prevent overfitting. Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit. The difference between the L1 and L2 is just that L2 is the sum of the square of the weights, while L1 is just the sum of the weights. Much of optimization theory has historically focused on convex loss functions because they\\'re much easier to optimize than non-convex functions: a convex function over a bounded domain is guaranteed to have a minimum, and it\\'s easy to find that minimum by following the gradient of the function at each point no matter where you start. For non- convex functions, on the other hand, where you start matters a great deal; if you start in a bad position and follow the gradient, you\\'re likely to end up in a local minimum that is not necessarily equal to the global minimum. You can think of convex functions as cereal bowls: anywhere you start in the cereal bowl, you\\'re likely to roll down to the bottom. A non-convex function is more like a skate park: lots of ramps, dips, ups and downs. It\\'s a lot harder to find the lowest point in a skate park than it is a cereal bowl.

QUESTION 3

In which lifecycle stage are test and training data sets created?

A. Model planning

B. Discovery

C. Model building

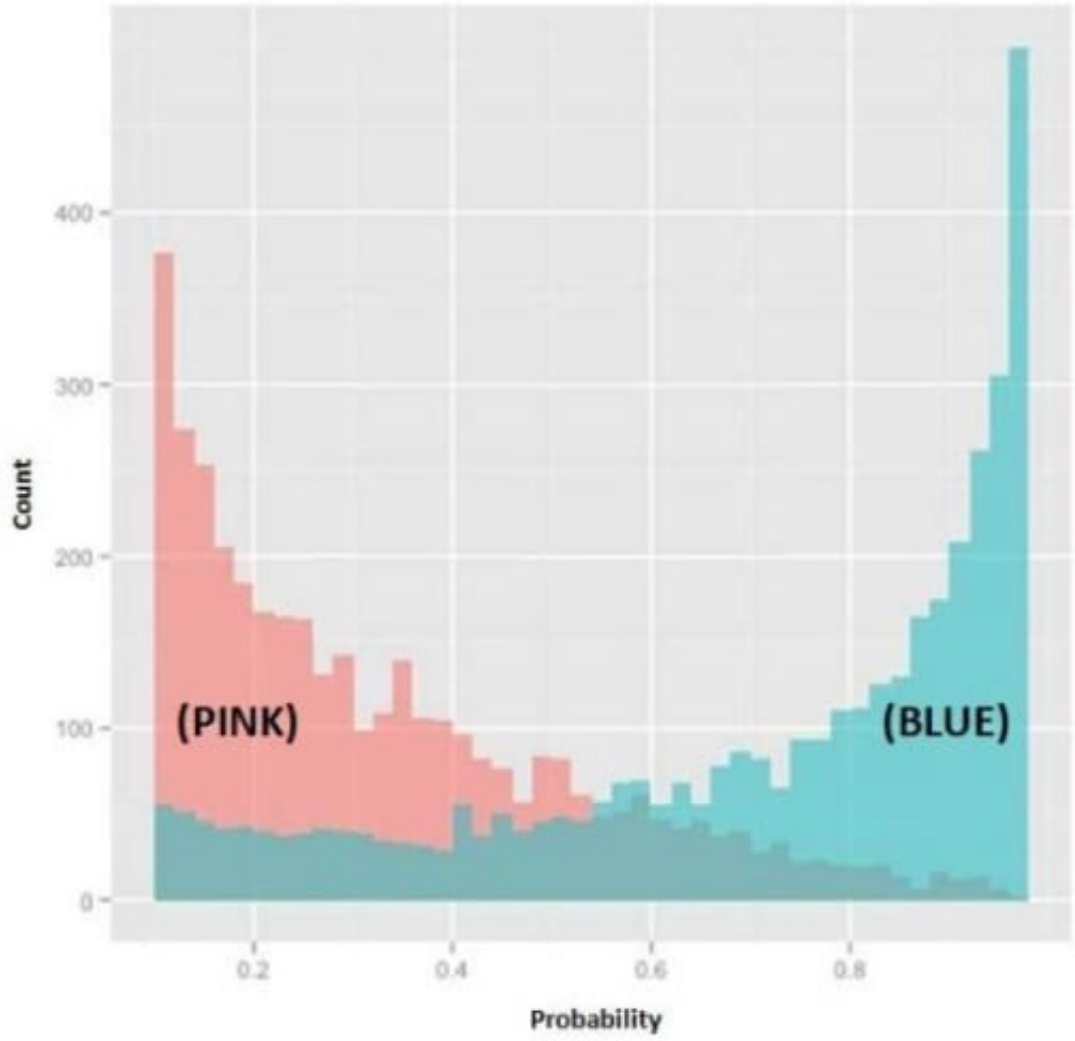
D. Data preparation

Correct Answer: C

Explanation: In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data. Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data Model planning: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models. Model building: In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable). Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders. Operationalize: In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

QUESTION 4

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 5

You are doing advanced analytics for the one of the medical application using the regression and you have two variables which are weight and height and they are very important input variables, which cannot be ignored and they are also

highly co-related.

What is the best solution for that?

A. You will take cube root of height

B. You will take square root of weight

C. You will take square of the height.

D. You would consider using BMI (Body Mass Index)

Correct Answer: D

Explanation: If multiple variables are highly co-related then it is better you consider using the either of the variable which correlates more (which is not in the given option) or go for the new variable which is a function of the both the variable in this case it could be BMI (Body Mass Index). Because it is a function of both weight and height as per the below formula. BMI = Weight/(Height * Height)

QUESTION 6

Clustering is a type of unsupervised learning with the following goals

A. Maximize a utility function

B. Find similarities in the training data

C. Not to maximize a utility function

D. 1 and 2

E. 2 and 3

Correct Answer: E

Explanation: type of unsupervised learning is called clustering. In this type of learning, The goal is not to maximize a utility function, but simply to find similarities in the training data. The assumption is often that the clusters discovered will match reasonably well with an intuitive classification. For instance, clustering individuals based on demographics might result in a clustering of the wealthy in one group and the poor in another. Clustering can be useful when there is enough data to form clusters (though this turns out to be difficult at times) and especially when additional data about members of a cluster can be used to produce further results due to dependencies in the data.

QUESTION 7

You are working with the Clustering solution of the customer datasets. There are almost 40 variables are available for each customer and almost 1.00,0000 customer\\'s data is available. You want to reduce the number of variables for clustering, what would you do?

A. You will randomly reduce the number of variables

B. You will find the correlation among the variables and from their variables are not co- related will be discarded.

C. You will find the correlation among the variables and from the highly co-related variables, you will be considering only

one or two variables from it.

D. You cannot discard any variable for creating clusters.

E. You can combine several variables in one variable

Correct Answer: CE

Explanation: When you are applying clustering technique and you find that there are quite a huge number of variables are available. Then it is better the find the co-relation among the variables and consider only one or two variables from the highly co-related variables. Because highly co-related variable will have the same effect, while creating the cluster. We can use scatter plot matrix among the variables to find the co-relation. You can also combine several variables into a single variable. For example if you have two values in the dataset like Asset and Debt than by combining these two values like Debt to Asset ratio and use it while creating the cluster.

QUESTION 8

You are using one approach for the classification where to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success, where agents might be rewarded for doing certain actions and

punished for doing others.

Which kind of this learning?

A. Supervised

B. Unsupervised

C. Regression

D. None of the above

Correct Answer: B

Explanation: Unsupervised learning seems much harder: the goal is to have the computer learn how to do something that we don\\'t tell it how to do! The approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note that this type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards. This approach nicely generalizes to the real world, where agents might be rewarded for doing certain actions and punished fordoing others.

QUESTION 9

You have data of 10.000 people who make the purchasing from a specific grocery store. You also have their income detail in the data. You have created 5 clusters using this data. But in one of the cluster you see that only 30 people are falling as below 30, 2400, 2600, 2700, 2270 etc."

What would you do in this case?

A. You will be increasing number of clusters.

B. You will be decreasing the number of clusters.

C. You will remove that 30 people from dataset

D. You will be multiplying standard deviation with the 100

Correct Answer: B

Explanation: Decreasing the number of clusters will help in adjusting this outlier cluster to get adjusted in another cluster.

QUESTION 10

Logistic regression is a model used for prediction of the probability of occurrence of an event. It makes use of several variables that may be.....

A. Numerical

B. Categorical

C. Both 1 and 2 are correct

D. None of the 1 and 2 are correct

Correct Answer: C

Explanation: Logistic regression is a model used for prediction of the probability of occurrence of an event. It makes use of several predictor variables that may be either numerical or categories.

QUESTION 11

Spam filtering of the emails is an example of

A. Supervised learning

B. Unsupervised learning

C. Clustering

D. 1 and 3 are correct

E. 2 and 3 are correct

Correct Answer: A

Explanation: Clustering is an example of unsupervised learning. The clustering algorithm finds groups within the data without being told what to look for upfront. This contrasts with classification, an example of supervised machine learning, which is the process of determining to which class an observation belongs. A common application of classification is spam filtering. With spam filtering we use labeled data to train the classifier: e-mails marked as spam or ham.

QUESTION 12

Which of the following are point estimation methods?

A. MAP

B. MLE

C. MMSE

Correct Answer: ABC

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 13

Select the correct option which applies to L2 regularization

A. Computational efficient due to having analytical solutions

B. Non-sparse outputs

C. No feature selection

Correct Answer: ABC

Explanation: :

The difference between their properties can be promptly summarized as follows:

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

QUESTION 14

Which of the following are advantages of the Support Vector machines?

A. Effective in high dimensional spaces.

B. it is memory efficient

C. possible to specify custom kernels

D. Effective in cases where number of dimensions is greater than the number of samples

E. Number of features is much greater than the number of samples, the method still give good performances

F. SVMs directly provide probability estimates

Correct Answer: ABCD

Explanation: Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, the method is likely to give poor performances.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

QUESTION 15

Regularization is a very important technique in machine learning to prevent overfitting. Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit. The difference between the L1 and L2 is...

A. L2 is the sum of the square of the weights, while L1 is just the sum of the weights

B. L1 is the sum of the square of the weights, while L2 is just the sum of the weights

C. L1 gives Non-sparse output while L2 gives sparse outputs

D. None of the above

Correct Answer: A

Explanation: Regularization is a very important technique in machine learning to prevent overfitting. Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit. The difference between

the L1 and L2 is just that L2 is the sum of the square of the weights, while L1 is just the sum of the weights. As follows:

L1 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} w_i$$