Online Questions - Valid Practice To your Databricks-Certified-Professional-Data-Scientist Exam (Updated 140 Questions) [Q15-Q30

Online Questions - Valid Practice To your Databricks-Certified-Professional-Data-Scientist Exam (Updated 140 Questions) Practice To Databricks-Certified-Professional-Data-Scientist - Remarkable Practice On your Databricks Certified Professional Data Scientist Exam Exam

NO.15 Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

Above is an example of
* Linear Regression
* Logistic Regression
* Recommendation system
* Maximum likelihood estimation
* Hierarchical linear models
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

NO.16 Projecting a multi-dimensional dataset onto which vector has the greatest variance?
* first principal component
* first eigenvector
* not enough information given to answer
* second eigenvector
* second principal component
Explanation

The method based on principal component analysis (PCA) evaluates the features according to the projection of the largest eigenvector of the correlation matrix on the initial dimensions, the method based on Fisher&#8217;s linear discriminant analysis evaluates. Them according to the magnitude of the components of the discriminant vector.

The first principal component corresponds to the greatest variance in the data, by definition. If we project the data onto the first principal component line, the data is more spread out (higher variance) than if projected onto any other line, including other principal components.

NO.17 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?
* Naive Bayes
* Logistic Regression
* Random Decision Forests
* Any one of the above

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.

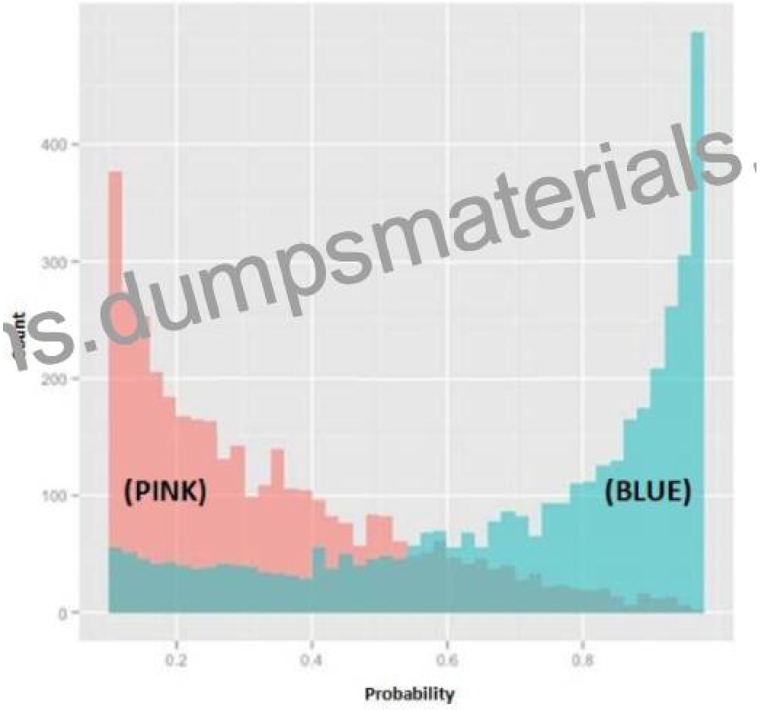
Support vector machines Naive Bayes Logistic regression Random decision forests

NO.18 You have used k-means clustering to classify behavior of 100, 000 customers for a retail store. You decide to use household income, age, gender and yearly purchase amount as measures. You have chosen to use 8 clusters and notice that 2 clusters only have 3 customers assigned. What should you do?
* Decrease the number of measures used
* Increase the number of clusters
* Decrease the number of clusters
* Identify additional measures to add to the analysis
Explanation

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.

NO.19 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?
* Linear Regression
* Logistic Regression
* Discriminant Analysis
* Association Rules

NO.20 Which of the following is a Continuous Probability Distributions?
* Binomial probability distribution
* Negative binomial distribution
* Poisson probability distribution
* Normal probability distribution

NO.21 You are having 1000 patients&#8217; data with the height and age. Where age in years and height in meters. You wanted to create cluster using this two attributes. You wanted to have near equal effect for both the age and height while creating the cluster. What you can do?
* You will be adding height with the numeric value 100
* You will be converting each height value to centimeters
* You will be dividing both age and height with their respective standard deviation
* You will be taking square root of height
Explanation

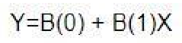
When you see the data age in years would have values like 50, 60r 70 90 years etc. And while calculating distance from centroid maximum possible value can be 90-0 and its square will be 8100.

While using heights in meter can be 2-0.5(1.5) meters and its square will be 2.25 only. So you can see age has more effect than height. Hence bringing the height on same level you can convert it into centimeters. Can bring data upto 200 centimeters and then it be more effective like square of 200 maximum.

However there is another approach is to divide the each value with its standard deviation, which will not have impact of the units e.g. age/sd of the age, which results in value without unit. This can also help in reducing the effect of units.

NO.22 Which of the following statement true with regards to Linear Regression Model?
* Ordinary Least Square can be used to estimates the parameters in linear model
* In Linear model, it tries to find multiple lines which can approximate the relationship between the outcome and input variables.
* Ordinary Least Square is a sum of the individual distance between each point and the fitted line of regression model.
* Ordinary Least Square is a sum of the squared individual distance between each point and the fitted line of regression model.
Explanation

Linear regression model are represented using the below equation



Y=B(0) + B(1)X

Where B(0) is intercept and B(1) is a slope. As B(0) and B(1) changes then fitted line also shifts accordingly on the plot. The purpose of the Ordinary Least Square method is to estimates these parameters B(0) and B(1).

And similarly it is a sum of squared distance between the observed point and the fitted line. Ordinary least squares (OLS) regression minimizes the sum of the squared residuals. A model fits the data well if the differences between the observed values and the model&#8217;s predicted values are small and unbiased.

NO.23 Which of the below best describe the Principal component analysis
* Dimensionality reduction
* Collaborative filtering
* Classification
* Regression
* Clustering

NO.24 Suppose you have made a model for the rating system, which rates between 1 to 5 stars. And you calculated that RMSE value is 1.0 then which of the following is correct
* It means that your predictions are on average one star off of what people really think
* It means that your predictions are on average two star off of what people really think
* It means that your predictions are on average three star off of what people really think
* It means that your predictions are on average four star off of what people really think

NO.25 Your customer provided you with 2. 000 unlabeled records three groups. What is the correct analytical method to use?
* Semi Linear Regression
* Logistic regression
* Naive Bayesian classification
* Linear regression
* K-means clustering

Explanation

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.

The algorithm has nothing to do with and should not be confused with k-nearest neighbor another popular machine learning technique.

NO.26 Select the choice where Regression algorithms are not best fit
* When the dimension of the object given
* Weight of the person is given
* Temperature in the atmosphere
* Employee status

Explanation

Regression algorithms are usually employed when the data points are inherently numerical variables (such as the dimensions of an object the weight of a person, or the temperature in the atmosphere) but unlike Bayesian algorithms, they&#8217;re not very good for categorical data (such as employee status or credit score description).

NO.27 In which lifecycle stage are test and training data sets created?
* Model planning

* Discovery
* Model building
* Data preparation

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.

NO.28 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?

* Naive Bayes
* Linear regression
* Random decision forests
* None of the above

Explanation

Bayes&#8217; theorem (also known as Bayes&#8217; rule) is a useful tool for calculating conditional probabilities.

NO.29 Which of the following is a correct example of the target variable in regression (supervised learning)?

* Nominal values like true, false
* Reptile, fish, mammal, amphibian, plant, fungi
* Infinite number of numeric values, such as 0.100, 42.001, 1000.743..
* All of the above

Explanation

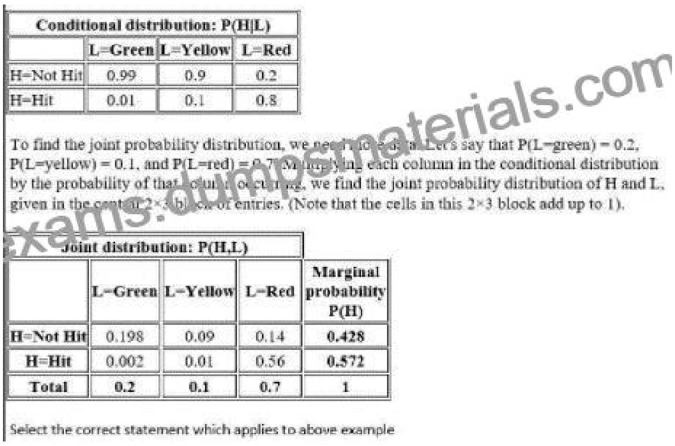
We address two cases of the target variable. The first case occurs when the target variable can take only nominal values: true or false; reptile, fish: mammal, amphibian, plant, fungi. The second case of classification occurs when the target variable can take an infinite number of numeric values, such as 0.100, 42.001,

1000.743, &#8230;. This case is called regression.

NO.30 Suppose that the probability that a pedestrian will be tul by a car while crossing the toad at a pedestrian crossing without paying attention to the traffic light is lo be computed. Let H be a discrete random variable taking one value from (Hit. Not Hit). Let L be a discrete random variable taking one value from (Red. Yellow.

Green).

Realistically, H will be dependent on L That is, P(H = Hit) and P(H = Not Hit) will take different values depending on whether L is red, yellow or green. A person is. for example, far more likely to be hit by a car when trying to cross while Hie lights for cross traffic are green than if they are red In other words, for any given possible pair of values for Hand L. one must consider the joint probability distribution of H and L to find the probability* of that pair of events occurring together if Hie pedestrian ignores the state of the light Here is a table showing the conditional probabilities of being bit. defending on ibe stale of the lights (Note that the columns in this table must add up to 1 because the probability of being hit oi not hit is 1 regardless of the stale of the light.)



Conditional distribution: P(H|L)			
	L=Green	L=Yellow	L=Red
H=Not Hit	0.99	0.9	0.2
H=Hit	0.01	0.1	0.8

To find the joint probability distribution, we need more data. Let's say that P(L=green) = 0.2, P(L=yellow) = 0.1, and P(L=red) = 0.7. Multiplying each column in the conditional distribution by the probability of that column occurring, we find the joint probability distribution of H and L, given in the central 2×3 block of entries. (Note that the cells in this 2×3 block add up to 1).

Joint distribution: P(H,L)				
	L=Green	L=Yellow	L=Red	Marginal probability P(H)
H=Not Hit	0.198	0.09	0.14	0.428
H=Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Select the correct statement which applies to above example

* The marginal probability P(H=Hit) is the sum along the H=Hit row of this joint distribution table, as this is the probability of being hit when the lights are red OR yellow OR green.
* marginal probability that P(H=Not Hit) is the sum of the H=Not Hit row
* marginal probability that P(H=Not Hit) is the sum of the H= Hit row

Explanation

The marginal probability P(H=Hit) is the sum along the H=Hit row of this joint distribution table, as this is the probability of being hit when the lights are red OR yellow OR green. Similarly, the marginal probability that P(H=Not Hit) is the sum of the H=Not Hit row

True Databricks-Certified-Professional-Data-Scientist Exam Extraordinary Practice For the Exam: https://www.dumpsmaterials.com/Databricks-Certified-Professional-Data-Scientist-real-torrent.html]