Vendor: Databricks

Exam Code: Databricks-Certified-Professional-Data-Scientist

Exam Name: Databricks Certified Professional Data Scientist

Version: DEMO

QUESTION 1
Suppose you have been given two Random Variables X and Y, whose joint distribution is already known, the marginal distribution of X is simply the probability distribution of X averaging over information about Y. It is the probability distribution of X when the value of Y is not known. So how do you calculate the marginal distribution of X

A. This is typically calculated by summing the joint probability distribution over Y.
B. This is typically calculated by integrating the joint probability distribution over Y
C. This is typically calculated by summing (In case of discrete variable) the joint probability distribution over Y
D. This is typically calculated by integrating(In case of continuous variable) the joint probability distribution over Y.

Answer: ABCD
Explanation:
Given two random variables X and Y whose joint distribution is known, the marginal distribution of X is simply the probability distribution of X averaging over information about Y. It is the probability distribution of X when the value of Y is not known. This is typically calculated by summing or integrating the joint probability distribution over Y. ' For discrete random variables, the marginal probability mass function can be written as Pr(X = x).
This is

$$\Pr(X=x)=\sum_{y}\Pr(X=x,Y=y)=\sum_{y}\Pr(X=x|Y=y)\Pr(Y=y),$$

where Pr(X = x,Y = y) is the joint distribution of X and Y, while Pr(X = x|Y = y) is the conditional distribution of X given Y In this case, the variable Y has been marginalized out. Bivariate marginal and joint probabilities for discrete random variables are often displayed as two- way tables. Similarly for continuous random variables, the marginal probability density function can be written as pX(x). This is

$$p_X(x)=\int_y p_{X,Y}(x,y)\ \mathrm{d}y=\int_y p_{X|Y}(x|y)\,p_Y(y)\ \mathrm{d}y,$$

where pX.Y(x.y) gives the joint distribution of X and Y while pX|Y(x|y) gives the conditional distribution for X given Y Again: the variable Y has been marginalized out.
Note that a marginal probability can always be written as an expected value:

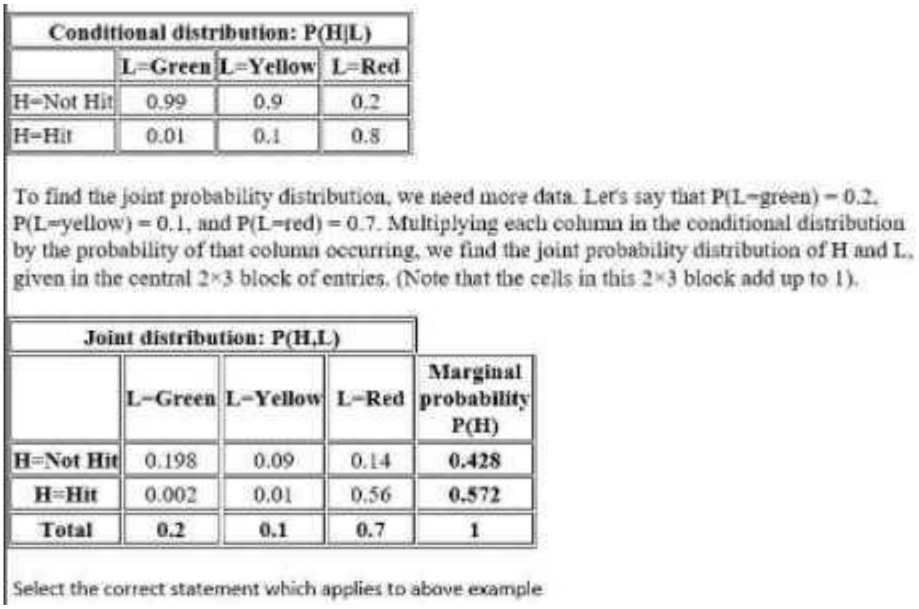
$$p_X(x)=\int_y p_{X|Y}(x|y)\,p_Y(y)\ \mathrm{d}y=\mathrm{E}_Y[p_{X|Y}(x|y)]$$

Intuitively, the marginal probability of X is computed by examining the conditional probability of X given a particular value of Y, and then averaging this conditional probability over the distribution of all values of Y This follows from the definition of expected value, i.e. in general

$$\mathbb{E}_Y[f(Y)] = \int_y f(y) p_Y(y)\, \mathrm{d}y$$

QUESTION 2
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

A. 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.
B. marginal probability that P(H=Not Hit) is the sum of the H=Not Hit row
C. marginal probability that P(H=Not Hit) is the sum of the H= Hit row

Answer: AB
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

QUESTION 3
You have modeled the datasets with 5 independent variables called A,B,C,D and E having relationships which is not dependent each other, and also the variable A,B and C are continuous and variable D and E are discrete (mixed mode).

Now you have to compute the expected value of the variable let say A, then which of the following computation you will prefer

A. Integration
B. Differentiation
C. Transformation
D. Generalization

Answer: A

QUESTION 4
RMSE measures error of a predicted

A. Numerical Value
B. Categorical values
C. For booth Numerical and categorical values

Answer: A

QUESTION 5
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

A. It means that your predictions are on average one star off of what people really think
B. It means that your predictions are on average two star off of what people really think
C. It means that your predictions are on average three star off of what people really think
D. It means that your predictions are on average four star off of what people really think

Answer: A

QUESTION 6
You are creating a regression model with the input income, education and current debt of a customer, what could be the possible output from this model.

A. Customer fit as a good
B. Customer fit as acceptable or average category
C. expressed as a percent, that the customer will default on a loan
D. 1 and 3 are correct
E. 2 and 3 are correct

Answer: C
Explanation:
Regression is the process of using several inputs to produce one or more outputs. For example The input might be the income, education and current debt of a customer The output might be the probability, expressed as a percent that the customer will default on a loan. Contrast this to

classification where the output is not a number, but a class.

QUESTION 7
In which of the scenario you can use the regression to predict the values

A. Samsung can use it for mobile sales forecast
B. Mobile companies can use it to forecast manufacturing defects
C. Probability of the celebrity divorce
D. Only 1 and 2
E. All 1 ,2 and 3

Answer: E
Explanation:
Regression is a tool which Companies may use this for things such as sales forecasts or forecasting manufacturing defects. Another creative example is predicting the probability of celebrity divorce.

QUESTION 8
RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a_______, as it is scale-dependent.

A. Between Variables
B. Particular Variable
C. Among all the variables
D. All of the above are correct

Answer: B
Explanation:
The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSE 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 9
You are creating a Classification process where input is the income, education and current debt of a customer, what could be the possible output of this process.

A. Probability of the customer default on loan repayment
B. Percentage of the customer loan repayment capability
C. Percentage of the customer should be given loan or not
D. The output might be a risk class, such as "good", "acceptable", "average", or "unacceptable".

Answer: D
Explanation:
Classification is the process of using several inputs to produce one or more outputs. For example the input might be the income, education and current debt of a customer The output might be a risk class, such as "good", "acceptable", "average", or "unacceptable". Contrast this to regression where the output is a number not a class.

QUESTION 10
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

Answer: A

QUESTION 11
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

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 12
Select the correct statement which applies to logistic regression

A. Computationally inexpensive, easy to implement knowledge representation easy to interpret
B. May have low accuracy
C. Works with Numeric values

Answer: ABC

QUESTION 13
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

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 14
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 15
In unsupervised learning which statements correctly applies

A. It does not have a target variable
B. Instead of telling the machine Predict Y for our data X, we're asking What can you tell me about X?
C. telling the machine Predict Y for our data X

Answer: AB
Explanation:
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 16
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: ABC
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?