Free Questions for Databricks-Certified-Professional-Data-Scientist by certscare

Shared by Juarez on 15-04-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You are working in a classification model for a book, written by HadoopExam Learning Resources and decided to use building a text classification model

for determining whether this book is for Hadoop or Cloud computing. You have to select the proper features (feature selection) hence, to cut down on the size of the feature space, you will use the mutual information of each word with the label of hadoop or cloud to select the 1000 best features to use as input to a Naive Bayes model. When you compare the performance of a model built with the 250 best features to a model built with the 1000 best features, you notice that the model with only 250 features performs slightly better on our test data.

What would help you choose better features for your model?

Options:

A- Include least mutual information with other selected features as a feature selection criterion

B- Include the number of times each of the words appears in the book in your model

C- Decrease the size of our training data

D- Evaluate a model that only includes the top 100 words

Correlation measures the linear relationship (Pearson's correlation) or monotonic relationship (Spearman's correlation) between two variables, X and Y. Mutual information is more general and measures the reduction of uncertainty in Y after observing X. It is the KL

distance between the joint density and the product of the individual densities. So Ml can measure non-monotonic relationships and other more complicated relationships
Mutual information is a quantification of the dependency between random variables. It is sometimes contrasted with linear correlation since mutual information captures nonlinear dependence.
Features with high mutual information with the predicted value are good. However a feature may have high mutual information because it is highly correlated with another feature that has already been selected. Choosing another feature with somewhat less mutual information with the predicted value, but low mutual information with other selected features, may be more beneficial. Hence it may help to also prefer features that are less redundant with other selected features.

Answer:

A

Question 2

Question Type: **MultipleChoice**

You are working as a data science consultant for a gaming company. You have three member team and all other stake holders are from the company itself like project managers and project sponsored, data team etc. During the discussion project managed asked you that when can you tell me that the model you are using is robust enough, after which step you can consider answer for this question?

Options:

A- Data Preparation

B- Discovery

C- Operationalize

D- Model planning

E- Model building

To answer whether the model you are building is robust enough or not you need to have answer below questions at least

- Model is performing as expected with the test data or not?
- Whatever hypothesis defined in the initial phase is being tested or not?
- Do we need more data?
- Domain experts are convinced or not with the model?

And all these can be answered when you have built the model and tested with the test data sets. Hence, correct option will be Model Building.

Answer:

E

Question 3

Question Type: MultipleChoice

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?

Options:

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)

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)

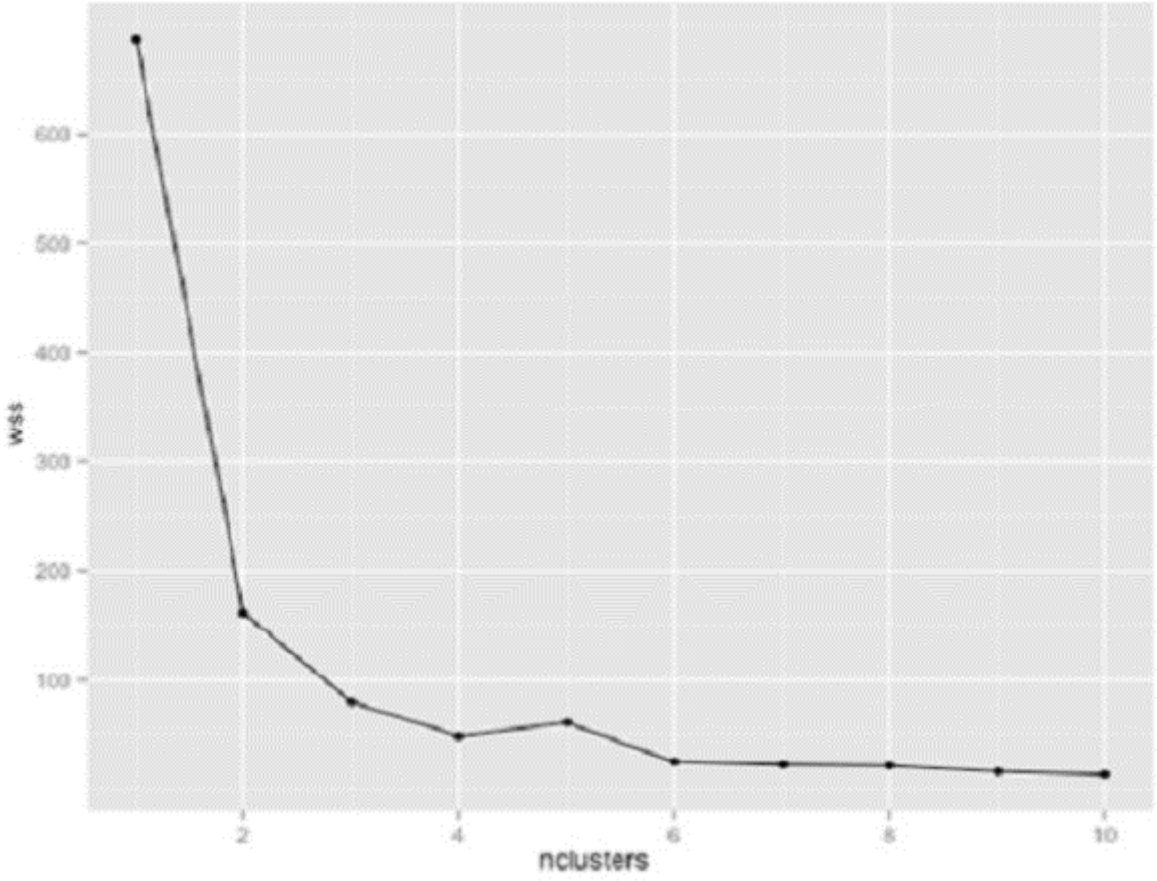
Answer:

D

Question 4

Question Type: **MultipleChoice**

Refer to the exhibit.



You are using K-means clustering to classify customer behavior for a large retailer. You need to determine the optimum number of customer groups. You plot the within-sum-of-squares (wss) data as shown in the exhibit. How many customer groups should you specify?

Options:

A- 2

B- 3

C- 4

D- 8

Answer:

C

Question 5

Question Type: MultipleChoice

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?

Options:

A- 1

B- 2

C- 0

D- n/2

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

Answer:

A

Question 6

Question Type: **MultipleChoice**

You are working on a Data Science project and during the project you have been gibe a responsibility to interview all the stakeholders in the project. In which phase of the project you are?

Options:

A- Discovery

B- Data Preparations

C- Creating Models

D- Executing Models

E- Creating visuals from the outcome

F- Operationnalise the models

During the discovery phase you will be interviewing all the project stakeholders because they would be having quite a good amount of knowledge for the problem domain you will be working and you also interviewing project sponsors you will get to know what all are the expectations once project get completed. Hence, you will be noting down all the expectations from the project as well as you will be using their expertise in the domain.

Answer:

A