AIML Online

Frequently Asked Questions in Problem Statement

Course: Unsupervised Learning

** Direct or Self-explanatory questions are not covered in this FAQ.*

Part A [30 marks]

2. A. Check and print feature-wise percentage of missing values present in the data and impute with the best suitable approach. [2 Mark]

Print each variable's percentage of missing values. If there are any missing values, impute those with a suitable approach.

2. B. Check for duplicate values in the data and impute with the best suitable approach. [1 Mark]

Check if the dataset has any duplicate rows. If there are any imputes with the suitable approach.

2. D. Visualize a scatterplot for 'wt' and 'disp'. Datapoints should be distinguishable by 'cyl'. [1 Marks]

Plot a scatterplot for 'wt' and 'disp'. In the scatter plot data points must be distinguishable by the classes present in 'cyl'. For this one can make use of the 'hue' function while plotting.

2. H. Check for unexpected values in all the features and datapoints with such values. [2 Marks] [Hint: '?' is present in 'hp']

Here as the hint given there is an unwanted data '?' present in the 'hp' variable. So here you have to find those values and convert them to nulls then you can drop those values using drop function.

This is not the same as 2.A. In 2.A you are finding % nulls from all columns and dropping them. Here you are supposed to find anomalies or unexpected values or unusual values from your data, convert them to nulls and next you have to drop them.

3. B. Plot a visual and find elbow point. [2 Marks]

Here plot a line plot to visualize K clusters (2 to 10) , next search for elbows in the plot.

3. C. On the above visual, highlight which are the possible Elbow points. [1 Marks]

Here highlight those elbow points from the above question by annotating them with arrows or any suitable markers.

3. D. Train a K-means clustering model once again on the optimal number of clusters. [3 Marks]

From 3.B you will understand which are the optimal number of clusters. Again, please build a K-means model on that K value or optimal number which you observed from 3.B.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

3. E. Add a new feature in the DataFrame which will have labels based upon cluster value. [2 Marks]

For this question, you are expected to create a new column inside your Data Frame by giving it a suitable name. Here the new column should be the labels which you get from the optimal number of clusters (Question 3.D).

3. F. Plot a visual and color the datapoints based upon clusters. [2 Marks]

From the above DataFrame with a new column of labels, choose two variables (e.g. wt V/S hp) plot a scatter plot with labels. So, that you can visualize how effectively clustering is done by the model.

3. G. Pass a new DataPoint and predict which cluster it belongs to. [2 Marks]

Here please create a new record or row and fit your optimal cluster model to it and the model will predict which cluster does this new record belong to. You have to create a synthetic row here.

Part B [30 marks]

1. C. Visualize a Pie-chart and print percentage of values for variable 'class'. [2 Marks]

Here please plot a single pie chart showing percentage of classes in the column class.

1.D. Check for duplicate rows in the data and impute with correct approach. [1 Marks]

Check if data has any duplicate rows. If present then impute with suitable approach.

2. A. Split data into X and Y. [Train and Test optional] [1 Marks]

You can fit your whole data to your model (model.fit(X,y)) Or you can split your data into train and test. It's your choice.

3. A. Train a base Classification model using SVM. [1 Marks]

Here train an SVM classification model svc.

3. B. Print Classification metrics for train data. [1 Marks]

Here print classification metrics for train data, if you have split your data into train and test. Otherwise print classification metrics of your whole data which you have fitted to your model.

3. C. Apply PCA on the data with 10 components. [3 Marks]

Here learners are expected to do PCA on all 10 components.

3. D. Visualize Cumulative Variance Explained with Number of Components. [2 Marks]

3. E. Draw a horizontal line on the above plot to highlight the threshold of 90%. [1 Marks]

Please plot n-components v/s cumulative explained variance and then draw a horizontal line at 90% cumulative variance.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

3. F. Apply PCA on the data. This time Select Minimum Components with 90% or above variance explained. [2 Marks]

Select those principal components which explains 90% and above variance and fit only those principal components on your original scaled data.

3. G. Train SVM model on components selected from above step. [1 Marks]

Fit an SVM model on the principal components selected from above step (3.G).

4. A. Train another SVM on the components out of PCA. Tune the parameters to improve performance. [2 Marks]

Please build another SVM model on n-components which explain 90% and above variance and further tune model parameters to improve performance.

Hint: You can use GridSearchCV for tuning hyper parameters. Some of the hyperparameters in SVM are C, gamma, kernel.

https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/

4. B. Share best Parameters observed from above step. [1 Marks]

From the above model, please observe and mention which are the best parameters.

5. A. Explain pre-requisite/assumptions of PCA. [2 Marks]

Please explain the prerequisites of PCA for your data, why it was necessary for your data to perform PCA.

5. B. Explain advantages and limitations of PCA. [3 Marks]

Please explain advantages and limitations of PCA.

General Queries

Outlier Treatment: *Outlier treatment is highly subjective. An observation is considered to be an outlier if that particular has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on the generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results.*

Since, it is not mentioned here to treat outliers. Prefer not to treat outliers.

**********HAPPY LEARNING**********

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited