AIML | MODULE PROJECT

Unsupervised Learning

TOTAL SCORE 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.
5. Any kind of Plagiarism will lead to 0 (zero) Marks.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and
2. '.html' (Jupyter Notebook > File > Download as > HTML)
5 Marks will be deducted if submission in any of the formats is missing.

Part A - 30 Marks

- **DOMAIN:** Automobile
- **CONTEXT:** The data concerns city-cycle fuel consumption in miles per gallon to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.
- **DATA DESCRIPTION:**

cylinders: multi-valued discrete	**acceleration**: continuous
displacement: continuous	**model year**: multi-valued discrete
horsepower: continuous	**origin**: multi-valued discrete
weight: continuous	**car name**: string (unique for each instance)
mpg: continuous	

- **PROJECT OBJECTIVE:** To understand K-means Clustering by applying on the Car Dataset to segment the cars into various categories.
- **STEPS AND TASK [30 Marks]:**

1. Data Understanding & Exploration: [5 Marks]

A. Read 'Car name.csv' as a DataFrame and assign it to a variable. [1 Mark]
B. Read 'Car-Attributes.json as a DataFrame and assign it to a variable. [1 Mark]
C. Merge both the DataFrames together to form a single DataFrame [2 Mark]
D. Print 5 point summary of the numerical features and share insights. [1 Marks]

2. Data Preparation & Analysis: [10 Marks]

A. Check and print feature-wise percentage of missing values present in the data and impute with the best suitable approach. [2 Mark]
B. Check for duplicate values in the data and impute with the best suitable approach. [1 Mark]
C. Plot a pairplot for all features. [1 Marks]
D. Visualize a scatterplot for 'wt' and 'disp'. Datapoints should be distinguishable by 'cyl'. [1 Marks]
E. Share insights for Q2.d. [1 Marks]
F. Visualize a scatterplot for 'wt' and 'mpg'. Datapoints should be distinguishable by 'cyl'. [1 Marks]
G. Share insights for Q2.f. [1 Marks]
H. Check for unexpected values in all the features and datapoints with such values. [2 Marks]
[Hint: '?' is present in 'hp']

3. Clustering: [15 Marks]

A. Apply K-Means clustering for 2 to 10 clusters. [3 Marks]
B. Plot a visual and find elbow point. [2 Marks]
C. On the above visual, highlight which are the possible Elbow points. [1 Marks]
D. Train a K-means clustering model once again on the optimal number of clusters. [3 Marks]
E. Add a new feature in the DataFrame which will have labels based upon cluster value. [2 Marks]
F. Plot a visual and color the datapoints based upon clusters. [2 Marks]
G. Pass a new DataPoint and predict which cluster it belongs to. [2 Marks]

Part B - 30 Marks

- **DOMAIN:** Automobile
- **CONTEXT:** The purpose is to classify a given silhouette as one of three types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.
- **DATA DESCRIPTION:** The data contains features extracted from the silhouette of vehicles in different angles. Four "Corgie" model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400 cars. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.
- All the features are numeric i.e. geometric features extracted from the silhouette.
- **PROJECT OBJECTIVE:** Apply dimensionality reduction technique – PCA and train a model and compare relative results.
- **STEPS AND TASK [30 Marks]:**

1. **Data Understanding & Cleaning: [5 Marks]**
   A. Read 'vehicle.csv' and save as DataFrame. [1 Marks]
   B. Check percentage of missing values and impute with correct approach. [1 Marks]
   C. Visualize a Pie-chart and print percentage of values for variable 'class'. [2 Marks]
   D. Check for duplicate rows in the data and impute with correct approach. [1 Marks]
2. **Data Preparation: [2 Marks]**
   A. Split data into X and Y. [Train and Test optional] [1 Marks]
   B. Standardize the Data. [1 Marks]
3. **Model Building: [13 Marks]**
   A. Train a base Classification model using SVM. [1 Marks]
   B. Print Classification metrics for train data. [1 Marks]
   C. Apply PCA on the data with 10 components. [3 Marks]
   D. Visualize Cumulative Variance Explained with Number of Components. [2 Marks]
   E. Draw a horizontal line on the above plot to highlight the threshold of 90%. [1 Marks]
   F. Apply PCA on the data. This time Select Minimum Components with 90% or above variance explained. [2 Marks]
   G. Train SVM model on components selected from above step. [1 Marks]
   H. Print Classification metrics for train data of above model and share insights. [2 Marks]
4. **Performance Improvement: [5 Marks]**
   A. Train another SVM on the components out of PCA. Tune the parameters to improve performance. [2 Marks]
   B. Share best Parameters observed from above step. [1 Marks]
   C. Print Classification metrics for train data of above model and share relative improvement in performance in all the models along with insights. [2 Marks]
5. **Data Understanding & Cleaning: [5 Marks]**
   A. Explain pre-requisite/assumptions of PCA. [2 Marks]
   B. Explain advantages and limitations of PCA. [3 Marks]