CS 482 MACHINE LEARNING

ASSIGNEMNT III

Nonlinear Models and Unsupervised Learning

For each step below, provide tables, visuals with explanation of what is being shown and results. Submit the report and the code. Do NOT copy and paste code into the report.

PART A: Nonlinear Models

Use the data from Kaggle below: Note that you might have to create an account to download this dataset. Use "download all" to download train and test set. It has 4 files with 163 columns.

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

1. **MEET THE DATA**: Provide a Meet the Data Section to introduce the data. Include the usual elements of

a) Number of features
b) Names of the features
c) Name of target
d) Number of samples
e) First five rows of the data

2. **SPLIT THE DATA INTO TRAINING AND TESTING:** Split the data into training and testing with 75-25 split, shuffle, random_state = 42.

3. **DATA PREPROCESSING:** Use the training data for preprocessing and not the test data. For this data set, there will be a good amount of data preprocessing and feature engineering.

a) Delete all columns that have unique values using code. Do not modify the .csv file. (list the deleted feature names.)

b) Fill in missing information in the data set if there are any. Explain why you chose to fill in the data the way you chose. Provide a table with which columns had missing information and how you filled it in and why you chose this scheme for filling in missing values

Feature Names with missing values	Percentage of values missing	How you chose to fill in missing values	Reasons for your choice

c) Converting Words to numeric values: List the columns with non-numeric values and explain what encoding you used and why.

Feature Names with word values	Encoding Used (Ordinal, one hot etc)	Reasons for your choice

c) Did you do any scaling (standard scalar, min-max scalar etc) for any numeric columns. If so, list them and explain why you chose to scale.

Feature Names of the numeric column	Type of scaling used	Reasons for your choice

4. FEATURE EXTRACTION

Use PCA (Principal component Analysis) on the features to make 10 components.

List the names the contribution of features the first two components.

5. MODEL DEVELOPMENT WITH PARAMETER TUNING:

a) Split the training data into training data and validation data (80:20) with shuffle is true and random_state=42

b) Now use SVM (rbf kernel with gamma from 1 to 10 with increment of 2 and Cost from 10 to 100 with increment of 10). State the best parameter and the RMSE and R^2 for the best parameter. Note that you MUST not use test data for parameter tuning

c) Neural network (with 1 hidden layers with varying number of units from 10 to 30 with increment of 2 with random initial weights, random_seed = 42) to get best number of hidden units for 1 layer network. Note that you MUST not use test data for parameter tuning

Model	Best Cost	Best Gamma	RMSE (Validation)	R^2 (Validation)
SVM -RBF				

Model	Best number of hidden units	RMSE (Validation)	R^2 (Validation)
NN 1-year			

6. MODEL EVALUATION

Use test data for model evaluation. If any column has missing data in test data, use the same values you used in the training data to fill in the missing values in test data.

For the two models that were tuned above with best parameters, state RMSE and R^2 for the test data.

Model	RMSE (Test)	R^2 (Test)
SVM -RBF		
NN-1-layer		

PART B: UNSUPERVISED LEARNING

Using the IRIS data set without the classification labels

a) Run the k-means clustering algorithm with 3 clusters.
b) Run the Agglomerative Clustering with 3 clusters
c) Run the DBScan algorithm with various eps until you form three clusters.

Provide for ARI and Silhouette Score for each the algorithms above.