CS6464: Concepts in Statistical Learning Theory SOFTWARE

ASSIGNMENT 2

Note: This assignment is an individual assignment.

Task 1 : Regression

PROBLEM STATEMENT

The assignment aims at predicting house prices given training and test house data of 20-dimensional features and comparing the performance of various regression methods.

TASKS:

Two regression models (one row for each group) as specified in Table 1 have to be trained using the training data (available in the file named "**kc_house_train_data.csv**") and the house prices should be predicted for the test data (available in the file named "**kc_house_test_data.csv**"). Perform 10-fold cross-validation. Compare the prediction quality between the two methods allotted.

INPUT DATA

- 20-dimensional housing data for training, 17385 samples
- 20-dimensional housing data for testing, 4230 samples

OUTPUT

- Compute the regression weights and interpret them based on the methods allotted.
- Plot the coefficient profiles of top 5 interesting features based on the largest change of the coefficients over iterations (as in Fig. 3.10 (a) in Hastie's book). Plot the coefficient profiles of each method separately. (Note: By iterations, we mean the iterations of the optimization function adopted (as in LASSO, ElasticNet, etc), or the steps (as in Stepwise regression).
- Evaluation of the models with Residual Sum of Squares (RSS) (or MSE) metric using the computed regression weights, predictors and outcome.

HINTS FOR EXCELLENCE

Additional observations and visualizations of the data and the attributes of the trained models will be given extra credit.

Roll no	Method 1	Method 2
ME17B140	Simple linear regression	Lasso regression
ME17B143	Lasso regression	Forward Stepwise Regression
ME17B112	Ridge Regression	Backward Stepwise Regression
CH20S011	Ridge Regression	ElasticNet Regression
EE17B113	Simple linear regression	Forward Stepwise Regression
EE17B114	Kernel Regression	Polynomial Regression
CS19M053	Backward Stepwise Regression	Simple linear regression
CS18M054	Lasso regression	Kernel Regression

Task 2 : Clustering

PROBLEM STATEMENT

You are given 2 dimensional datasets containing a few clusters.

a. Use K-means clustering with 5 different values of k (in range 1-10) and choose the best k using Elbow method.
b. Plot the original data. Also plot the clusters obtained with different colors for 3 values of k, including the best one.
c. Use GMM (Gaussian Mixture Model) clustering with different numbers of Gaussians (k in range 1-10) and choose the best k.
d. Plot the clusters with different colors for 3 values of k, including the best one.
e. Report the mean vector and covariance matrix of the Gaussians for the best k value.

INPUT DATA

Roll No.	Dataset for task-2
EE17B113	Data 1
EE17B114	Data 2
CS19M053	Data 3
CS18M054	Data 4
ME17B140	Data 1

ME17B143	Data 2
ME17B112	Data 3
CH20S011	Data 4

HINTS FOR EXCELLENCE

Additional observations and visualizations of the data and the attributes of the trained models will be given extra credit.

DATASET FOLDER LINK:
https://drive.google.com/drive/folders/1YcOwL0hwHwNf5zD7kXuvOfyEoJAn9Rct?usp=sharing

References:

- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4
- https://towardsdatascience.com/kernel-regression-from-scratch-in-python-ea0615b23918