School of Computer Science Engineering and Technology

Course- BTech	Type- Specialization Core
Course Code: CSET 301	Course Name- AIML
Year- 2022-23	Semester- 4th
Date-	Batch-

Lab Assignment W3L2

Objective: To understand the process of K-Fold validation by implementing it in the task of predicting the Violent Crimes Per Population in USA.

1. Dataset: Download the dataset from the link https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

The dataset contains 127 features and target (Violent Crimes Per Population). Pre-process the dataset as follows:

a. Drop: The first five attributes are non-predictive, hence drop them.

b. Missing Value Imputation: Dataset contains a lot of missing values in it.

Missing values are represented using “?”. Those values can be predicted using different methods such as replace by global constant, mean, median, mode, value from k-nearest sample, etc. using SimpleImputer() class in Scikit-Learn.

c. Define X matrix (independent features) and y vector (target feature).

2. **Perform K-Fold cross validation on the performance of a Linear Regression model. The general procedure is defined as follows:**

a. Shuffle the dataset randomly.

b. Split the dataset into k groups

c. For each unique group:

i. Take the group as a hold out or test data set

ii. Take the remaining groups as a training data set

iii. Fit a model on the training set and evaluate it on the test set

iv. Retain the evaluation scores such as such as mean squared error, mean absolute error, median absolute error, R2 score.

v. Discard the model

d. Summarize the skill of the model using the average of model evaluation scores.

Hint: Use sklearn.model_selection.KFold() class to generate K different folds.

3. Compare the performance of the following models using suitable graphs.

a. Linear Regression without K-fold cross-validation.

b. Linear Regression with 5-fold cross validation.

c. Linear Regression with 10-fold cross validation.

4. a simpler way of K-fold cross validation is to use the helper function cross_val_score() defined in the module sklearn.model_selection. Perform the same with the help of this function. (20)

Additional Fun:

Interested students can analyse the performance of the model using other cross-validation methods such as Stratified K-Fold, Leave One Out and Leave P Out etc.

Useful links

1. https://vitalflux.com/k-fold-cross-validation-python-example/#:~:text=K%2Dfold%20cross%2Dvalidation%20is%20used%20for%20model%20tuning%2F,with%20the%20most%20optimal%20performance

2. https://towardsdatascience.com/what-is-k-fold-cross-validation-5a7bb241d82f