Prof.Sangsan Lee, TA:Shinkook Cha
School of Mechanical and Control Engineering
Handong Global University, Pohang

March 18, 2024

ASSIGNMENT 2 — Regression and Classification

1 Part 1: Regression

- Download the `Abalone` dataset provided from the UC Irvine Machine Learning Repository by using the **ucimlrepo** package: `https://archive.ics.uci.edu/dataset/1/abalone`
- Divide the `Abalone` dataset into train and test set. (Note: test_size=0.2 and random_state=24)
- Apply appropriate preprocessing steps before feeding the dataset into model
- Create a Regression Model which predicts the age of the Abalone given a set of features from the test set
- Calculate the MAE (Mean Absolute Error) and the RMSE (Root Mean Squared Error) of the prediction to the actual values
- Students are highly encouraged to explore different types of regression models and tweak the hyperparameters of the given regression model
- **Bonus points are given if the test set prediction MAE or RMSE is below 1.44 and 2.25 respectively**

2 Part 2: Supervised Learning - Classification

- Divide the `ObesityDataSet.csv` into train, validation and test sets (train:validation:test = 7:1:2, random_state=24)
- Apply appropriate preprocessing steps before feeding the dataset into the model
- Create at least two classification models (e.g., SVC, Random Forest) which predicts the obesity level of a given patient
- Output a confusion matrix of the model using the test set
- **Minimum accuracy of the classification model should be 87.0% and bonus points will be given if the model accuracy is above 93.0%**
- Information regarding the dataset is detailed here `https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data`

3 Part 3: Unsupervised Learning

- Concatenate the train and validation sets from 'Part 2: Classification' to create a "new training set".
- Using K-Means Clustering, form k number of clusters in the "new training set".
- Now, count the number of sample labels in each cluster and classify the cluster as the majority label.

- For each sample in the test set from 'Part 2: Classification', calculate the euclidean distance between the sample and the centroids of each cluster and classify the test sample as the label of the closest cluster.
- Iterate the four steps above for different values of k ($k \geq$ **Number of unique labels**) (e.g., k: 4,5,6...)
- Print the number of correctly classified samples and find the optimal value of k

Submitted by Prof.Sangsan Lee, TA:Shinkook Cha on March 18, 2024.