Guvi Data Science Project

Chandan J

22 november 2023

1 Project 1 : Kannada MNIST - Classification Problem

Problem Statement: This is an extension of clasic MNIST classification problem. Instead of using Hindu numerals, lets use a recently-released dataset of Kannada digits. This is a 10 Class classification problem.
Kannada is a language spoken predominantly by people of Karnataka in southwestern India. The language has roughly 45 million native speakers and is written using the Kannada script. `https://en.wikipedia.org/wiki/Kannada`
Dataset can be downloaded from the link : `https://www.kaggle.com/datasets/higgstachyon/kannada-mnist`.
All details of the dataset curation has been captured in the paper titled: Prabhu, Vinay Uday. "Kannada-MNIST: A new handwritten digits dataset for the Kannada language."`https://arxiv.org/abs/1908.01242`

Procedure:

1. Extract the dataset from the npz file from the downloaded dataset or from the web.There are 60000 images for training and 10000 images for test. Each image is of the size 28X28.
2. Perform PCA to 10 components. So now we have train and test images in 10 dimension instead of 28X28 dimension.
3. Now apply the the following models:
    - Decision Trees
    - Random forest
    - Naive Bayes Model
    - K-NN Classifier
    - SVM

4. For each of this method produce the following metrics:
    - Precision, Recall, F1 - Score
    - Confusion Matrix
    - RoC - AUC curve
5. Try to repeat the same experiment for different component size : 15,20,25,30

2 Project 2 : Toxic Tweets Dataset : NLP Problem

This dataset has a collection of Tweets. Its labelled as Toxic - 1, Non toxic - 0. Apply the NLP methods to predict the toxicity of the tweets. Download the dataset from the following Kaggle Compitation `https://www.kaggle.com/datasets/ashwiniyer176/toxic-tweets-dataset`. All the credits to the original collectors.

Procedure:

1. Convert the CSV file to the panda data frame.
2. Convert the text to the following.
    - Bag of Words
    - TF-IDF
3. For the obtained features, apply the following methods of prediction.
    - Decision Trees
    - Random forest
    - Naive Bayes Model
    - K-NN Classifier
    - SVM
4. For each of this method produce the following metrics:
    - Precision, Recall, F1 - Score
    - Confusion Matrix
    - RoC - AUC curve

3 Project 3 : Regression Problem 1

In this problem, the task is to predict the current health (as given by the target variable) of an organism given the measurements from two biological sensors measuring their bio-markers (negative indicates that it is lesser than the average case). With this data, you are expected to try our linear regression models on the training data and report the following metrics on the test split: (a) Mean Squared Error, (b) Mean Absolute Error.

Dataset: In Folder Shared look for file p1-train.csv to train the model and p1-test.csv to test the model. Last column is the target value.

4 Problem 4: Regression Problem 2

Here, you are expected to predict the lifespan of the above organism given the data from three sensors. In this case, the model is not linear. You are expected to try several (at least 3) non-linear regression models on the train split and report the following metrics on the test split (a) Mean Squared Error, (b) Mean Absolute Error

Dataset: In Folder Shared look for file p2-train.csv to train the model and p2-test.csv to test the model. Last column is the target value.

Models to use: Suppot Vector Regression, Linear Regression (Use this for both the regression problem)