Lab Assignment 7

Pattern Recognition and Machine Learning

Part 1:-

Dataset

The dataset is available at https://github.com/mishravipul/data/blob/main/vehicle.csv

Description

The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS, which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness. Four "Corgie" model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

Original source: https://www.kaggle.com/rajansharma780/vehicle

Objective

To classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.

Features

1. compactness float average perimeter**2/area
2. circularity float average radius**2/area
3. distance_circularity float area/(av.distance from border)**2
4. radius_ratio float (max.rad-min.rad)/av.radius
5. pr_axis_aspect_ratio float (minor axis)/(major axis)
6. max_length_aspect_ratio float (length perp. max length)/(max length)
7. scatter_ratio float (inertia about minor axis)/(inertia about major axis)
8. elongatedness float area/(shrink width)**2
9. pr_axis_rectangularity float area/(pr.axis length*pr.axis width)

10. max_length_rectangularity float area/(max.length*length perp. to this)
11. scaled_variance_major_axis float (2nd order moment about minor axis)/area
12. scaled_variance_minor_axis float (2nd order moment about major axis)/area
13. scaled_radius_gyration float (mavar+mivar)/area
14. skewness_major_axis float (3rd order moment about major axis)/sigma_min**3
15. skewness_minor_axis float (3rd order moment about minor axis)/sigma_maj**3
16. kurtosis_minor_axis float (4th order moment about major axis)/sigma_min**4
17. kurtosis_major_axis float (4th order moment about minor axis)/sigma_maj**4
18. hollows_ratio float (area of hollows)/(area of bounding polygon)

Target

19. vehicle_class string Predictor Class. Values: Opel, Saab, Bus, Van

Tasks:

- Obtain the multi-class dataset from the given link
- Load the dataset
- Apply pre-processing techniques: Encoding, Scaling
- Divide the dataset into training (70%) and testing (30%)
- Build your own random forest model from scratch (using individual decision tree model from sklearn)
- Train the random forest model
- Test the random forest model
- Train and test the random forest model using sklearn.
- Compare the performance of both the models

Useful links:

- https://machinelearningmastery.com/implement-random-forest-scratch-python/
- https://towardsdatascience.com/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249
- https://www.analyticsvidhya.com/blog/2018/12/building-a-random-forest-from-scratch-understanding-real-world-data-products-ml-for-programmers-part-3/

Part-2

Congressional Voting Classification

AdaBoost is an ensemble learning method (also known as “meta-learning”) which was initially created to increase the efficiency of binary classifiers. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strong ones.

Data Set

The data can be downloaded from https://github.com/mishravipul/data/blob/main/house_votes84.csv

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

Original source: https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

Features:

1. handicapped-infants: 2 (y,n)
2. water-project-cost-sharing: 2 (y,n)
3. adoption-of-the-budget-resolution: 2 (y,n)
4. physician-fee-freeze: 2 (y,n)
5. el-salvador-aid: 2 (y,n)
6. religious-groups-in-schools: 2 (y,n)
7. anti-satellite-test-ban: 2 (y,n)
8. aid-to-nicaraguan-contras: 2 (y,n)
9. mx-missile: 2 (y,n)
10. immigration: 2 (y,n)
11. synfuels-corporation-cutback: 2 (y,n)
12. education-spending: 2 (y,n)
13. superfund-right-to-sue: 2 (y,n)

14. crime: 2 (y,n)
15. duty-free-exports: 2 (y,n)
16. export-administration-act-south-africa: 2 (y,n)

Target

17. Class Name: 2 (democrat, republican)

Objective

The main objective is to predict whether congressmen is Democrat or Republican based on voting patterns by using the decision tree with the adaboost.

Tasks:

- Obtained the dataset
- Apply pre-processing operations
- Train Adaboost model from scratch and test the model
- Train Adaboost model using sklearn
- Compare the performance of Adaboost, Random Forest and Decision Trees

Helpful resources

- Concise and informative blog post about adaboost: https://medium.com/analytics-vidhya/implementing-an-adaboost-classifier-from-scratch-e30ef86e9f1b