Summer 2021: CSEE5590 – Special Topics

Python_Lesson_2_Part_2: Machine Learning

Lesson Overview:
In this lesson we will introduce classification.
b. Classification algorithm
c. Scikit learn
d. Advanced concept related to machine learning algorithm like overfitting, underfitting, cross validation, evaluation for clustering methods

Use Case Description:
k-nearest neighbor classifier

Programming elements:
Classification

Data Set:

Dataset: Glass
Dataset description: https://www.kaggle.com/uciml/glass
he original glass identification dataset from UCI machine learning repository is a classification dataset.
The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence, if correctly identified. This dataset contains attributes regarding several glass types (multi-class). The name of target Column is **Type.**

Assignment:
(Titanic Dataset**)**

1. Find the correlation between 'survived' (target column) and 'sex' column for the Titanic use case in class.
    a. Do you think we should keep this feature?
2. Do at least two visualizations to plots to describe or show correlations. (e.g.: Survived: Class and gender).

(Glass Dataset**)**

1. Implement Naïve Bayes method using scikit-learn library.
    a. Use the glass dataset available in Link also provided in your assignment.
    b. Hold a small percentage of the data set for validation.
    c. Use **train_test_split** to create training and testing part.
2. Evaluate the model on testing part using score and
3. Use the model to Predict the classes of the validation set & mean accuracy.

```
classification_report(y_true, y_pred)
```

1. Implement linear SVM method using scikit library
    a. Use the glass dataset available in Link also provided in your assignment.
    b. Hold a small percentage of the data set for validation.
    c. Use **train_test_split** to create training and testing part.
2. Evaluate the model on testing part using score and
3. Use the model to Predict the classes of the validation set & mean accuracy.

```
classification_report(y_true, y_pred)
```

Which algorithm you got better accuracy? Can you justify why?

Online Submission Guidelines (for Online students):

1. Submit your source code and documentation to GitHub and represent the work in a ReadMe file properly (submit your screenshots as well. The screenshot should have both the code and the output)
2. Comment your code appropriately
3. Video Submission (1 – 3 min video showing the demo of the assignment, with brief voice over on the code explanation)

Note: *Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: https://catalog.umkc.edu/special-notices/academic-honesty/*