AIN429 Data Mining Laboratory

Assignment 4: Classification
Date Issued : 03.01.2023
Date Due : 10.01.2023

Aim of the Experiment

In this assignment, we will focus on classification, which is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. You are required to implement the 3 different classification algorithms. The assignment should be implemented as a single Jupyter Notebook. Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results. At the end of this exercise, you will become familiar with classification method using Python libraries.

Classification

Classification is a supervised machine learning process that involves predicting the class of given data points. Those classes can be targets, labels or categories. In classification, a program uses the dataset or observations provided to learn how to categorize new observations into various classes or groups. For instance, 0 or 1, red or blue, yes or no, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes.

Classification Algorithms can be further divided into the mainly two category:

- Linear Models
    - Logistic Regression
    - Support Vector Machines
    -
- Non-linear Models
    - K-Nearest Neighbors
    - Kernel SVM
    - Naïve Bayes
    - Decision Tree Classification
    - Random Forest Classification

Types of Classification in Machine Learning

1. Lazy Learners

Lazy learners store the training data and wait until testing data appears. When it does, classification is conducted based on the most related stored training data. Compared to eager learners, lazy learners spend less training time but more time in predicting.
Examples: K-nearest neighbor and case-based reasoning.

2. Eager Learners

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Because of this, eager learners take a long time for training and less time for predicting.
Examples: Decision tree, naive Bayes and artificial neural networks.

Experiment

1. Download the dataset. The dataset will be shared on the Piazza group.
2. Perform preprocessing steps that may be necessary to clean or filter the data.
3. Analyze the dataset using tables and graphs.
4. Clearly explain analysis results.
5. Split the datasets into training and test sets.
6. Apply the 3 different classification algorithms of your choice.
7. Present the classification results (classification accuracy, confusion matrix).
8. Compare the performance of classification algorithms using tables and graphs.
9. Summarize and interpret your results.
10. You should submit your codes and report as a single Jupyter notebook.

Background information

We provide with you some references related to classification .

- https://www.datacamp.com/blog/classification-machine-learning
- https://www.edureka.co/blog/classification-in-machine-learning/
- https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning
- https://www.analyticsvidhya.com/blog/2021/09/a-complete-guide-to-understand-classification-in-machine-learning
- https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2
- https://medium.com/analytics-vidhya/building-classification-model-with-python-9bdfc13faa4b

Grading

You will present your projects during laboratory hours.

- Import dataset and Preprocessing (%15)
- Visualization (%15)
- Implementing methods (%40)
- Report (%30)

REMARKS:

- Submission format:
    - studentID_name_surname_hw4.ipynb

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use an online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ Deadline is 23:59. No other submission method (such as; CD or email) will be accepted.
- Do not submit any file via email related to this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.