AIN429 Data Mining Laboratory

Assignment 4: Classification and Clustering

Due Date: 19/12/2024 23:59

Aim of the Experiment

In this assignment, you will explore **classification**, a supervised machine learning method where the goal is to predict the class label of input data. You will also explore **clustering** (an unsupervised learning method) to analyze and group data. You will implement three classification algorithms each, evaluate their performance, and compare the results. The assignment should be implemented as a single Jupyter Notebook, with clear documentation using comments and markdown cells to explain the code and results.

Part 1: Dataset Import and Preprocessing (5 Marks)

Tasks:

1. **Dataset Loading**: Download the dataset and load it into your Jupyter Notebook.
2. **Preprocessing**:
    - Handle missing values and outliers, if any.
    - Normalize or standardize features as necessary for the chosen algorithms.
3. **Exploratory Analysis**:
    - Analyze the dataset using summary statistics, tables, and visualizations to understand its structure and distributions.

Part 2: Visualization and Data Analysis (5 Marks)

Tasks:

1. Visualize the dataset using techniques such as:
    - Histograms or box plots to show feature distributions.
    - Pair plots or scatter plots to illustrate relationships between features.
    - A heatmap for the correlation matrix to identify relationships between numerical variables.
2. Clearly interpret your visualizations and provide insights about the dataset.

Part 3: Implementing Classification Algorithms (35 Marks)

Tasks:

1. **Split the dataset**:
    - Divide the dataset into training and testing sets.
2. **Choose and Apply Three Algorithms**:
    - Implement three classification algorithms of your choice (e.g., Logistic Regression, Decision Tree, K-Nearest Neighbors, etc.).
    - Train each model on the training set and test it on the testing set.
3. **Evaluation**:
    - Use a classification report (precision, recall, F1-score) and a confusion matrix to evaluate the performance of each model.
    - Visualize the performance using tables or graphs.
    - Dataset has many features, some features might be making classification too obvious, find most important features and check if they disproportionately define certain classes.
    - Remove or transform features that make predictions overly obvious.

Part 4: Implementing Clustering Algorithms (35 Marks)

Tasks:

1. **Apply Three Clustering Algorithms**:
    - Implement three clustering algorithms of your choice (e.g., K-Means, Hierarchical Clustering, DBSCAN).
    - Discuss the strengths and weaknesses of each algorithm with respect to the dataset.
2. **Normalization and Feature Selection**:
    - Apply normalization and feature selection methods and assess their effects on clustering performance.
3. **Evaluation**:
    - Use metrics to evaluate the clustering results.
    - Visualize the clusters using scatter plots or dendrograms.
    - Compare clustering results with and without the features that made classification "too obvious".

Part 5: Comparison and Analysis (20 Marks)

Tasks:

1. Compare the performance of the three algorithms based on their metrics and explain why one might outperform the others for classification and clustering separately.
2. Discuss the strengths and weaknesses of each algorithm in relation to the dataset.
3. Summarize your findings and propose which algorithms are best suited for classification and clustering separately.

Grading Breakdown

- **Dataset Import and Preprocessing**: 5%
- **Visualization and Data Analysis**: 5%
- **Implementing Classification Algorithms:** 35%
- **Implementing Clustering Algorithms:** 35%
- **Comparison and Analysis**: 20%

Note: Since this assignment is a combination of two assignments, it will be worth two assignments.

Presentation

You will do an online presentation of your implementation of this assignment.

Submission Requirements

- **Format**: Submit your assignment as b`<studentID.zip>` containing:
  - A single Jupyter Notebook (`assignment_4.ipynb`).
- **Late Penalty**: 10% per day, up to 2 days.
- **Submission Method**: Use https://submit.cs.hacettepe.edu.tr.
  - **Not getting 1 from the submit system will lose you points!**
- **Important** 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.
- **Questions** You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.

Helpful Links

- 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://scikit-learn.org/stable/modules/clustering.html

- https://www.javatpoint.com/clustering-in-machine-learning
- https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
- https://github.com/krasserm/machine-learning-notebooks
- https://machinelearningmastery.com/clustering-algorithms-with-python/