LAB-EXPERIMENT-2-29-07-2024(SWE4012)

Lab Question: Concept Learning in Python

Total Marks: 100

Objective:

To understand and implement the concept of concept learning through a real-world problem by using Python. You will be working on a dataset, applying concept learning techniques, and analyzing the results.

Problem Statement:

You are tasked with predicting whether a given email is "Spam" or "Not Spam" based on various features of the email. The dataset you will be working with contains information about emails, including various attributes that might indicate whether an email is spam or not.

Dataset:

The dataset `email_data.csv` contains the following columns:

1. `subject_length` - The length of the email subject line (integer).

2. `num_links` - The number of hyperlinks in the email (integer).

3. `num_attachments` - The number of attachments in the email (integer).

4. `contains_offer` - A binary feature indicating if the email contains a special offer (0 = No, 1 = Yes).

5. `is_spam` - The target variable indicating if the email is spam (0 = Not Spam, 1 = Spam).

Tasks:

1. Data Preprocessing (20 marks):

- Load the dataset from `email_data.csv`.
- Display few rows of the dataset
- Check for any missing values and handle them appropriately.
- Perform exploratory data analysis (EDA) to understand the distribution of features and target variable.
- Normalize or standardize the features if necessary.

2. Concept Learning (30 marks):

- Implement the concept learning algorithm to determine the concept of spam emails. Use the Find-S algorithm for this task.

- Write a Python function `find_s_algorithm` that takes the dataset and outputs the most specific hypothesis (concept) for identifying spam emails.

```python
def find_s_algorithm(data):
    # Implement the Find-S algorithm here
    # Return the most specific hypothesis
    pass
```

3. Model Training and Evaluation (30 marks):

- Split the dataset into training and testing sets (80% training, 20% testing).

- Apply the concept learned from the Find-S algorithm to classify the test set.

- Evaluate the performance of your concept learning model using accuracy, precision, recall, and F1 score. Report these metrics.

```python
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
# Load and split data
# Implement evaluation metrics here
```

4. Interpretation and Discussion (20 marks):

- Discuss the concept learned by the Find-S algorithm. How well does it capture the concept of spam emails based on the given features?

- Provide insights into the limitations of the concept learning approach and potential improvements.

Submission Guidelines:

-Name of the submission file should be LAB-EXP-2-Your Registration Number.docx

- Submit a Jupyter notebook or Python script containing your implementation (word file with outputs).

- Include comments in your code explaining each step.

- Ensure that your notebook/script runs without errors and includes the final output and discussion.

Additional Notes:

- Use Appropriate Python libraries such as `pandas`, `numpy`, and `sklearn` for data manipulation and evaluation.

- Make sure to include visualizations where necessary to support your analysis.

Marking Scheme:

- Data Preprocessing: 20 marks

- Concept Learning Implementation: 30 marks

- Model Training and Evaluation: 30 marks

- Interpretation and Discussion: 20 marks

Dr.Trilok Nath Pandey

SCOPE,VIT,Chennai