Machine Learning - Assignment 1

Publication date: 24/05/2024
Due date: 13/06/2024

Data

In this exercise, you will be working with a dataset containing information about various laptops listed on eBay. The dataset includes details on the brand, specifications, condition, and price of each laptop. The data has 2,939 entries divided into 10 columns. Below is an explanation of each feature in the dataset.

Variables

Column	Description
Brand	The manufacturer of the laptop.
Product_Description	A raw description of the laptop, extracted from eBay.
Screen_Size	The diagonal size of the laptop's display, measured in inches.
RAM	The amount of Random Access Memory (RAM) in the laptop, measured in gigabytes (GB).
Processor	The type and generation of the laptop's central processing unit (CPU).
GPU	The graphics processing unit (GPU) present in the laptop.
GPU_Type	Indicates whether the GPU is integrated or dedicated.
Resolution	The display resolution of the laptop screen. Resolutions in the dataset are written in the following format ‘width x height’.
Condition	The physical and operational state of the laptop, as one of the following options: New, Open box, Excellent - Refurbished, Very Good - Refurbished, Good - Refurbished.
Price	The cost of the laptop in USD.

Data Split

- **Training set**: rows 0 to 2,057 (inclusive).
- **Validation set**: rows 2,058 to 2,498 (inclusive).
- **Test set**: rows 2,499 to 2,938 (inclusive).

Requirements

Section A - Coding (40 pts)

Decision Tree

Implement a **Decision Tree** classifier and regressor algorithm in Python.

Random Forest

Implement a **Random Forest** classifier and regression algorithm in Python. For the regressor, at test time, implement average.

Section B - Data Viz & Preparation (10 pts)

Data Visualization

Explore the data using visualizations. The goal of this section is to get insights on the data which may or may not be relevant for the following sections.
Plots should have an informative main title, axis labels and a legend. For each plot, provide a short description of key observations. You are required to have a minimum of three plots.

Data Prep

Some features have missing values, handle them accordingly and explain your choice of dealing with them.

Section C - Implementation (30 pts)

Classification

Use both models from section A and predict the laptop condition as one of two options
(1) *'New'* which includes the categories 'New' and 'Open Box'.
(2) *'Refurbished'* which includes the categories 'Excellent - Refurbished', 'Very Good - Refurbished', and 'Good - Refurbished'.
Using the following features: Brand, Screen_Size, RAM, Processor, GPU, GPU_Type, Resolution, Price.

Regression

Use both models from section A and predict the laptop price, using the following features: Brand, Screen_Size, RAM, Processor, GPU, GPU_Type, Resolution, Condition.

Section D - Comparison (20 pts)

Sklearn Models

Implement the models from section C (including hyperparameter tuning) using the built-in functions from the sklearn library.

Comparison

Compare the results of your program and the built-in sklearn models in terms of metrics and runtime. If there are large differences, suggest an explanation to why.
Compare the results of the random forest regressor when using average and median at test time, comment on the differences.

Section E - Bonus (20 pts)

Screen Resolution

Split the Resolution column into two different numerical columns of height and width, **implement** both the classification and regression tasks again using your models and **compare** the results to Section B results. **Discuss** the changes.

Classification Metric

Report the sensitivity and specificity metric of section C. Is there a significant difference between the scores? Suggest an explanation as to why this may be the case. Suggest a method to improve the score.

Random Forest Regression Test

Change your implementation of testing the random forest regression to median (section A), rerun the implementation in Section C. Compare the new results with using the average. Discuss the changes and reflect on them.

Guidelines

Please read the following section carefully before submitting the assignment.

Coding Guidelines

- Each model should be trained on the training and validation data.
- Each model should predict the label of the test data and report the following measures, you can use the built-it functions in Sklearn.
    - Accuracy for classification.
    - MSE for regression.
- Impurity measures are Gini for classification and SSR for regression.
- Use familiar packages with explicit explanations. Except for parts where you were explicitly asked to use the Sklearn library, do not use this library or any other machine learning library).
- Try to minimize the usage of loops, lists and other inefficient programming habits. Numpy library has a lot of useful built-in functions. In this case, consult with the library documentation and your chosen search engine.
- If you have installed any libraries beyond those presented in the exercises, please specify this in the report.
- The code should run without warnings or errors.
- Good documentation is **critical**.
- Indicate the exercise sections in the code as well.
- Use meaningful variable names.
- Do not use reserved words.
- Use constants where possible.

Submission Guidelines

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including all the sections. One in **.ipynb** format and one in **.html** format. **Both files should also include the program's outputs**.
- The files' names should be of the form: **ML_HW1_ID1_ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submission will not be accepted.

Grading

You can get more than 100 points for the exercise. The exercise will be graded according to the following criteria:

- Correctness.
- Clarity.
- Efficiency of implementation.
- Elegance of implementation.

Self-learning

As we mentioned at the beginning of the course, self-learning is an important part of the course. Treat all sources of information carefully and critically.
Usage of LLMs is allowed, but reference it when it was used.
You can and should consult with other students in the course, but each pair must write their own work.
It is reasonable to assume that not all results and algorithms will be identical.

Questions and Reception hours

- Please post your questions on the exercise forum in Moodle, after you have read the previous posts. Professional questions sent by email will not be answered.
- If you want to schedule a reception hour with one of the instructors - please send your questions by email in advance.
- In any other case (personal questions, request for an extension with a justified reason, etc.) please email the instructor.

Good luck