Machine Learning - Assignment 3

Publication date: 09/07/2024

Due date: 15/08/2024

Data

The Wine Quality Dataset, sourced from the UCI Machine Learning Repository, contains physicochemical properties of wine samples, as well as their quality ratings (continues number between 0 and 10). The dataset contains 6,497 instances.

Data source: [Wine Quality - UCI Machine Learning Repository]

A detailed explanation of each column can be found in the Kaggle page of the dataset.

Data import:

```
pip install ucimlrepo
```

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
wine_quality = fetch_ucirepo(id=186)

# data (as pandas dataframes)
X = wine_quality.data.features
y = wine_quality.data.targets

# metadata
print(wine_quality.metadata)

# variable information
print(wine_quality.variables)
```

Requirements

Section A - Data Exploration & Visualization *(20 pts)*

1. Explore the data using tables, visualizations, and other relevant methods (use at least 3 different types of graphs).

   Your graphs should answer the following questions:

   - What is the distribution of wine quality scores?
   - How do different physicochemical properties correlate with wine quality?

   Except for these questions, add at **least 3 more** visualizations of your own that show interesting insights on the data, explain which questions they answer.
2. Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.
   a. Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.
   b. For sections B & C a categorical target is needed, create a categorical wine quality feature. Explain your choice of conversion and number of categories, make sure the resulting categories are all almost equally represented.

Section B - Dimensionality Reduction *(40 pts)*

In this section, use the physicochemical properties of wine and apply dimensionality reduction on them, in order to identify the different wine quality **groups**.

1. Apply at least 2 dimensionality reduction algorithms.
2. For each algorithm result:
   a. Create a scatter plot for the new data and color each observation according to the quality of the wine. Describe your findings (which wines are most similar to one another).

b. Which features are the **most effective** to separate the wine by their quality? Explain your findings.
c. Which features are the **least effective** to separate the wines? Explain your findings.
d. Using the algorithm components that explain the majority of the variance in the data, identify the features that are most strongly represented in each component (in absolute values). Show it visually.

3. Choose one of the algorithms above (the one with better separation), and the results of task 2.b above (identifying the most effective features). Reapply your dimensionality reduction algorithm, but this time without using the most effective feature.
    a. Explain what changed in the results, and how the separation was affected.
    b. Explore and show visually which features now are the most effective.
    c. Create a biplot using the two algorithm implementations (task1 and task3). Interpret the biplot: examine the position of the data points in the biplot and their relationships to the variables (features).
    Interpret the biplot by considering the following aspects: proximity of data points, angle, and direction of vectors, variables' contribution, etc.
4. **Bonus** *5 pts*: Using a dimensionality reduction algorithm, find outliers in your dataset. Print a list of the outliers and explain how you found them and what they have in common.

Section C - Classification of Wine Quality (*20 pts)*

Predicting the quality of wine. In this section, the objective is predicting the categorical quality of wine (you can use the same categories you built in the previous section).

1. Apply **SVM** and at least 2 more machine learning algorithms.
    a. Provide feature importance for each model (if possible) and explain if the important features make sense.
    b. The implementation should include parameter tuning.

c. Compute accuracy for each model and provide sensitivity and specificity measurements for every class in each model. Is there a class in which one of the measurements is relatively low? If so, explain why, and implement a suitable method to improve the results and explain it.
d. Compare and **discuss** the performance results between the models.

Section D - Regression of Wine Quality (*20 pts)*

Predicting the quality of wine. In this section, your goal is to build a regression model to predict the numerical quality of the wine column.

1. Apply at least 3 different machine learning algorithms.
    a. Provide feature importance for each model and explain if the important features make sense.
    b. The implementation should include parameter tuning.
    c. Report suitable measurements to evaluate the performance of each model
    d. Compare and **discuss** the results.

Section E - Bonus (*15 pts)*

<u>Task 1</u>

The wine experts are interested in identifying excellent and poor wine bottles, being:

- Excellent wine: quality of 8, 9, 10.
- Poor wine: quality of 0, 1, 2, 3, 4.
- Normal wine: all the rest.

Suggest an untraditional method for identifying excellent and poor wine (not classification or regression model)

a. Explain your method of choice and your intuition.
b. Implement the suggested method.
c. The implementation should include parameter tuning (if applicable).
d. Report suitable measurements to evaluate the performance of the method.

Task 2

In the *sectionE-bonus-task2-data zip*, you should find two CSV files, each containing the rows belonging to red and white wine.

a. Suggest and implement a classification model to classify the red/white groups.
b. Explain any/all data preprocessing techniques you chose to implement.
c. Report the metrics of the trained model. Discuss ways to enhance its performance.
d. Are there are features that are unique per wine group?

Section F - Performance Bonus (*5 pts)*

Machine learning models that outperformed other students' models for either the classification or regression tasks may get additional points, as long as the non-standard methodology to obtain superior results is also explained.

Guidelines

Please read the following section carefully before submitting the assignment.

Coding Guidelines

- Each model should be trained on the training and validation data.
- Use familiar packages with explicit explanations.
- 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_HW3_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 correctness, clarity, efficiency of implementation, and 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