2104618 Project Assignment for Second Semester of Academic Year of 2566

Problem Statement

The problem is about the quality of fruit. You are asked to develop the classification models to predict the fruit quality from the provided features.

The dataset contains seven features which are information about various attributes of fruit, providing insights into its characteristics and the label which is the quality of fruit as follows:

Features (Attributes)

- Size: Size of the fruit
- Weight: Weight of the fruit
- Sweetness: Degree of sweetness of the fruit
- Crunchiness: Texture indicating the crunchiness of the fruit
- Juiciness: Level of juiciness of the fruit
- Ripeness: Stage of ripeness of the fruit
- Acidity: Acidity level of the fruit

Predict variable (desired target)

- quality – 1 if the fruit quality is good, 0 if the fruit quality is bad ok

Instruction

- The dataset contains 4,000 examples and is randomly divided into three parts – training data (2,400 samples – 60%), cross validation data (800 samples – 20%) and test data (800 samples – 20%). ✓
- Name of each column can be seen from the provided MS-Excel file.
- Develop the forecasting models to predict whether the fruit quality is good or bad by using the training data. After that you compare the models using the cross validation data to select the best model. Finally, you determine the generalization error of the selected model using the test data. ✓
- You can use any supervised learning algorithms – e.g. logistic regression (with / without regularization), neural networks, support vector machine, random forest, decision tree, gradient boosting, XGBoost, LSTM, and so on.
- You should develop **at least** 3 algorithms to compare. In each algorithm, you can develop several models by varying the hyperparameters or structures. I did 4.
- You do not have to use all features. You can use just a subset of the provided features if you want. ✓ I used it all.
- You should explore the features in the preprocessing step. If there are any missing values, you can impute them, e.g. replace with mean, median, mode, previous value, next value, etc. ✓ Data is already cleansed.
- You can use any degree of polynomial for any selected feature and you can add interaction terms if necessary. Nah

- You can use any algorithm to train the models – e.g. scipy.optimize module, gradient descent, etc.
- You can use any python library.
- Feel free to set the values of hyperparameters by yourself – e.g. regularization parameter, learning rate, threshold, number of hidden layers, number of neurons in each hidden layer and so on. Please clearly state the hyperparameters and their values you use in each model.
- You can use GridSearchCV, RandomSearchCV or BayesSearchCV to find the optimal combination of hyperparameters.
- If you use the neural networks, you can use any activation function – e.g. ReLU, ELU, PReLU, Sigmoid, tanh, etc.
- It would be good to perform the machine learning diagnostic – e.g. check bias/variance, plot learning curve, plot debugging, etc.
- You may evaluate your models by any criterion – e.g. classification accuracy, classification error, F_1 score, precision, recall. Please give the reason to support your decision for choosing the selected criterion.
- Prepare the report and hand in it on the presentation days – Thursday, April 25, 2024 (for Section 1) and Sunday, May 5, 2024 (for Section 9). Besides, please send the jupyter notebook files of all models to instructor's e-mail at nantachai.k@chula.ac.th.
- The report should have
    1. Table of Contents
    2. Table of Tables
    3. Table of Figures
    4. Problem Statement
    5. Literature Review
    6. Methodology
    7. Results and Discussion
    8. References
- Prepare the presentation. Each presentation will be 15 minutes, followed by a 5-minute period for questions and discussion.
- **Honor code: You agree to do this project assignment by yourself without the help of others.**