ELG 5255 Applied Machine Learning Fall 2021

Assignment 1 (Multiclass Classification)

Start Date: $Sep\ 22^{nd}\ 2021$

Due Date: $Oct\ {}^{2}st\ 2021\ 23:59$ **Eastern Time (US and Canada)**

Submission

You must submit your assignment on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit a assignment passed the deadline. It is student's responsibility to ensure that the assignment has been submitted properly. A mark of 0 will be assigned to any missing assignment.

Goal

This assignment aims to implement One-versus-Rest (OvR) strategy transforming multiclass classification problems to multiple binary classification problems.

Dataset

During this assignment, Seeds dataset is used. Training and test splits are provided in csv file format.

In python:

```python
import pandas as pd
seeds_train = pd.read_csv("file path/seeds_train.csv")
seeds_test = pd.read_csv("file path/seeds_test.csv")
```

Seeds dataset has 7 attributes (features) and 3 classess (Kama, Rosa and Canadian) that are named as numerical index 1,2 and 3, respectively. In order to visualize data and have better intuition, first and fifth features, that are more meaningful, will be used.

Before building OvR strategy, the performance of binary classification models (Perceptron and Support Vector Machine (SVM)) should be compared; hence, the first class (Kama) should be dropped to form 2 class dataset.

One-versus-Rest (OvR)

OvR strategy involve training a set of binary classclassifiers for each class. During testing process, each classifier will predict the confidence of each class and the one with highest confidence will be selected.

Training:

```
Inputs: X, y, estimator
    yBin = binarize(y)
    build a list of estimators for each class
Output: a list of estimators
```

Prediction:

```
Inputs: X, a list of estimators
    argmax of each estimator's confidence score on X
```

Questions

Note the difference among Seeds dataset (with 7 features and 3 classes), 2D Seeds dataset (with 2 features and 3 classes). Please use the provided training and testing set.

Please submit your code and report (including screen shot of code and the relevant figures).

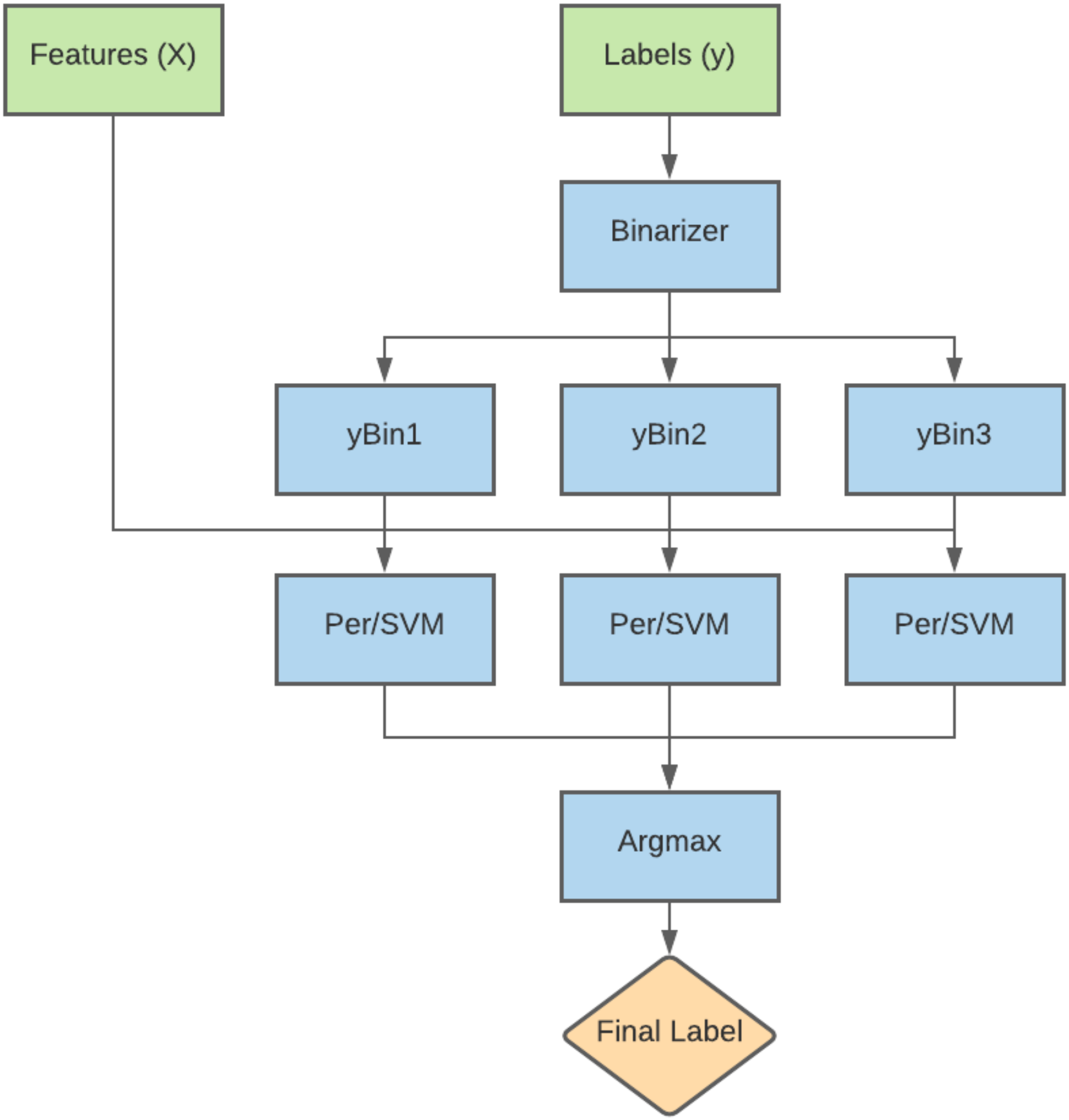
1. Load the Seeds dataset.

2. Form 2 class-Seeds dataset by removing first class (Kama), compare performance of Perceptron and SVM on testing set. Provide accuracies, confusion matrix for both model and make commnets on the performance. (**8 marks**)

3. Build OvR-Perceptron and OvR-SVM and test on Seeds testing dataset (which contains 3 classes).(**18 marks for each classifier; 54 marks in total**)

   For each binary classifier:

   - Obtain the binarized labels.1 for positive class, -1 for negative class (OvR). (*2 marks*)
   - Obtain the Percepton's confusion matrix and accuracy (*3 marks*)
   - Obtain the SVM's confusion matrix and accuracy (*3 marks*)
   - Plot Perceptron's decision boundary (*4 marks*)
   - Plot SVM's decision boundary (*4 marks*)
   - Compare performance of two models and make comments for each class at the end. (*2 marks*)



4. Use argmax to aggregate confidence scores and obtain the final label and obtain the performance (i.e., confusion matrix, accuracy, plotting correct and wrong prediction points) of OvR-Perceptron (**5 marks**) and OvR-SVM (**5 marks**)
    - Accuracy (*1 marks* for each model)
    - Confusion Matrix (*2 marks* for each model)
    - Plotting correct and wrong prediction points (*2 marks* for each model)

5. Improve an alternative aggregation strategy instead of existing argmax function based OvR, using the third step's results, obtain the performance (i.e., confusion matrix, accuracy, plotting correct and wrong prediction points) of your own strategy, and explain why your strategy is better/worse than existing OvR (**20 marks**)
    - Aggregation strategy design *(10 marks)*
    - Explanation of your own strategy *(3 marks)*
    - Confusion matrix, accuracy, plotting correct and wrong prediction points *(5 marks)*
    - Performance comparison of argmax and your own strategy (*2 Marks*)

6. Provide a conclusion section on your report. Include overview of what you have done and learnt during the assignment. Aim no less than one third of a page and no more than half page. (**8 Marks**)
    - Models (Perceptron and SVM)
    - OvR strategy
    - Argmax
    - Aggregation Strategy etc.