Exercise – 4: DS203-2024-S2

Submissions due by: February 28, 2025, 11:55pm

This exercise is aimed at:

- Getting introduced to and running various **classification** algorithms on a given data set and understanding their relative characteristics, performance, and advantages.
- Calculating, effectively documenting, and understanding various classification metrics and developing an approach towards effectively using them.
- Creating and consolidating multiple plots to compare the results of different algorithms.
- Get introduced to the relevant functions of the Python library: **scikit-learn (sklearn).**

In this exercise you will be working with two datasets – **circles** and **moons.**

Processing steps:

1. You are provided with some utility functions to interact with the datasets (E4.ipynb). Understand and document what each function does in the report.

Section 1 - 2-class classification

(Work with the moons dataset for questions 2 to 5)

2. Fit the following models on the dataset and generate the decision boundary for each (you can use `sklearn` functions with their default arguments, unless otherwise specified)
    a. Logistic Regression *(Has already been fit as an example)*
    b. SVC – with `'linear'` kernel (what is `'linear'`?)
    c. SVC – with `'rbf'` kernel (what is `'rbf'`?)
    d. Random Forest Classifier
    e. Neural Network Classifier – with `hidden_layer_sizes=(5)`
    f. Neural Network Classifier – with `hidden_layer_sizes=(5,5)`

3. Compute and show the confusion matrices for each of the models. Calculate the accuracy, precision, recall and F1 score for each of the above trained models manually (show calculations), and tabulate them. Interpret the numbers.

4. We will now perform grid search to obtain the best setting of hyperparameters (C and `gamma`) for SVC (with `'rbf'` kernel).
   Generate all possible combinations with **C ∈ {0.1, 1, 10}** and **`gamma` ∈ {0.1, 1, 10}** and attach the 9 plots showing the decision boundaries obtained. What do the plots say about the roles of these hyperparameters? (think in terms of under/overfitting). Which is the best setting among the 9 tested? Support your answer with accuracy and F1 scores.

5. We see the Logistic Regression does not work very well since the two classes are not linearly separable. One way to fix this is to add higher-order features (x^2, x^3 etc.) so that the decision boundary is not linear. Looking at the dataset, what do you think is the minimum degree of added features needed for the classification to perform well? Support your answer by adding the required features (see `PolynomialFeatures()` from `sklearn`) and fitting the model again.

Section 2 - Multi-class classification

(Work with the circles dataset for questions 6 to 7)

6. Use the following algorithms / variants to process the dataset:
    a. SVC – with `'linear'` kernel
    b. SVC – with `'rbf'` kernel
    c. Decision Tree Classifier

d. Random Forest Classifier – with `min_samples_leaf=1`
e. Random Forest Classifier – with `min_samples_leaf=5`
f. Neural Network Classifier – with `hidden_layer_sizes=(5,5)`

Generate the decision boundary for each model used above.

7. Create a table with the following metrics for each model (similar to the 'Classification Report' mentioned in class), and then compare the models:
   - Accuracy
   - Precision (per class), Precision (micro, macro, weighted average)
   - Recall (per class), Recall (micro, macro, weighted average)
   - F1-score (per class), F1-score (micro, macro, weighted average

   (Hint: The following functions may be used: `accuracy_score`, `precision_score`, `recall_score`, `f1_score`)

8. ***BONUS**: To sample a point uniformly from an annulus, instead of picking* $r \in \{r1, r2\}$ *and* $\Theta \in [0, 2\pi)$ *uniformly randomly, we instead pick* $r^2 \in \{r1^2, r2^2\}$ *randomly (see line 2 of* `sample()`*), and then take a square root. Think about why this ensures a better distribution of points in an annulus. You can support your answer using scatter plots showing distribution of points sampled from an annulus, using both methods, and discuss your observations.*

9. List your major learnings from this exercise.

10. Your submissions should include the following:
    a. A PDF document with all the above analyses and comments. Ensure that you include the required figures and tables (i.e. metrics data) in your report, along with the explanations and analysis.
    b. Your Python source file (.py file). Please DO NOT upload Jupyter Notebooks – they get bulky!
    c. Name of the PDF should be **E4-your-roll-number.pdf** and the name of the Python source file should be **E4-your-roll-number.py**
    d. Upload the PDF and the source file to the assignment submission point E4.

oooOOOooo