CPEN 355 Homework 3

Aloysio Kouzak Campos da Paz
Student ID 58687526

Collaboration statement I completed the assignment with the help of the internet and ChatGPT for understanding the concepts behind the assignment and for generating Python code.

Here is a link to the chat I had with ChatGPT:
https://chat.openai.com/share/8e7d9274-ff4f-4f59-9919-d927310bfd65

1 Coding Practice

We will train three different classification models using the same dataset to compare their accuracies:

- A linear kernel support vector machine (SVM)
- A Radial Basis Function (RBF) kernel SVM
- A random forest classifier

We will use the same algorithm to train the three models. Here's some pseudo code explaining it

```
for parameter in list of parameters:
  instatiate model
  fit the model using parameter
  get prediction
  evalutate prediction accuracy using testing data
  save accuracy in accuracies list

plot model accuracy vs. the parameter that trained that model
```

Then we will cross validate the results to select the best hyper-parameters for each model

Note We train all the machine learning models using normalized features. We normalize the features by making their mean 0 and variance 1.

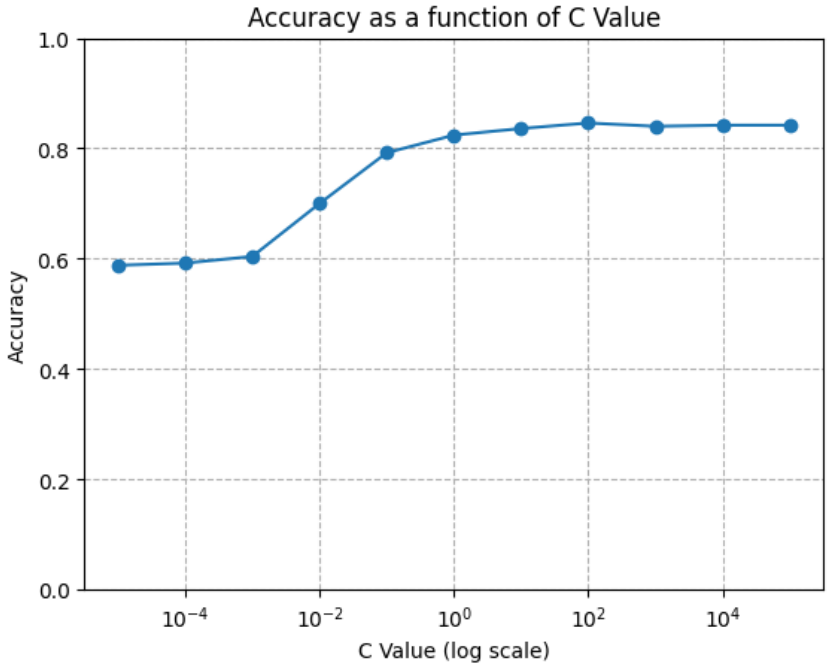
1.1 Question 1 a: linear kernel SVM

First, we will train a linear kernel SVM using the function sklearn.svm.LinearSVC

We can vary C in the range of

$$C = (10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{5}) \tag{1}$$

and see how the accuracy of the prediction changes with each value of C.



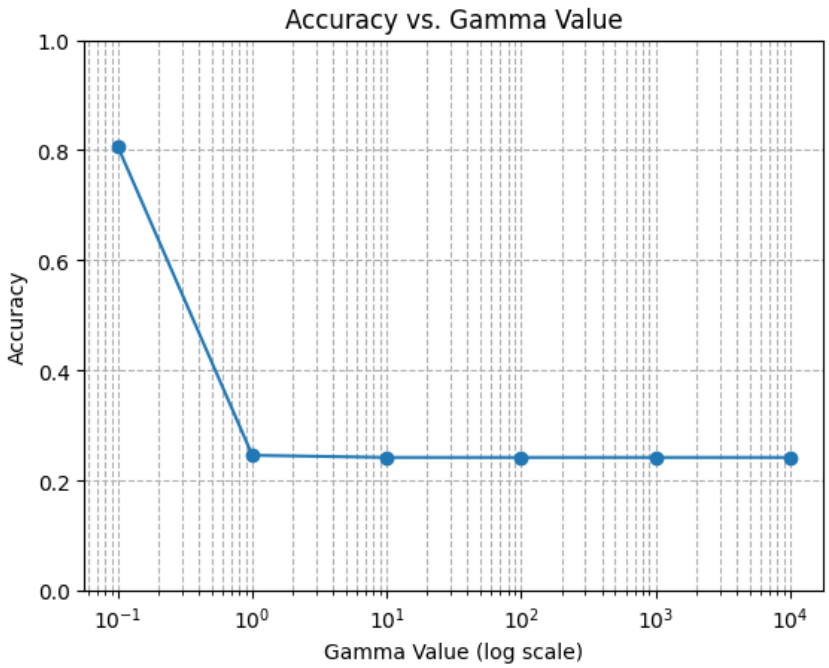
We can see that higher C values increase the the accuracy of a linear kernel SVC. However, we can see that the accuracy stops increasing rapidly for values of C greater than 0.1. The accuracy reaches a plateau a little above 0.8.

1.2 Question 1 b: RBF kernel SVM

Now we will train a model called RBF kernel SVM. This model has a parameter gamma, which we will vary in the range of

$$\gamma = (10^{-1}, 10^{0}, ..., 10^{4}) \tag{2}$$

and see how the accuracy of the prediction changes with each value of gamma.



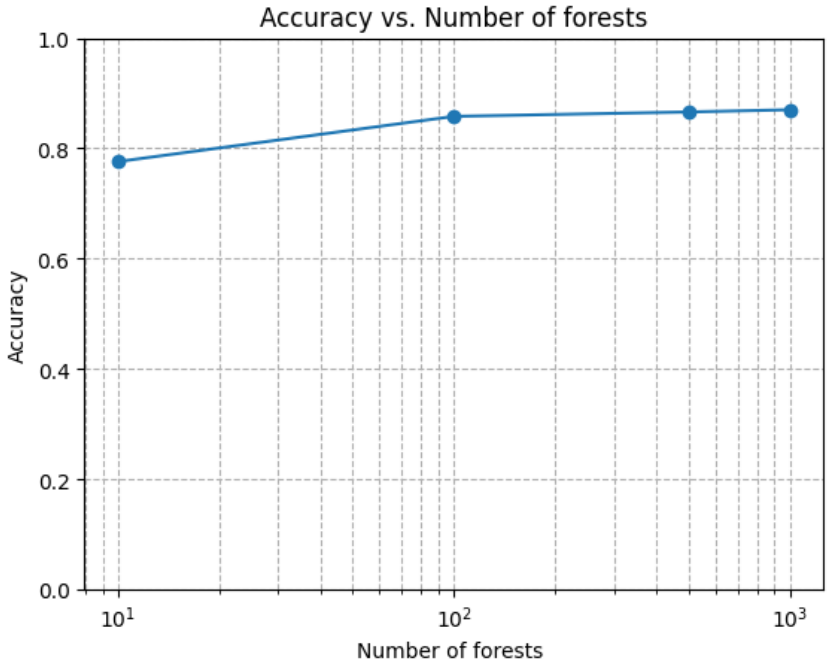
We can observe that the accuracy decreases with larger values of gamma. The accuracy goes from about 0.8 to about 0.22 and plateaus at 0.22 for all gamma values equal 1 or greater.

1.3 Question 1 c: random forest classifier

Now we will train a random forest classifier while changing the number of trees in the range of

$$(10, 100, 500, 1000) \tag{3}$$

and see how the accuracy of the prediction changes with the number of trees



We can notice that as the number of forests increases the accuracy increases, however the accuracy does not increase much after the number of forests is 100 and above.

1.4 Question 1 d: 5-fold cross validation for selecting hyper-parameters

Now we implement 5-fold cross-validation for selecting the optimal hyper-parameters for the lsvc, rsvc and random forest. We will use the grid search method. Here is the algorithm I used written in pseudocode.

```
Import the relevant libraries: GridSearchCV, and KFold

Configure the kfold cross-validation procedure

Define search space for parts lsvc, rsvc and random forest
    using the same range of parameters from part a, b and c

Define the lsvc, rsvc and rf models

Perform a grid search

Get the best performing model fit on the whole training set

Evaluate model on the test set

Print the results for the optimal hyperparameters and accuracy
    on test data
```

In the following table we summarize our results:

Parameter	**Value**
LinearSVC	
Optimal C	100
Accuracy on test data	0.846
RSVC	
Optimal gamma	0.1
Accuracy on test data	0.806
RandomForest	
Optimal number of trees	500
Accuracy on test data	0.866

Table 1: Optimal Parameters and Accuracies for Different Models

2 Short Answer Questions

2.1 Question 2a

Answers:

A and C

2.2 Question 2b

Answers:

C