CENG 499

Introduction to Machine Learning

Fall 2022-2023

Homework 3

Mustafa Aydogdu

2380111

1 Part 2 - Data Preprocessing, SVM, Kernel Functions

1.1 Data-set 1

For this part, 4 different configurations of parameters have been tested with Support Vector Classifier.

Configurations:

- 1^{st} Configuration → Kernel: Linear, C:1
- 2^{nd} Configuration → Kernel: Linear, C:5
- 3^{rd} Configuration → Kernel: RBF, C:1
- 4^{th} Configuration → Kernel: RBF, C:5

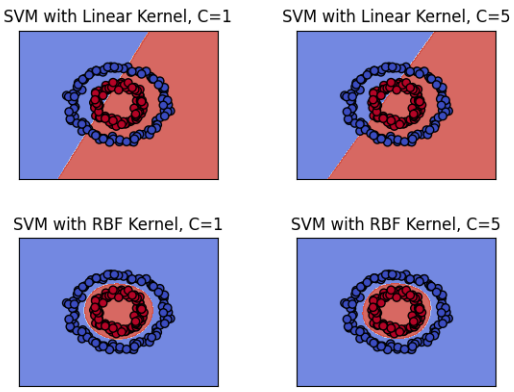

Figure 1: Plots with decision boundaries of the SVM model

1.2 Data-set 2

In this part, 12 different combinations of kernel and c values have been tested with 95% confidence interval.

```
Configuration: {'C': 5, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.010
Configuration: {'C': 5, 'kernel': 'rbf'} | mean accuracy: 0.94 | Confidence interval: 0.94 +/- 0.016
Configuration: {'C': 10, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.006
Configuration: {'C': 10, 'kernel': 'rbf'} | mean accuracy: 0.94 | Confidence interval: 0.94 +/- 0.012
Configuration: {'C': 50, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.006
Configuration: {'C': 50, 'kernel': 'rbf'} | mean accuracy: 0.93 | Confidence interval: 0.93 +/- 0.015
Configuration: {'C': 100, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.010
Configuration: {'C': 100, 'kernel': 'rbf'} | mean accuracy: 0.92 | Confidence interval: 0.92 +/- 0.026
Configuration: {'C': 500, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.006
Configuration: {'C': 500, 'kernel': 'rbf'} | mean accuracy: 0.91 | Confidence interval: 0.91 +/- 0.037
Configuration: {'C': 1000, 'kernel': 'linear'} | mean accuracy: 0.95 | Confidence interval: 0.95 +/- 0.006
Configuration: {'C': 1000, 'kernel': 'rbf'} | mean accuracy: 0.91 | Confidence interval: 0.91 +/- 0.036
```

Figure 2: Confidence intervals for each hyperparameter

2 Part 3 - Method Comparison

In this part, 2 different configurations of each method have been tested, which are K-Nearest Neighbor, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier.

KNN configurations:

- 1^{st} Configuration → Weight: Uniform, N: 9
- 2^{nd} Configuration → Weight: Uniform, N: 29

SVC configurations:

- 1^{st} Configuration → Kernel: Linear, C: 5
- 2^{nd} Configuration → Kernel: Linear, C: 500

DT configurations:

- 1^{st} Configuration → Criterion: Entropy, Max depth: 2
- 2^{nd} Configuration → Criterion: Entropy, Max depth: 5

RF configurations:

- 1^{st} Configuration → Number of trees in the forest: 100, Criterion: Gini
- 2^{nd} Configuration → Number of trees in the forest: 100, Criterion: Entropy

```
-----KNN-----
Configuration 1: N=9, weights = uniform
Mean Accuracy: 0.721, Confidence Interval: , 0.721 +/- 0.005

Configuration 2: N=29, weights = uniform
Mean Accuracy: 0.721, Confidence Interval: , 0.721 +/- 0.004

Overall KNN Accuracy: 0.722, Accuracy Confidence Interval: 0.722 +/- 0.008, F1 Score: 0.823, F1 Score Confidence Interval: 0.823 +/- 0.006

-----SVM-----
Configuration 1: kernel = linear, C = 5
Mean Accuracy: 0.743, Confidence Interval: , 0.743 +/- 0.004

Configuration 2: kernel = linear, C = 500
Mean Accuracy: 0.743, Confidence Interval: , 0.743 +/- 0.004

Overall SVM Accuracy: 0.747, Accuracy Confidence Interval: 0.747 +/- 0.008, F1 Score: 0.826, F1 Score Confidence Interval: 0.826 +/- 0.005

-----Decision Tree-----
Configuration 1: criterion = entropy, max_depth = 2
Mean Accuracy: 0.700, Confidence Interval: , 0.700 +/- 0.005

Configuration 2: criterion = entropy, max_depth = 5
Mean Accuracy: 0.692, Confidence Interval: , 0.692 +/- 0.006

Overall Decision Tree Accuracy: 0.696, Accuracy Confidence Interval: 0.696 +/- 0.009, F1 Score: 0.797, F1 Score Confidence Interval: 0.797 +/- 0.013

-----Random Forest-----
Configuration 1: n_estimators = 100, criterion = gini
Mean Accuracy: 0.754, Confidence Interval: , 0.754 +/- 0.002

Configuration 2: n_estimators = 100, criterion = entropy
Mean Accuracy: 0.755, Confidence Interval: , 0.755 +/- 0.002

Overall Random Forest Accuracy: 0.756, Accuracy Confidence Interval: 0.756 +/- 0.003, F1 Score: 0.840, F1 Score Confidence Interval: 0.840 +/- 0.003
```

Figure 3: Results with confidence intervals of 95%

As seen in the results, best performing method is, with highest overall accuracy and highest F1 score, the Random Forest Classifier.