Derya Tınmaz 2380947

CENG 499 - Introduction to Machine Learning

Homework 3 Report

Part 2 - dataset1

For this part we were expected to run SVM model of Scikitlearn with different configurations. I run the model with 6 different configurations and provided their corresponding resulting plots:

1- radial basis function kernel with C=1.0

2- radial basis function kernel with C=10.0

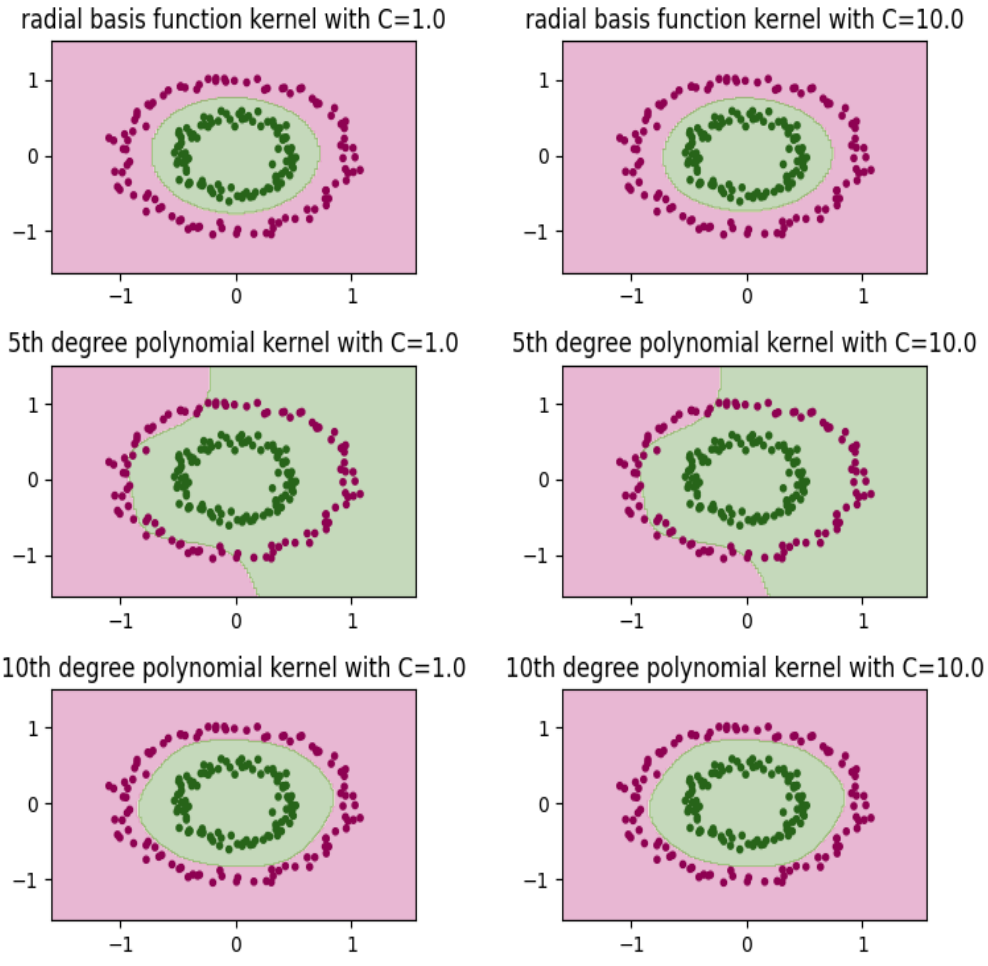
3- 5th degree polynomial kernel with C=1.0

4- 5th degree polynomial kernel with C=10.0

5- 10th degree polynomial kernel with C=1.0

6- 10th degree polynomial kernel with C=10.0

Among the configurations 3 and 4 are not very good at classification for this dataset. 1 and 2 are better than 5 and 6 because the boundary between classes is centered. 1 and 2 look similar, they both are good at classification for this dataset.



Derya Tınmaz 2380947

Part 2 - dataset2

For this part we were expected to perform cross-validation to find the best hyper parameter values for this dataset. I considered accuracy performance metric while determining the best hyper parameter values on dataset2. The table below shows the results for the hyper parameter I used in grid search.

	kernel functions	C value	accuracy score	confidence interval
1	polynomial (with degree 5)	10	87.412	(87.188, 87.636)
2	radial basis	10	94.249	(94.093, 94.406)
3	sigmoid	10	78.773	(78.114, 79.431)
4	polynomial (with degree 5)	100	91.266	(91.031, 91.500)
5	radial basis	100	91.923	(91.404, 92.442)
6	sigmoid	100	77.881	(77.492, 78.271)

According my results from table, configuration 2: radial basis function kernel with C=10 gives the highest accuracy score for this dataset.

Part 3

Tables of hyper parameter search results for each algorithm and their best configuration is shown:

KNN

	metric	n neigbors	mean	confidence interval
1	cosine	11	71.773	(71.120, 72.426)
2	cosine	51	71.314	(70.905, 71.722)
3	euclidean	11	71.825	(71.189, 72.462)
4	euclidean	51	71.341	(70.946, 71.737)

SVM

	C	kernel function	mean	confidence interval
1	10	polynomial	70.843	(69.934, 71.752)
2	10	radial basis	72.701	(71.841, 73.562)
3	100	polynomial	70.421	(69.548, 71.294)
4	100	radial basis	71.815	(71.045, 72.586)

Derya Tınmaz 2380947

Decision Tree

	criterion	max depth	mean	confidence interval
1	gini	30	67.302	(66.642, 67.962)
2	gini	50	67.326	(66.619, 68.032)
3	entropy	30	67.431	(66.863, 67.998)
4	entropy	50	67.587	(67.030, 68.143)

Random Forest

	n estimators	max depth	mean	confidence interval
1	100	30	75.328	(75.017, 75.639)
2	100	50	75.430	(75.133, 75.727)
3	200	30	75.263	(74.967, 75.560)
4	200	50	75.408	(75.101, 75.716)

Table of algorithm grid search:

	f1 scores	confidence interval	accuracy scores	confidence interval
knn	82.8	(82.3, 83.2)	71.8	(70.9, 72.7)
svm	81.8	(80.9, 82.7)	73.0	(71.9, 74.2)
decision tree	76.2	(75.1, 77.3)	67.0	(65.8, 68.3)
random forest	84.1	(83.6, 84.5)	75.6	(75.0, 76.3)

According to my grid search results, random forest algorithm is the best algorithm to classify the given dataset. 100 estimators and max depth 50 is the best hyper parameter configuration for this algorithm for the given dataset.