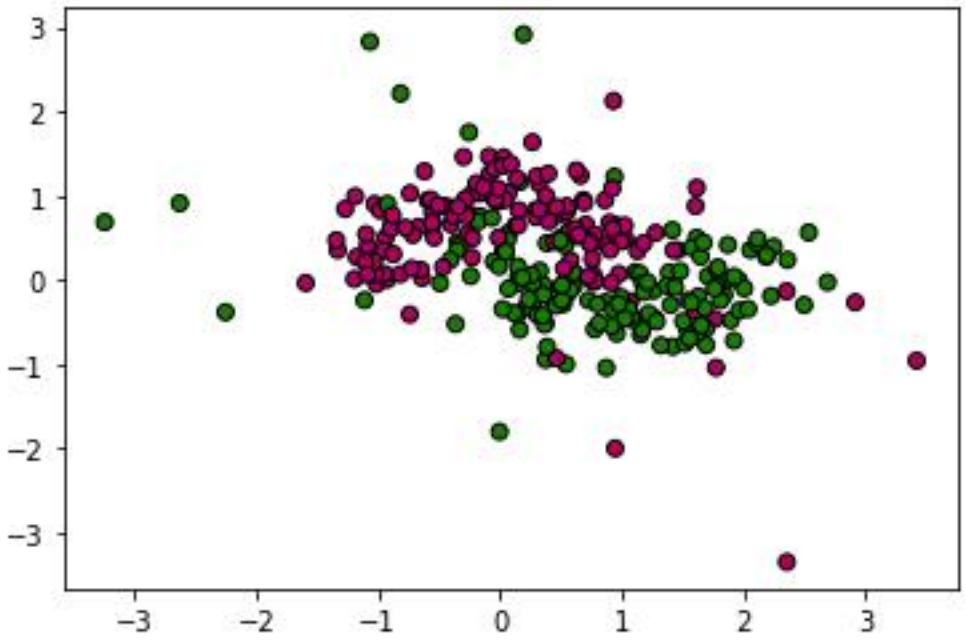
CS6375 Assignment 4 Report

Haoda le
HXL180046

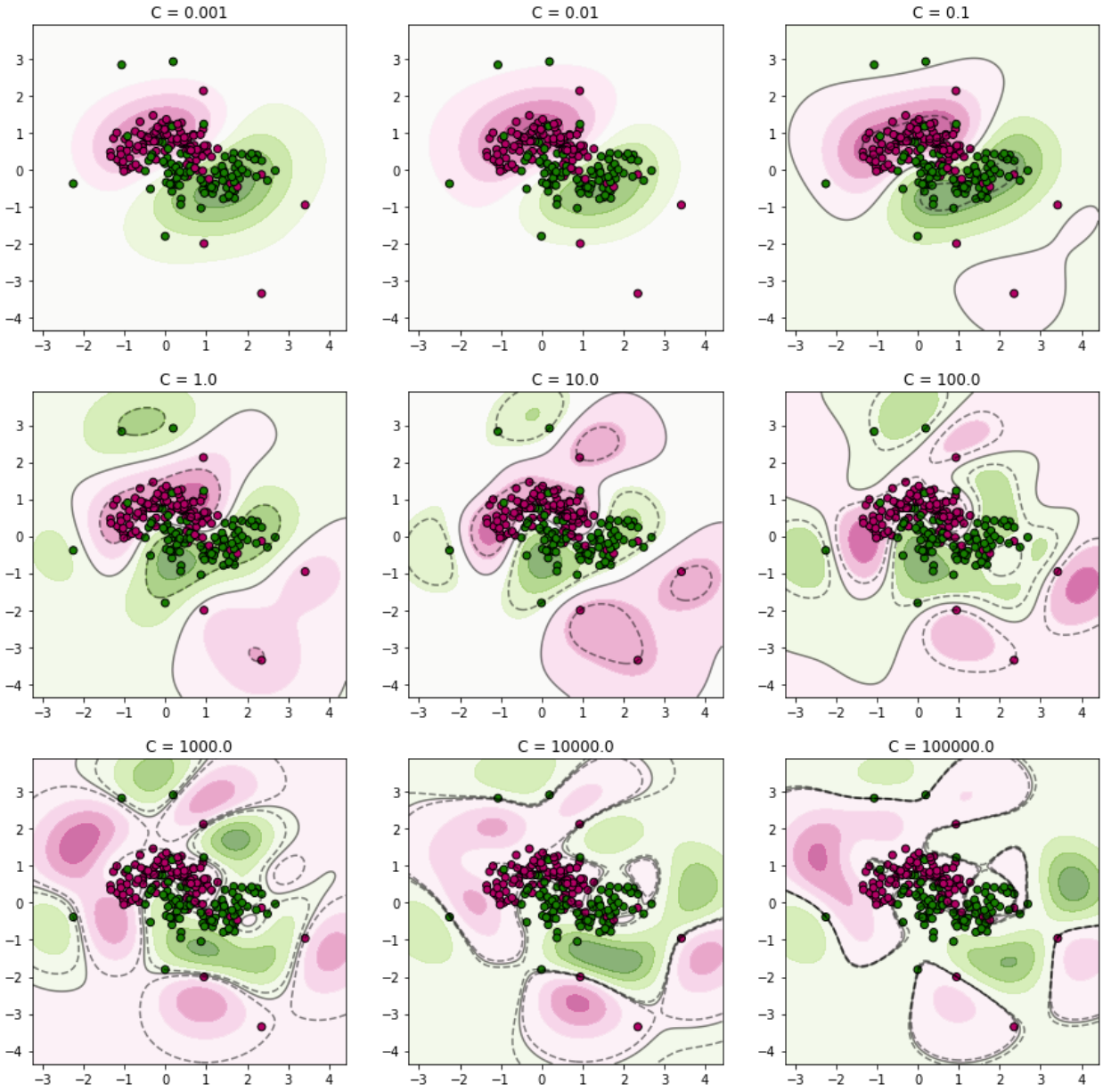
1. **Support Vector Machines with Synthetic Data**

Scatter plot of randomly generated data shows below:

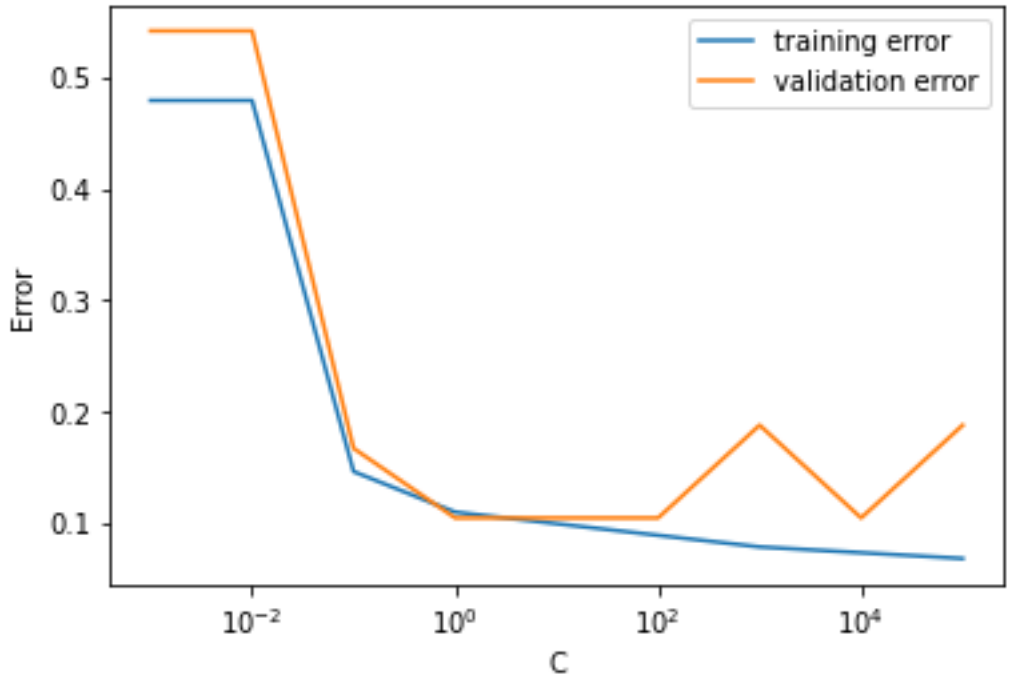


1.a

Visualization of the models shows below:



The training error and the validation error shows below:



Discussion:
When C increases, the training error decreases all the way down, and validation error decrease at first, then when C > 1, the validation error increase a little bit and start to vibrate. The mode becomes more accurate and the partition becomes more complicated when C increases.
In terms of minimizing the SVM's object function, when C increases, the hinge loss for each training example should be smaller. So that the model becomes more precise.

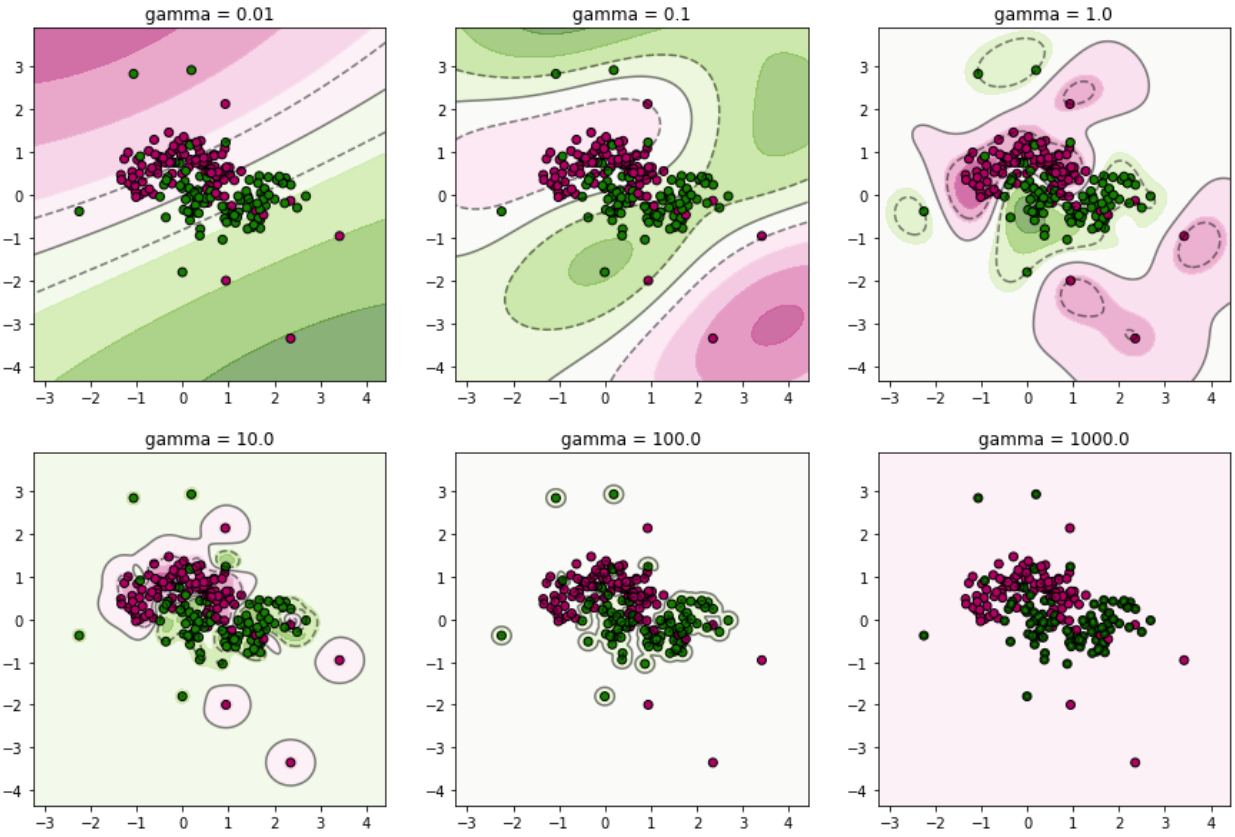
Final Model Selection:
We print all the validation error, and according to the result we choose C_best = 1.0, so the test set accuracy is 0.833.

validation error for different C:
{0.001: 0.5416666666666667, 0.01: 0.5416666666666667, 0.1: 0.16666666666666663, 1.0: 0.10416666666666663, 10.0: 0.10416666666666663, 100.0: 0.10416666666666663, 1000.0: 0.1875, 10000.0: 0.10416666666666663, 100000.0: 0.1875}

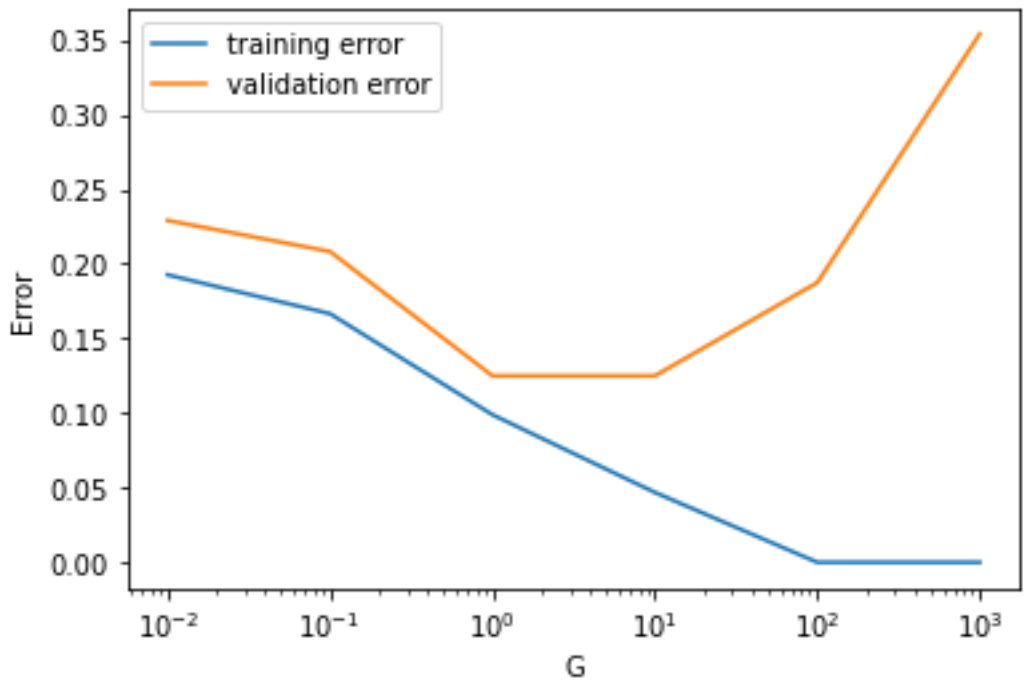
C_best: 1.0
accuracy: 0.8333333333333334

1.b

Visualization of the models shows below:



The training error and the validation error shows below:



Discussion:
When gamma increases, the training error decreases all the way down, and validation error decrease at first, then when gamma > 1, the validation error no longer decreases and then start to increase. When gamma increases, the model becomes less accurate and the partition become simpler. In terms of the functional form of the RBF kernel, when gamma increases, the squared Euclidean distance has more effect on k(x, z), so that the model becomes less precise.

Final Model Selection:
We print all the validation error, and according to the result we choose gamma_best = 1.0, so the test set accuracy is 0.833.

validation error for different gamma:
{0.01: 0.22916666666666663, 0.1: 0.20833333333333337, 1.0: 0.125, 10.0: 0.125, 100.0: 0.1875, 1000.0: 0.35416666666666663}

Gamma_best: 1.0
accuracy: 0.8333333333333334

2. **Breast Cancer Diagnosis with Support Vector Machines**

The table below shows the training error and validation error. For each table, the row stands for C, the column stands for gamma, so we have the different combination.

trnErr

	0.001	0.01	0.1	1.0	10.0	100.0
0.01	0.372	0.372	0.372	0.372	0.372	0.372
0.1	0.307	0.05	0.035	0.372	0.372	0.372
1.0	0.047	0.029	0.012	0.0	0.0	0.0
10.0	0.027	0.012	0.0	0.0	0.0	0.0
100.0	0.015	0.003	0.0	0.0	0.0	0.0
1000.0	0.006	0.0	0.0	0.0	0.0	0.0
10000.0	0.0	0.0	0.0	0.0	0.0	0.0

valErr

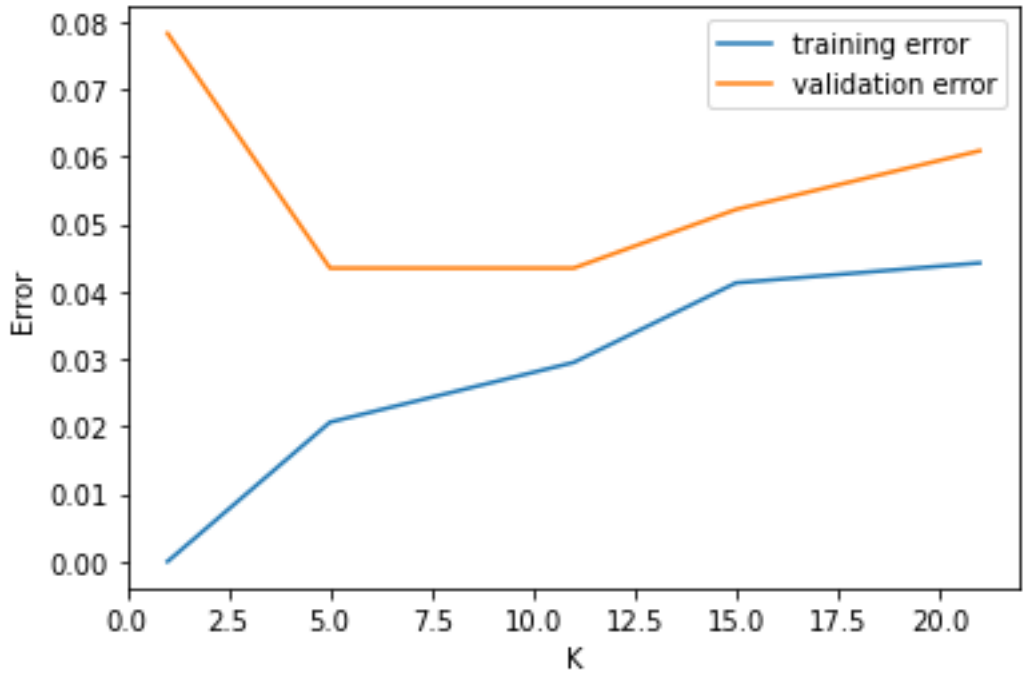
	0.001	0.01	0.1	1.0	10.0	100.0
0.01	0.374	0.374	0.374	0.374	0.374	0.374
0.1	0.304	0.07	0.078	0.374	0.374	0.374
1.0	0.061	0.061	0.043	0.374	0.374	0.374
10.0	0.035	0.043	0.035	0.374	0.374	0.374
100.0	0.035	0.026	0.035	0.374	0.374	0.374
1000.0	0.035	0.026	0.035	0.374	0.374	0.374
10000.0	0.026	0.026	0.035	0.374	0.374	0.374

Final Model Selection:
According to the validation error table above, we can choose C_best = 10000, and gamma_best = 0.001, so the test set accuracy is 0.939.

C_best: 1000.0
Gamma_best: 0.001
accuracy: 0.9391304347826087

3. **Breast Cancer Diagnosis with -Nearest Neighbors**

The training error and the validation error shows below:



Final Model Selection:
We print all the validation error, and according to the result, we choose K_best = 5, so the test set accuracy is 0.957.

validation error for K:
{1: 0.07826086956521738, 5: 0.04347826086956519, 11: 0.04347826086956519, 15: 0.05217391304347829, 21: 0.060869565217391286}

K_best: 5
accuracy: 0.9565217391304348

Discussion:
I would prefer KNN for this classification task, because the accuracy for KNN is better than SVMs.