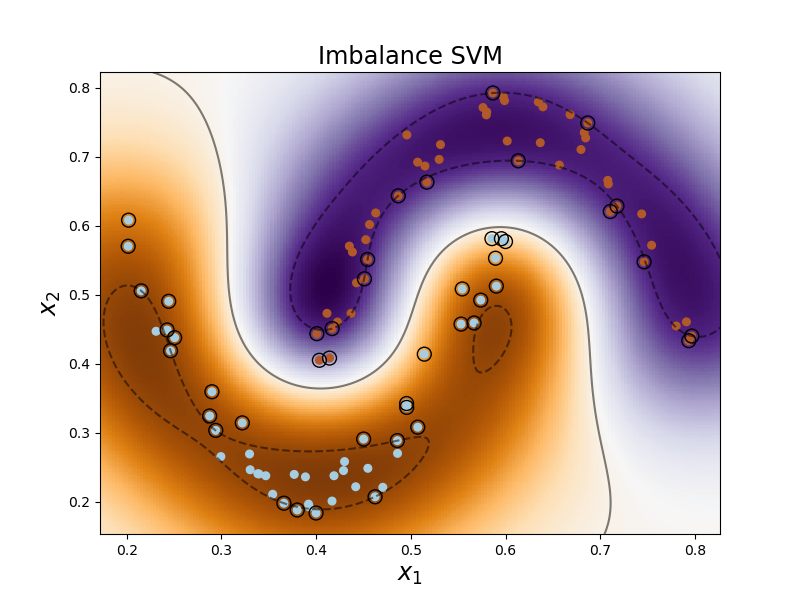
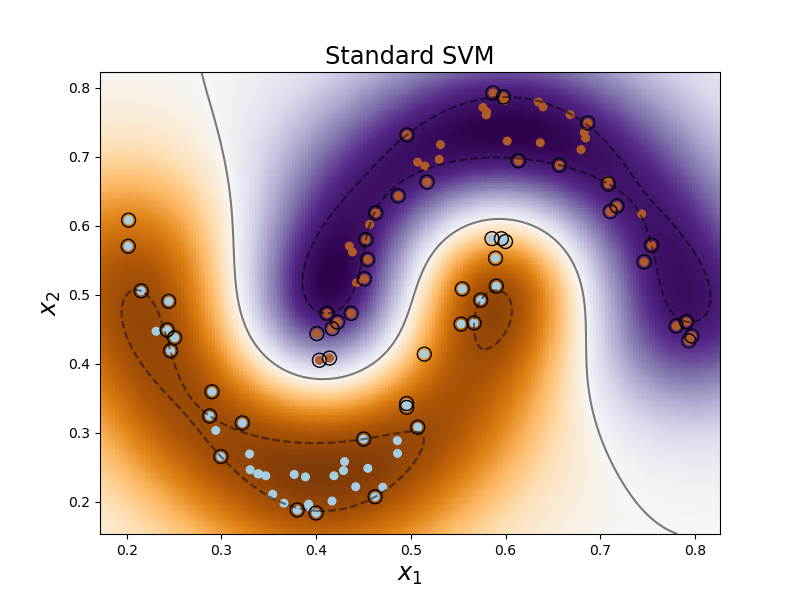
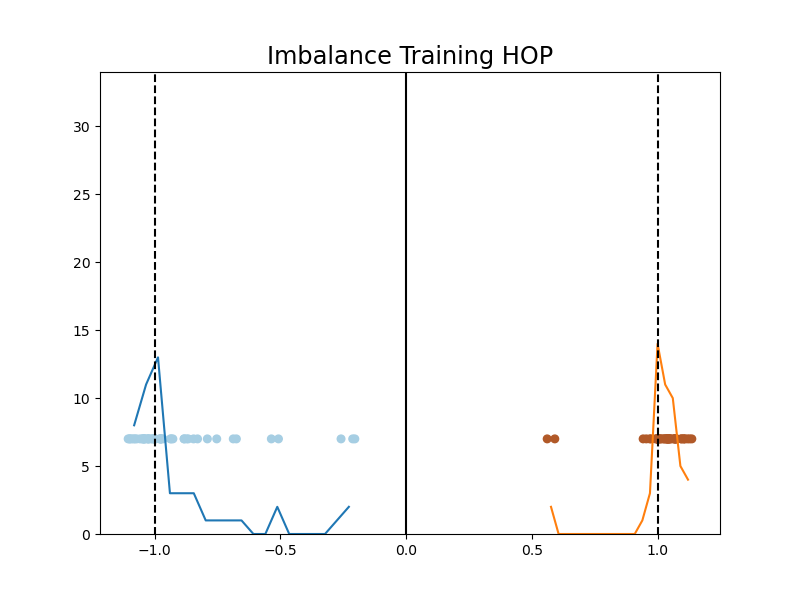
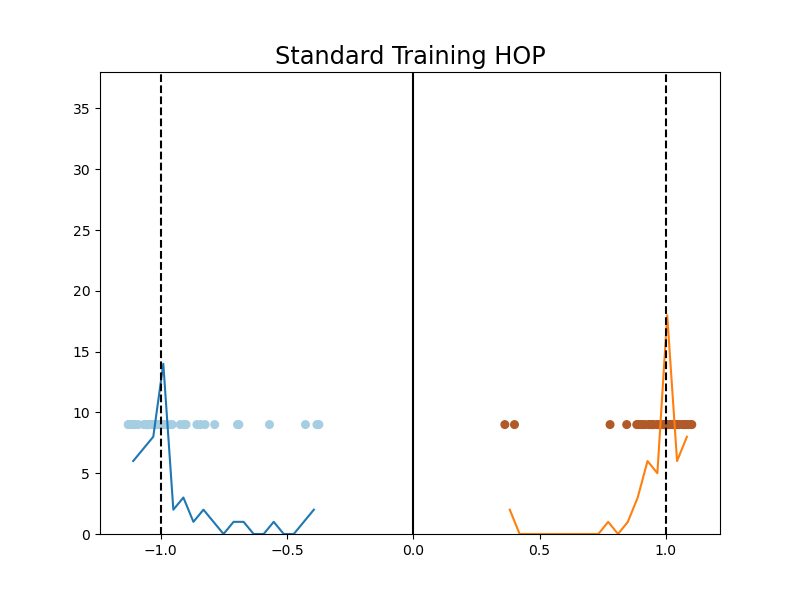
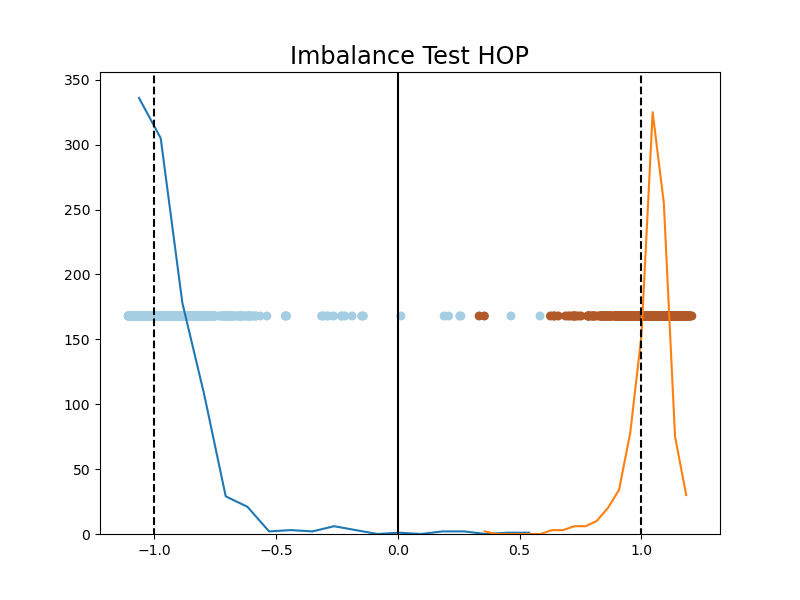
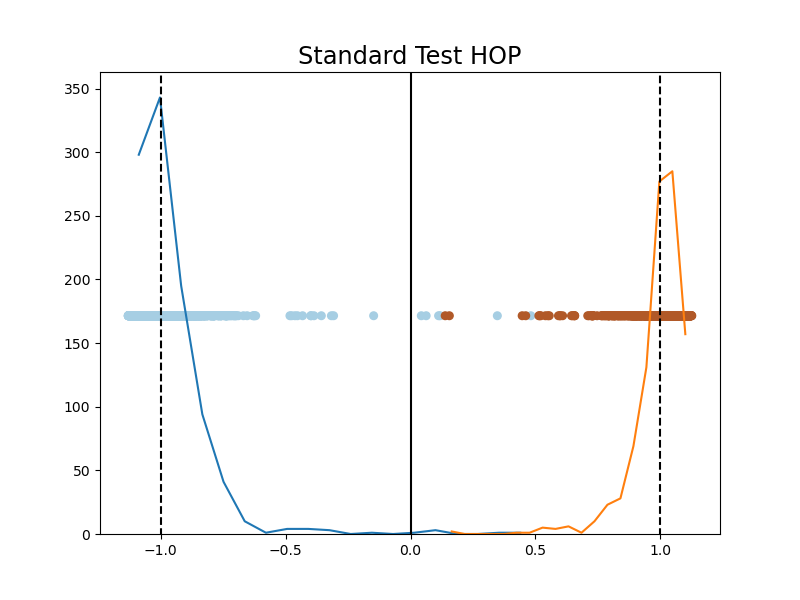
Homework 3

Problem 2







For the hyperbola dataset (see lecture set 9.2 on how to generate the dataset), the number of training, validation, and test sample is set to 100, 100, and 2000, respectively. For model selection, the range of SVM parameter C is set as [], whereas the range of kernel parameter is set as [48, 56, 64, …, 88, 96]. For the imbalance SVM case, the weight ratio between +1 and -1 is set to 2:1. Figures above show the decision boundary and histogram of projection (for both training and test dataset) of standard and imbalance RBF-SVM. Because the hyperbola dataset is separable (in the higher-dimensional space), both standard and imbalance decision boundary looks very similar to each other, although some minor differences can still be observed (e.g. decision boundary is slightly closer to the -1 class samples (blue) at the center for the imbalance case). The training, test, optimal C, and optimal gamma for standard and imbalance SVM are summarized in the table below:

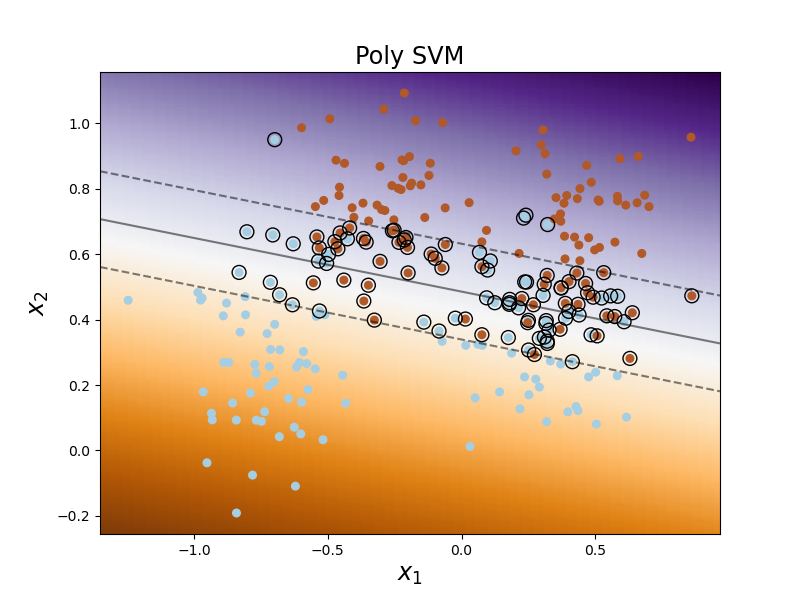
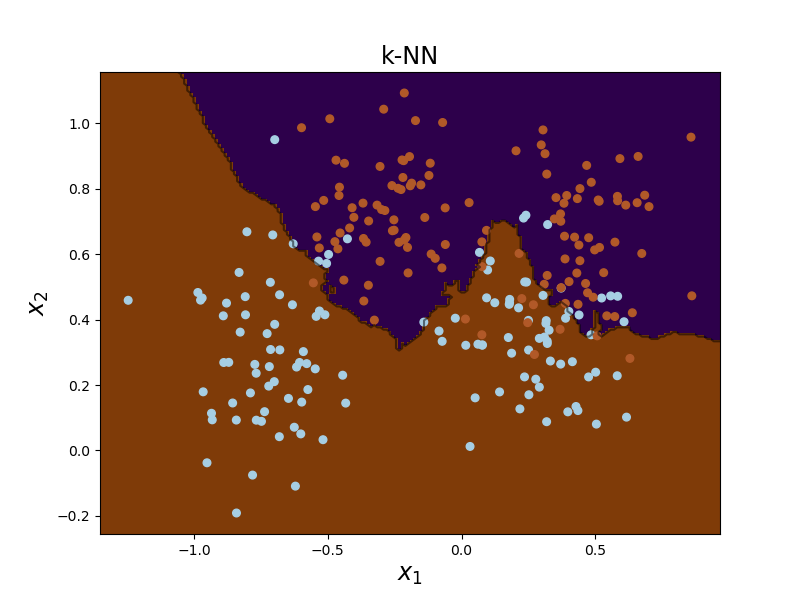
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Optimal C** | **Optimal** | **Training Error** | **Test Error** |
| Standard SVM | 0.25 | 56 | 0 | 0.003 |
| Imbalance SVM | 0.25 | 48 | 0 | 0.0035 |

Problem 3

The result for the Ripley’s dataset and the high-dimensional dataset using k-NN and polynomial SVM are shown in the table below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | k-NN | | | Polynomial SVM | | | |
|  | Optimal K | Train Error | Test Error | Optimal C | Optimal  Degree | Train Error | Test  Error |
| Ripley’s Dataset | 13 | 0.12 | **0.097** | 2 | 1 | 0.15 | 0.109 |
| High-Dimensional Dataset | 9 | 0.16 | 0.27 | 0.25 | 3 | 0 | **0.15** |

For Ripley’s dataset, the number of training, validation, and test sample is set to 250, 250, and 750, respectively. During model selection, the range of k is set as [3, 5, 7, …, 15] for the k-NN classifier, whereas the range of C and polynomial degree is set as [ and [1,2,3,4] respectively, for the polynomial SVM. The same range of parameters for k-NN and polynomial SVM is also used for the high-dimensional dataset. For the Ripley’s dataset, the k-NN achieves lower training and test error than the polynomial SVM. This result indicates the k-NN has better generalization than the polynomial SVM, for the Ripley’s dataset. On the other hand, for the high-dimensional dataset, the polynomial SVM achieve much lower training and test error than the k-NN. This is because the high-dimensional dataset contains many irrelevant features (such as ), and k-NN is ill-suited to deal with dataset with irrelevant features (all features contributes equally). In contrast, polynomial SVM is much more effective against irrelevant features.



Decision Boundary of optimal kNN and polynomial SVM for Ripley’s dataset.

Problem 4

Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Optimal C | Optimal | Training Error | Test Error |
| Small Training Size  (Averaged over 100 realization) | - | - | 0 | 0.1680.04 |
| Large Training Size | 0.25 | 8 | 0 | 0.042 |
| Large Training Size (5% output noise) | 0.25 | 8 | 0 | 0.114 |

Model selection

All optimal modeling results are chosen based on given range of parameter C = [ and ], and the optimal parameter are shown in the table above.