CS7641 Fall 2018 Assignment1 Supervised learning

Xiaoxi Wang (xwang738)

email: xwang738@gatech.edu

Summary

In this report, different supervised learning algorithms (Decision tree, k-Nearest Neighbors, Support Vector Machine, Boosting, Neural Networks) were applied to two data sets from the UCI machine learning repository. Programming was done using Python with the scikit-learn library. For each algorithm, hyperparameter tuning was performed over one parameter using cross validation curves or two parameters with cross validation heatmaps. The selected models were then evaluated with learning curve over sample size [1]. Finally, different algorithms were compared in a quantitative manner based on prediction scores (accuracy, AUC and F), training time and prediction time.

Datasets

Two data sets were selected from the UC Irvine Machine Learning Depository [2]. Both datasets were binary classification problems (suitable for the scope of this assignment).

The HTRU2 dataset

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South) [3]. Pulsars are a rare type of Neutron star that produce radio emission detectable on earth. The data set describes a problem of labeling pulsar candidates based on 8 different features of the detected emission pattern. The data set is interesting as it contains a large number of instances, moreover, it is biased towards the negative examples (90.8%), which helps practice analyzing biased datasets.

The Breast Cancer Wisconsin data set

This data set represents a problem of diagnosing whether a breast lump is benign or malignant [4,5]. Digitized images were collected with a fine needle aspirate (FNA) of the breast mass of interest, and features were extracted from these images describing characteristics of the cell nuclei. This data set contains a non-trivial size of instances (n=699) and little noise. Among the samples, the two categories were almost evenly distributed with 34.5% “malignant” and 65.5% “benign”.

	Task	#Instances	#Attributes	Attribute type	Positive v.s. Negative examples
HTRU2	Binary Classification	17898	8	Numeric	9.2% v.s. 90.8%
Breast Cancer Wisconsin	Binary Classification	683*	9**	Numeric	34.5% v.s. 65.5%

Table 1. Comparison of the two data sets used in this report

*Removed 16 instances with missing data.

**ID number attribute was dropped.

Data preprocessing: Both data sets were spitted into training (70%) and testing (30%), and 5 folds cross validation was performed using the training data. The test data were "untouched" till the final evaluations and comparisons of different algorithms. For the HTRU2 dataset, several feature values were not in comparable ranges, so the data were normalized using standard scaler, which removes the mean and scales to unit variance [6]. Moreover, because the HTRU2 data set is severely biased towards negative examples, F1 score metrics (2*Precision*Recall/(Precision + Recall)) was used to tune and evaluate each model.

Results and Discussion

1. **Decision Tree**

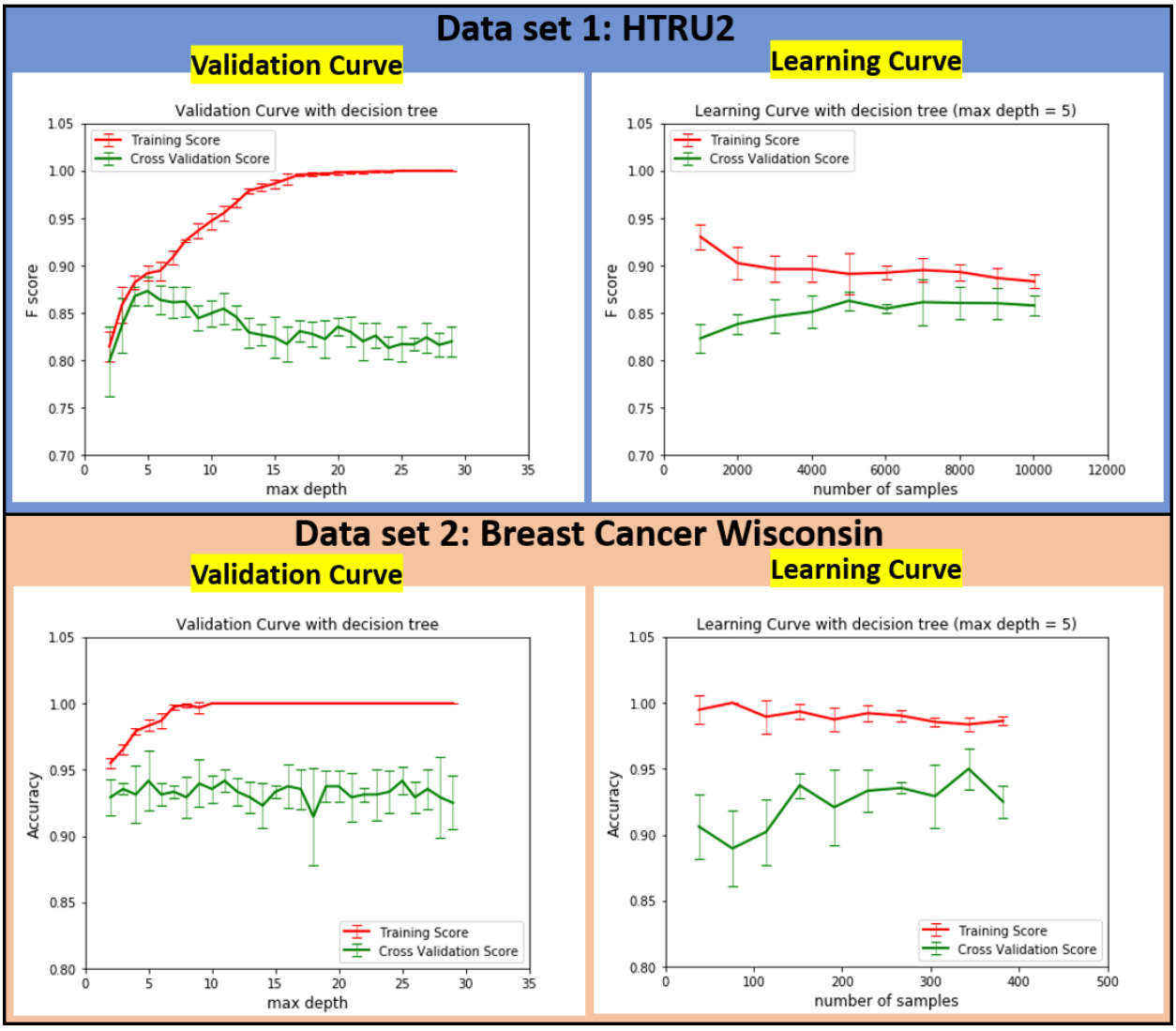

Figure 1. Decision Tree: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

Pruning was performed by restricting the maximal depth of the tree. For both datasets, increasing tree maximal depth allows the learning algorithm to fit training data perfectly, however, the cross validation

curve drops after reaching a maximal depth of 5 for HTRU2, indicating overfitting. Using maximal depth of 5, The learning curve converges for HTRU2, suggesting that the model was good with no high bias or high variance issues.

For the breast cancer data, very simple decision tree (maximal depth <3) was able to achieve high cross validation accuracy (>90%), and increasing tree complexity didn't help much. As revealed by the learning curve, the relatively large gap between training and cross validation indicate high variance, and more data would help considering that there are less than 500 training samples.

2. ***k*-Nearest Neighbors**

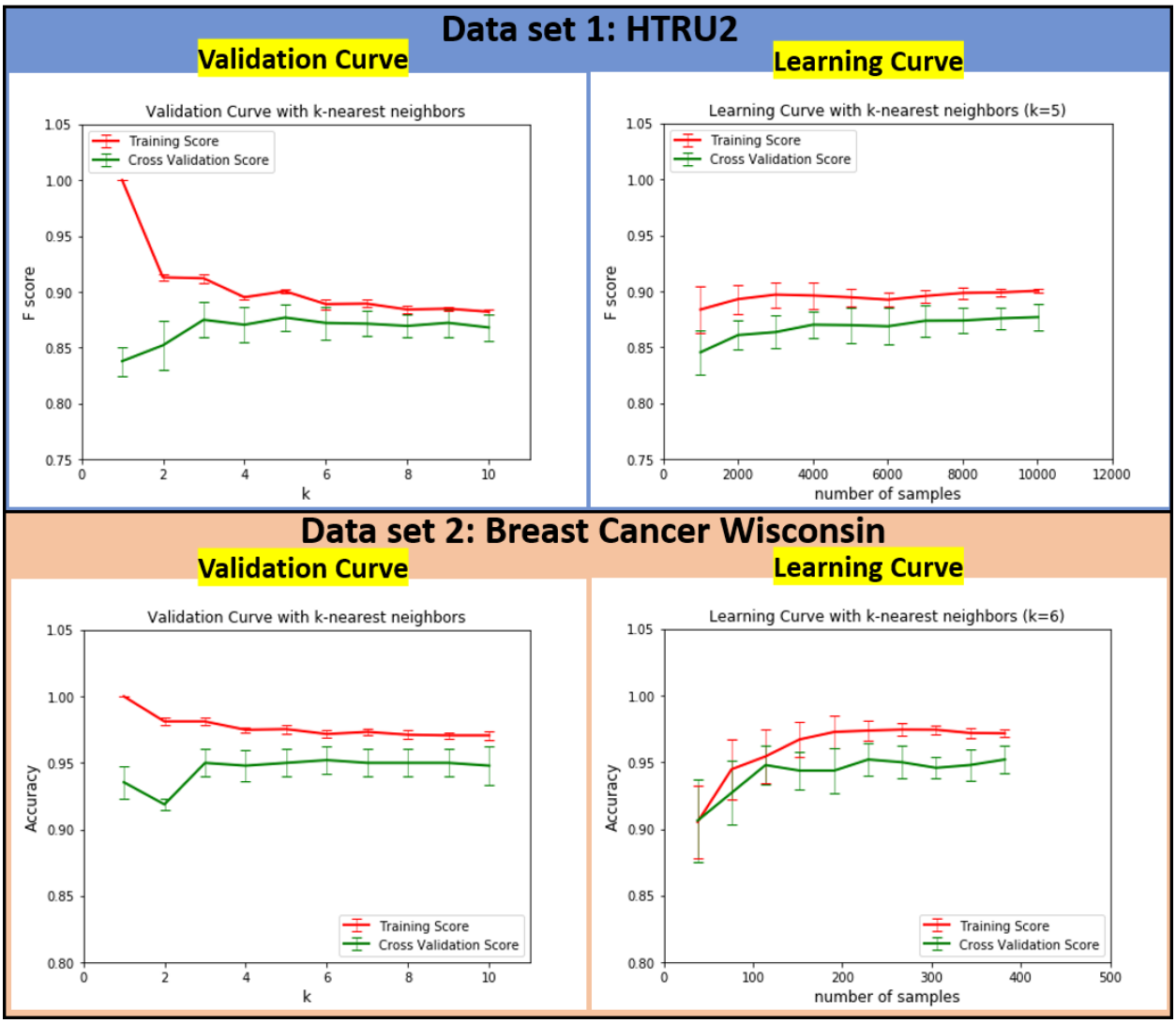

Figure 2. k-Nearest Neighbors: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

For both datasets, training scores reache 100% at k=1 and decrease as the value of k grows, in contrast, the cross validation scores first increase and then plateau after reaching a k value of about 3. As revealed by the learning curves, both training and cross validation scores increase over larger sample

size. For the breast cancer data set, learning and cross validation may converge better as sample size continue to increase, considering there are only 500 instances in this set.

3. Support Vector Machine

3.1 linear kernel

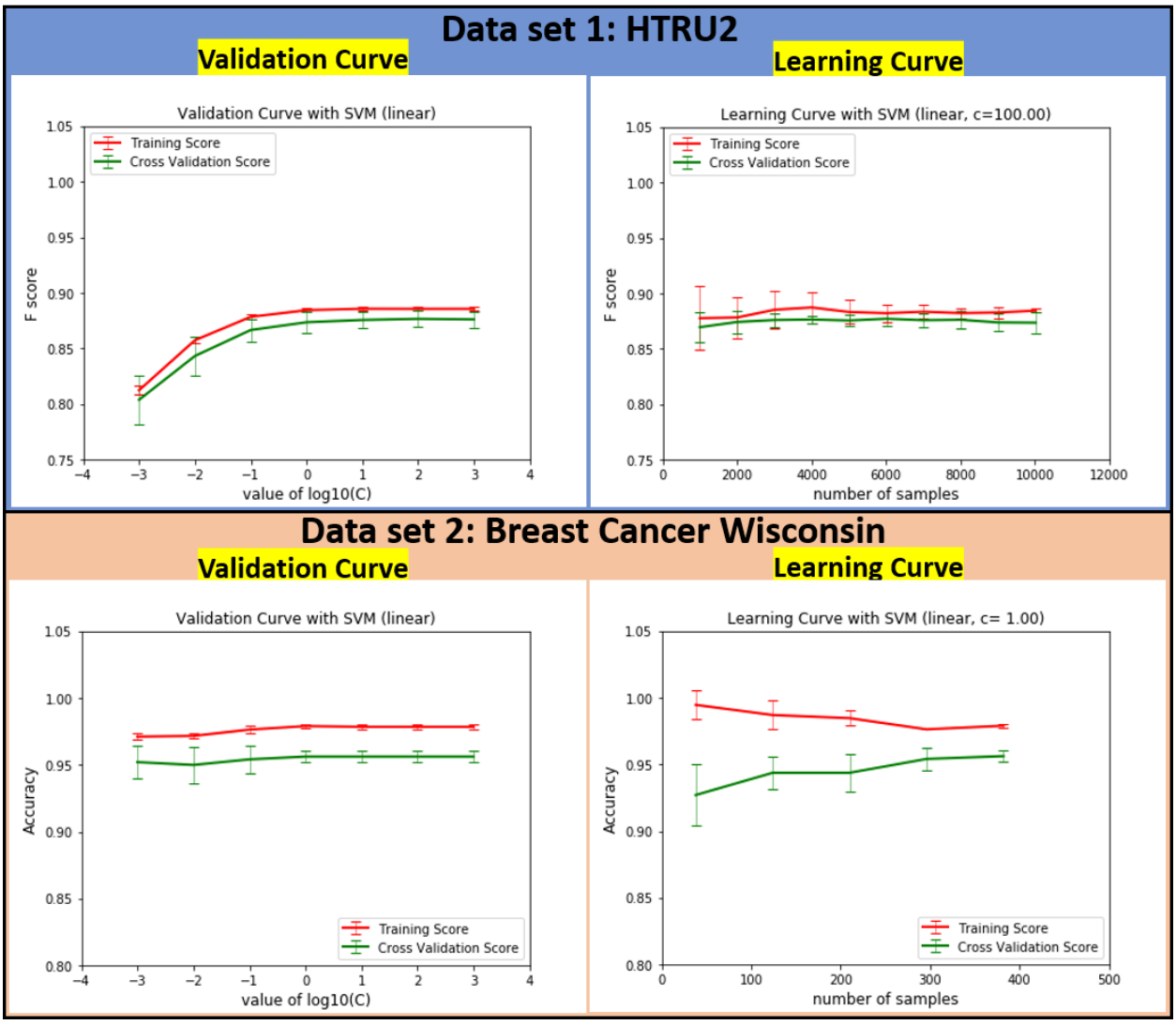

Figure 3.1. SVM with linear kernel: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

The linear kernel SVM was tuned over the value of C. Larger C means high penalty for misclassified samples and may results in overly complex model with a small margin. For the HTRU2 dataset, both training and cross validation scores increase and plateau when C reaches 0.1. Training and cross validation curves are close and do not change significantly over sample size, indicating high bias. The above results suggest non-linear separability of the data.

Liner kernel SVM performs well with the breast cancer data, as both training and cross validation scores reaches 0.95 even with very small values of C, suggesting the data were linear separable, and there is

little noise in the data. As sample size increases, the gap between training and cross validation curves gradually drops, suggesting that increasing sample size lowers variance.

3.2 RBF kernel

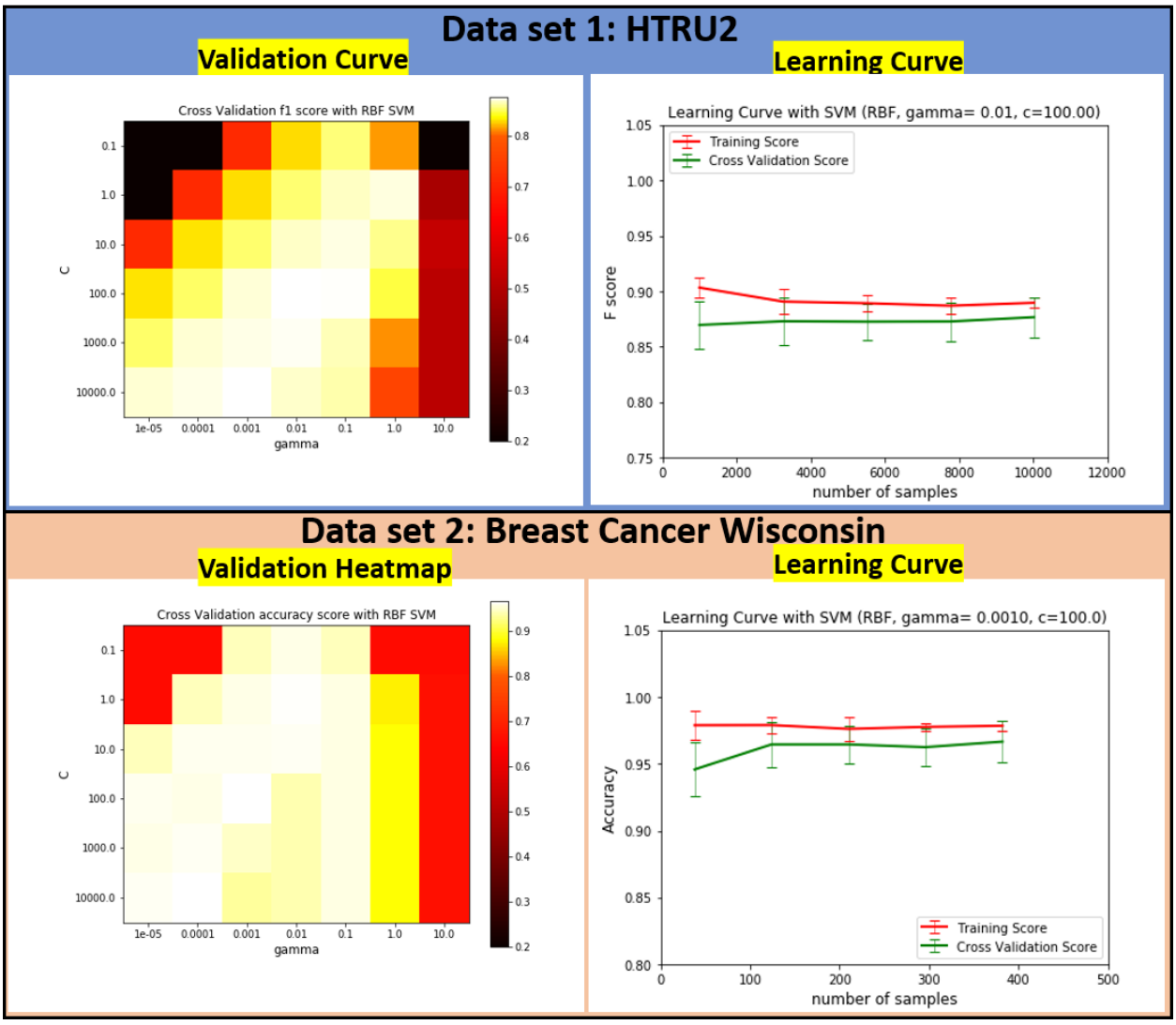

Figure 3.2. SVM with RBF kernel: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

For SVM with Radial basis function (RBF) kernel, hyperparameter tuning was performed over two factors, C and gamma. Larger gamma represents higher standard deviation of the kernel function. RBF performs well for both data sets even with small sample size (even if only use 10% data for HTRU2 and 20% data for breast cancer data set), with high accuracy and little gap between training and cross validation scores. The results suggest that RBF kernel perform well for both data sets, and it is not suffering from high bias or overfitting (high variance) issues.

4. Boosting

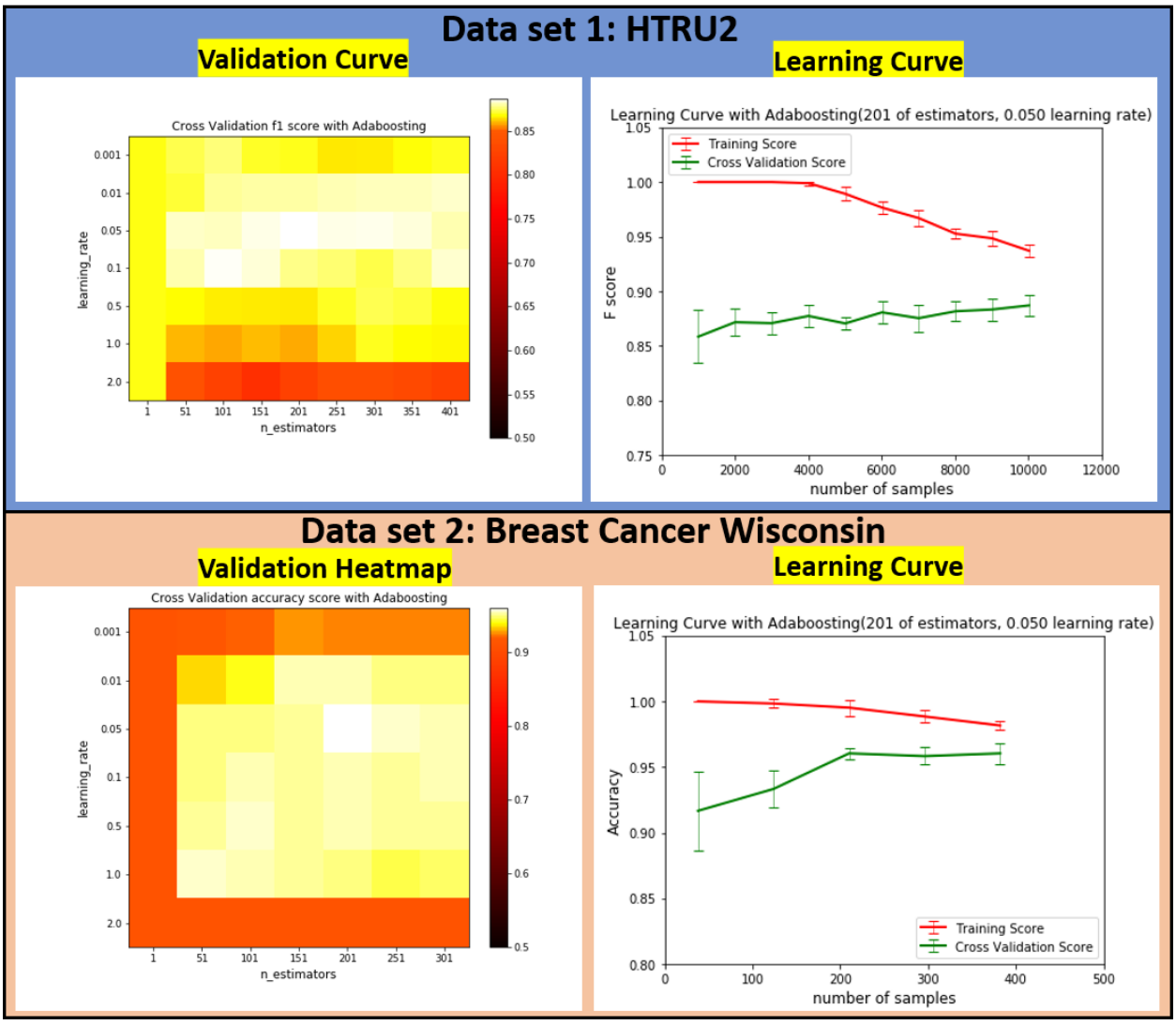

Figure 4. Adaboosting: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

For adaboosting, decision tree was used as the base classifier with more aggressive pruning (max depth 3 for HTRU2 and 1 for breast cancer data set). The algorithm was tuned over two hyperparameters, the number of estimators and learning rate. Model complexity increases as number of estimators increases, and lower learning rate shrinks the contribution of each estimator. For both data sets, cross validation scores increase as number of estimators increases and peak at n=200. As sample size increases, training and cross validation curves gradually converge, but there is still high variance for the HTRU2 data set at the current sample size of 10000 and the breast cancer data set at the current samples size of about 400, suggesting further increasing sample size would help with the overfitting problem in adaboosting.

5. **Neural Networks**

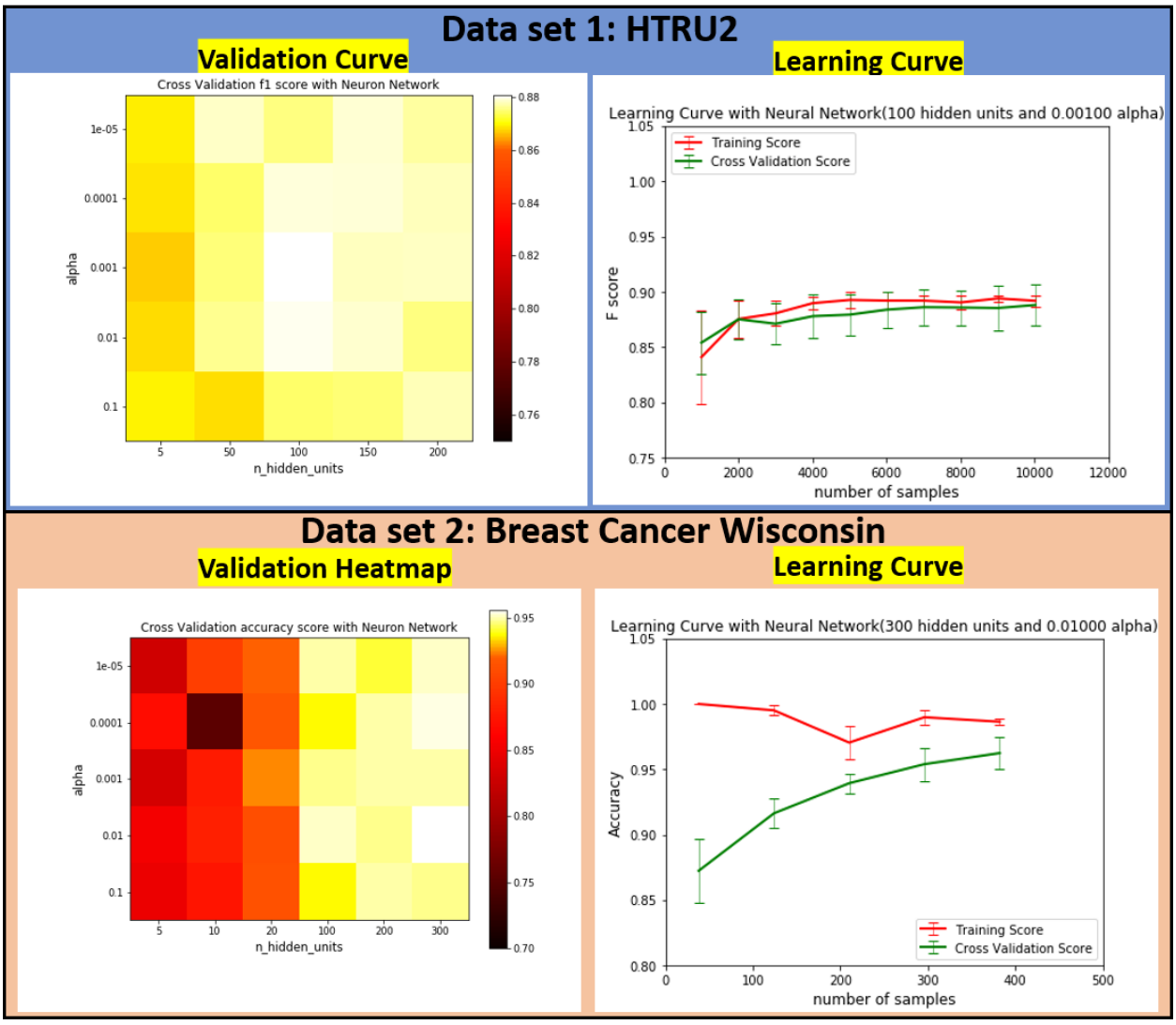

Figure 5. Neural Networks: Validation curve and learning curve
(5 folds cross validation. Error bars represent standard deviation)

For neural networks, I used relatively simple networks by setting the number of hidden layers to 1 with varying number of hidden units. The model was also optimized over the number of hidden units and the value of alpha, a parameter for regularization term. Generally, increasing alpha may overcome high variance by encouraging smaller weights. For the HTUR2 data set, neural networks yields high F score, and tuning these two hyperparameters does not significantly affect the algorithm performance. The inability to reach F1 score of higher than 0.9 probably indicates noise in the data. For the breast cancer data set, increasing number of hidden unites significantly increases the accuracy of the algorithm. And the cross validation learning curve gradually converges to training learning curve as sample size increases.

6. Comparison of the different algorithms.

Learning Algorithms	Test AUC Score	Test Accuracy	Test F Score	Training Time	Test Time	Hyperparameters
			HTRU2 Data Set			
Decision Tree	0.89898	0.97821	0.85921	0.0132626	0.00346114	Max depth = 5
k-Nearest Neighbor	0.90994	0.97952	0.86967	0.0091312	0.126501	k=5
SVM (linear kernel)	0.90952	0.98063	0.87560	3.70600	0.0311281	C=100
SVM (RBF kernel)	0.91270	0.98082	0.87753	0.454213	0.0916295	C=100 gamma=0.01
Boosting	0.91762	0.98045	0.87691	6.54262	0.102999	Decision tree with max_depth = 3 200 estimators Learning rate 0.05
Neural Networks	0.91987	0.98082	0.87953	4.57695	0.00571682	alpha=0.001 1 hidden layer with 100 nodes
			Breast Cancer Wisconsin Data Set			
Decision Tree	0.94077	0.94634	0.92617	0.0006914	0.00035675	Max depth = 5
k-Nearest Neighbor	0.95128	0.95610	0.93960	0.0007125	0.00180639	k=6
SVM (linear kernel)	0.94744	0.95122	0.93333	0.0027717	0.00044157	C=1
SVM (RBF kernel)	0.96744	0.96585	0.95425	0.0023225	0.00058448	C=100 gamma=0.001
Boosting	0.93795	0.94634	0.92517	0.309154	0.0149498	Decision tree with max_depth = 1 200 estimators Learning rate 0.05
Neural Networks	0.93795	0.94634	0.92517	1.75296	0.00088109	alpha=0.01 1 hidden layer with 300 nodes

Table 2. Comparison of different learning algorithms

For evaluating different algorithms, previously "untouched" test data were used. As the HTRU2 data set is severely biased towards the negatives, and the breast cancer data set is moderately imbalanced, different score metrics were measured including AUC score, F score and accuracy.

For the HTRU2 data set, SVM with RBF kernel, boosting and neural networks perform better than decision tree, k-nearest neighbors and SVM with linear kernel. This suggests that the data are non linear separable. SVM with RBF kernel outcompetes boosting and neural networks because of a significantly shorter training time, which is only 10% of that of neural networks, and 7% of that of

boosting. The HTRU2 data set does not suffer much from high variance, probably due the large number of instances.

For the breast cancer data set, simple methods such as k-Nearest Neighbors and decision tree with small max_depth yields high accuracy, F and AUC scores. This suggests that the data is linear separable and there is little noise in the data. SVM with RBF kernel performs best, with trivial training and test time. In contrast, boosting and neural networks show relatively low accuracy, despite the cost of significantly longer training time, and they both suffer from high variance probably due to the small sample size (<700).

Reference

1. https://www.coursera.org/learn/machine-learning/home/welcome
2. Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
3. R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach MNRAS, 2016.
4. W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
5. O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
6. http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html