**CS7641 Fall 2018 Assignment1 Supervised learning**

Xiaoxi Wang (xwang738)

email: xwang738@gatech.edy

your explanations should be both thorough and concise

**Summary**

**Datasets**

**Why they are interesting**

Nontrival, allows tuning and comparing of different algorithms

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Task | #Instances | #Attributes |  |  |  |  |  |  |
| HTRU2 | Binary Classification |  |  |  |  |  |  |  |  |
| Breast Cancer Wisconsin | Binary Classification | 683[2] | 9[3] | Real |  |  |  |  |  |

**Table 1.**

[2] Removed 16 instances with missing data

[3] ID number attribute was dropped

Scaling

Splitting into training (70%) and testing (30%)

**Results and Discussion**

1. Decision Tree

Pruning

The HTRU2 dataset

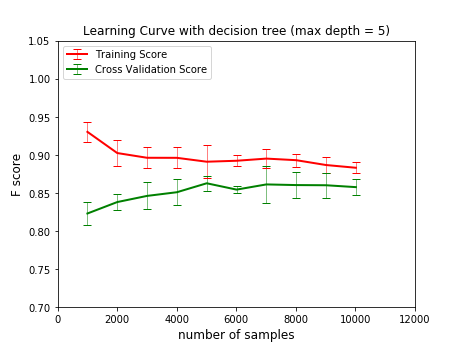
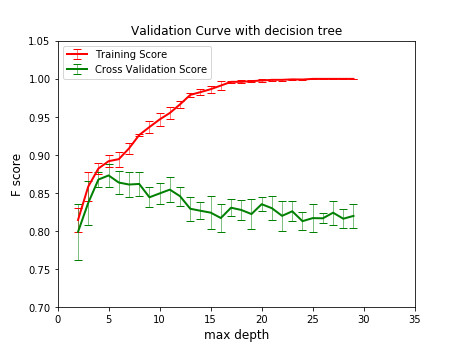


Figure 1.1 Figure 1.2

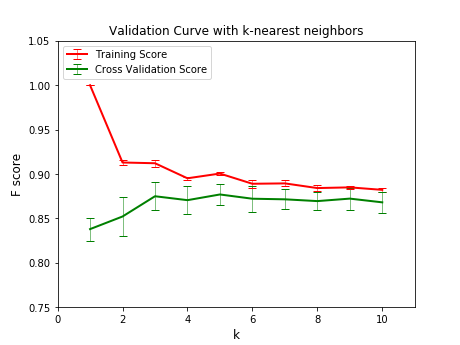
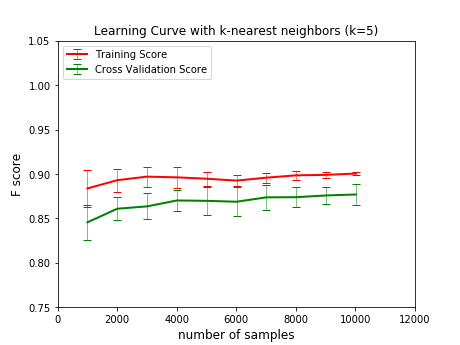
The cancer dataset

Figure 1.3 Figure 1.4

1. *k*-Nearest Neighbors

*The effects of different values of k*

The HTRU2 dataset

**

*The cancer dataset*

1. Support Vector Machine
   1. linear kernel

The HTRU2 dataset

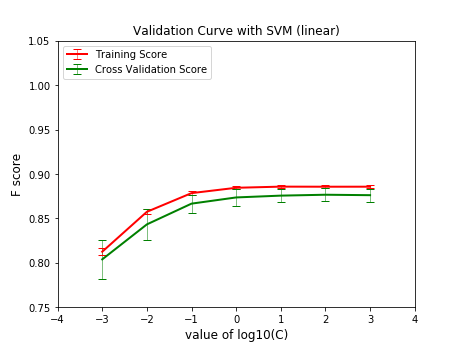
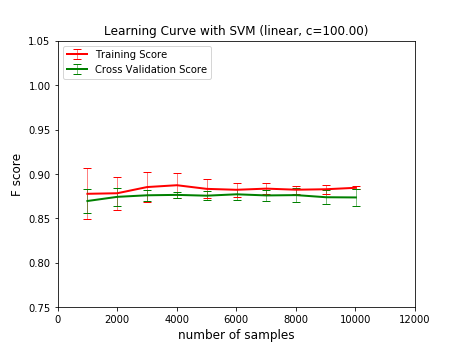


Figure 3.1 Figure 3.2

* 1. RBP kernel

1. Boosting

Base classifier decision tree with more aggressive pruning (max depth 3 for HTRU2 and 1 for breast cancer data set)

1. Neural Networks

1 hidden layer # of hidden unites the value of alpha, activation function

1. Compare and contrast the different algorithms.

Which algorithm performed best? How do you define best?

What sort of changes might you make to each of those algorithms to improve performance?

How fast were they in terms of wall clock time? Iterations?

How much performance was due to the problems you chose? How about the values you chose for learning rates, stopping criteria, pruning methods, and so forth (and why doesn't your analysis show results for the different values you chose?)?

5 folds cross validation for hyperparameter tuning

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Model Complexity curve over k

Reference

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.   
[[Web Link]](http://rexa.info/paper/b98475235164960529ad2ff9fda3816e9335cf8a)   
  
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.   
[[Web Link]](http://rexa.info/paper/90e988e83c7f06d2797b41580569c1f9a13f6749)