CS 4641: Supervised Learning

Bradley Reardon

February 10, 2019

1 Introduction

The purpose of this assignment is to evaluate various supervised learning techniques in the context of two classification problems. We will focus on analyzing and adjusting five learning algorithms using two publically-available data sets, through cross-validation and hyperparameter adjustments. Then, we can use our adjusted models to compare supervised learning techniques in the context of each of our classification problems.

2 Classification problems

For the purposes of this report, data sets were obtained from the UCI Machine Learning Repository. Each data set downloaded was processed with a custom script in the corresponding folder, **process.py**, which randomly separated the data into an 80% training data and 20% test data split using the **train_test_split** function from scikit-learn. The split data was then serialized using Python's **pickle** module, to ensure that the training and test sets remained constant for evaluation.

For the purposes of this assignment, cross-validation scoring of the various algorithms for each classification problem will be shown as the average score of 5-fold cross validation.

2.1 Car data set

The Car Evaluation Database was created in June, 1997 by Marko Bohanec. It contains 1728 instances and six attributes. The purpose of this database is to attempt to classify a car as an unacceptable, acceptable, good, or very good choice based on factors including cost of ownership, comfort, and safety. Full details about the data set can be found at the source link below.

Note that this specific data set contains only categorical attributes. As scikit-learn does not support non-continuous variables, a one-hot encoder was used to re-encode categorical features into multiple dimensions.

The problem at hand for this dataset is determining the acceptability for a car purchase based on the aforementioned attributes. Because this dataset notes that the instances in the data set completely cover the attribute space, this data set is interesting in particular due to its usefulness in comparing the optimization of different supervised learning algorithms.

Source: https://archive.ics.uci.edu/ml/datasets/car+evaluation

2.2 Breast Cancer Wisconsin data set

The Breast Cancer Wisconsin data set was donated to the UCI Machine Learning Repository in 1992, and contains data from one doctor's clinical cases, gathered from January 1989 to November 1991. In total, there are 699 instances signifying information about breast tumors such as clump thickness, uniformity in shape and size, and other screening factors. Data points are identified by two classes – benign or malignant. The features of the data points are encoded as 9 continuous attributes rating the screening factor from 1 to 10.

This data set contains unknowns in the form of question marks in the data. To deal with this, missing values were imputed, calculating missing data points based on the mean of other points in the specific column of the missing attribute.

The problem at hand for this dataset is determining whether a tumor is benign or malignant based on tumor screening characteristics identified in the data set.

Source: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

3 Decision trees with pruning

3.1 Parameter selection

GINI criterion was selected because – no discernible difference in training results, GINI is computationally faster

Max depth, min samples per leaf methodology... Car seemed to have much lower accuracy with lower depth – perhaps because the data set was complete. Chose max depth of 9 as a result. Cancer responsed better to more aggressive pruning with a max

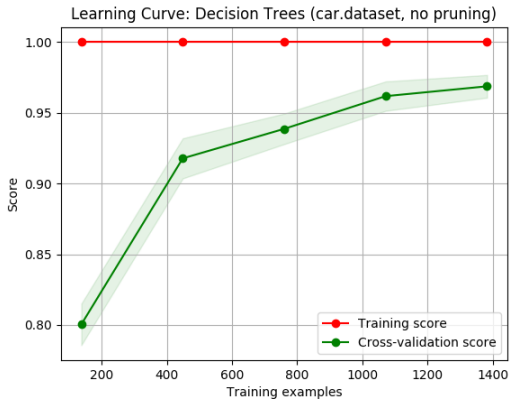
3.2 Performance

Wall clock time (pruning vs. nopruning)

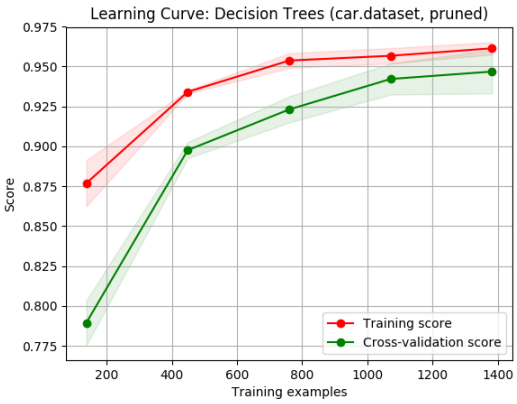
Learning curves for the car and breast cancer data sets, with and without pruning enabled, can be found in Figure 1 on the following page.

4 Neural networks

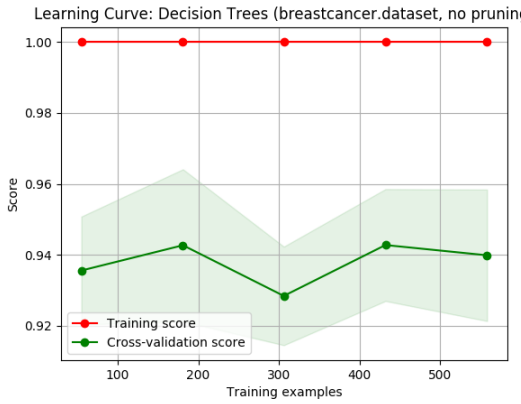
TODO



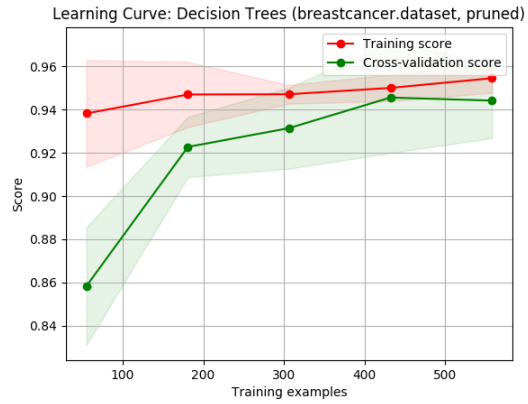
(a) Car data set, no pruning



(b) Car data set, with pruning



(c) Breast cancer data set, no pruning



(d) Breast cancer data set, with pruning

Figure 1: Learning curves for the car and breast cancer data sets using a decision tree classifier, with and without pruning.

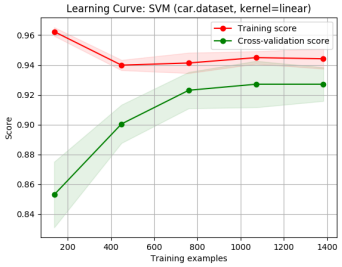
5 Boosting

TODO

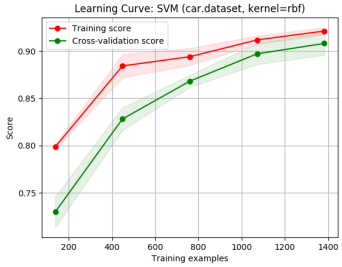
6 Support vector machines

6.1 Performance

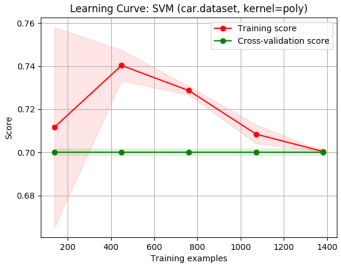
Learning curves for the car and breast cancer data sets can be found in Figure 2 on the next page.



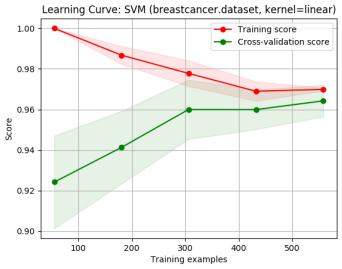
(a) Car data, linear kernel



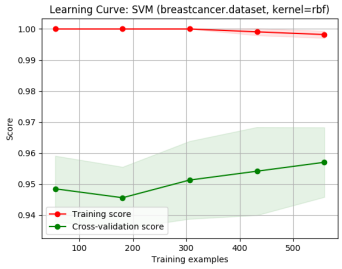
(b) Car data, RBF kernel



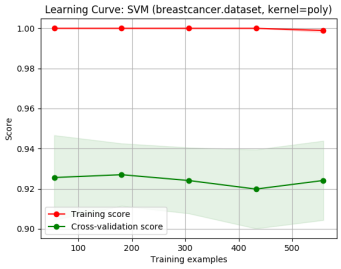
(c) Car data, poly kernel



(d) Breast cancer, linear kernel



(e) Breast cancer, RBF kernel



(f) Breast cancer, poly kernel

Figure 2: Learning curves for the car and breast cancer data sets using an SVM classifier, with linear, RBF, and poly kernels.

7 k-nearest neighbors

TODO

7.1 Parameter selection

car k=11 cancer k=5

Lowest k with highest cross-validation

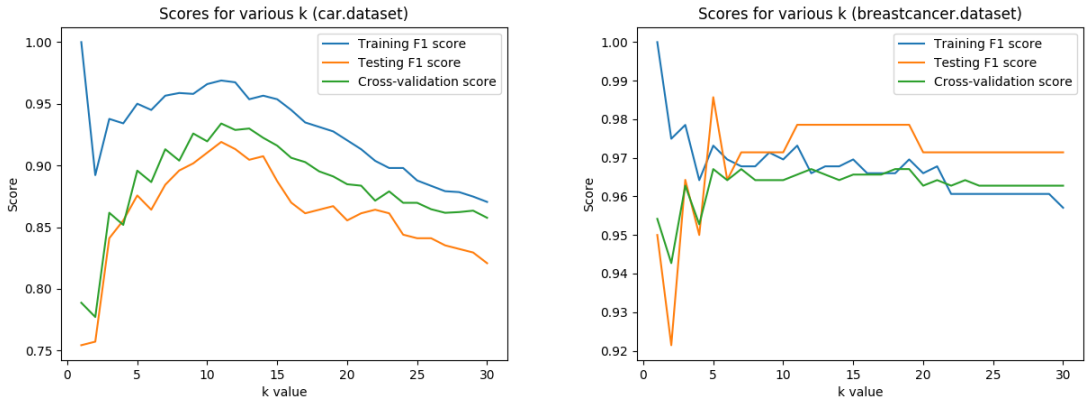
TODO Figure 3 on the following page

7.2 Performance

TODO Figure 4 on the next page

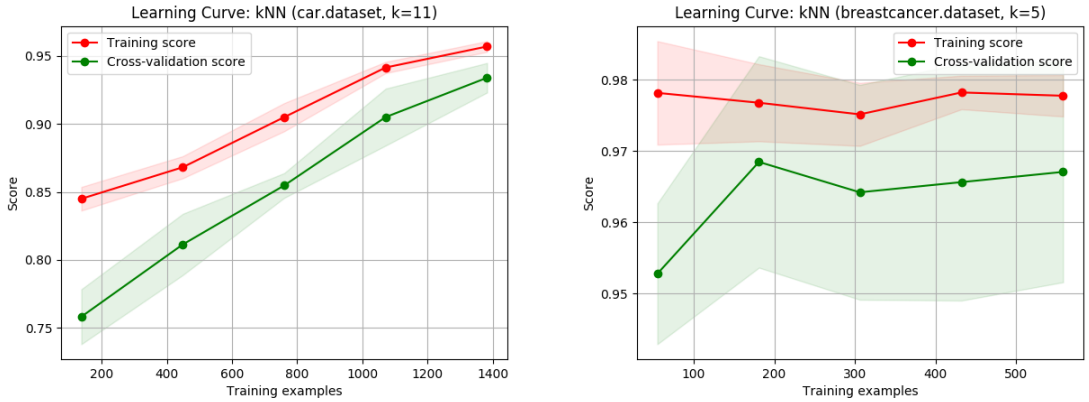
8 Analysis

TODO



(a) Car data set **(b)** Breast cancer data set

Figure 3: Testing k-nearest neighbors with various values of k



(a) Car data set, $k = 11$ **(b)** Breast cancer data set, $k = 5$

Figure 4: TODO TODO

9 Discussion

TODO

10 Conclusion

TODO