

Georgia Institute of Technology

Hui Xia (hxia40)

903459648

Assignment 1: Supervised Learning

CS7641 Machine Learning, 2019 Fall

**1. Introduction**

Supervised learning is a major category in machine learning. The target of this assignment is to gain understanding of commonly used algorisms under various circumstances. In this assignment, two datasets are analyzed using five algorithms (Decision trees with some form of pruning, Neural networks, Boosting, Support Vector Machines, and k-nearest neighbors), as per required. In this report, I will firstly describe the dataset used for this assignment, then describe and discuss the learning curve analysis and model complexity analysis performed on these datasets, using the said algorithms.

**2. Experiment Methods**

2.1. Datasets

In this work, two datasets of similar size are analyzed. One dataset is a slice from Fashion-MNIST. Each row of this dataset describes a 28 x 28 grayscale pixel image (each pixel is represented with a value of 0-255), thus, the training and test datasets have (28 x 28) = 784 columns, i.e. X1-X784 [ref]. Each row is also associate with a label y (value of 0-9) that respectively represents different 10 kinds of clothing. All of the y values have equal proportion in this dataset. For this assignment, the first 2000 rows in the train dataset, and the first 500 rows in the test dataset are used. The total size of the training and testing dataset is 5.5 MB.

Another dataset is the Epileptic Seizure Recognition Data Set (ESR) from UCI Machine Learning Repository [ref]. This dataset is collected as the recording of brain activity via electric signals. Each row of the dataset describes 178 data points (i.e. X1-X178) happened within 1 second in a person’s brain. Each row is associated with a label y that represent 5 different status (eyes open, eyes closed, healthy area, tumor area, and seizure) of the person tested. All of the five kinds of y status have equal proportion in this dataset. For this assignment, the first 6000 rows in the dataset are split into a 5000-row train and 1000-row test dataset. The total size of the training and testing dataset is 5.5 MB.

I found that there are several interesting points among these two datasets. First, as MNIST is an image-based database while ESR is not, it is interesting to find if more complex model, such as artificial neuron network (ANN), will perform better (compared with other algorithms) in MNIST compared with in ESR. Second, the ESR dataset is not only of my personal research interest (I hold a Ph.D. in biomedical engineering), but also a ‘weak’ classification dataset that is well-known as hard to gain good performance. It will be interesting to find out that if I can find out a way to build a high-performance algorithm. Third, as described above, both of the datasets are used for a classification problem, and I purposely sliced them into same storage size. This fact itself make the compare between the two datasets interesting - not only on algorithm performance, but also on running time.

2.2. Experiment design

The training set from both of the datasets are subjected to a classification study against their respective y values. In details, a learning curve study is firstly performed on the datasets using five algorithms (decision tree, boosting using decision tree, neural networks, support vector machines, and k-nearest neighbor) under default parameters. The classifier and the related default hyper-parameters are listed in **Table 1**. In the learning curve study, a cross-validation is performed using 10-section ShuffleSplit. The training score and cross-validation score is plotted against 50 different sample sizes (from 2% to 100% of the size of the training set). For the iterative algorithms (i.e. boosting, ANN, and SVM), the learning curve study is also performed as plotting the training score and cross-validation score against iterations. The iterations of boosting and SVM are controlled using their respective hyperparameters n\_estimatorsand max\_iter, while the iterations of ANN is realized using its partial\_fit method.

**Table 1**. Classifier and default hyperparameters adopted in first learning curve study.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Classifier | Hyperparameter 1 | | Hyperparameter 2 | | Iterator | |
| name | default value | name | default value | name | default value |
| Decision tree | sklearn.tree.  DecisionTreeClassifier | *max\_depth* | *None* | *min\_samples\_leaf* | *25* |  |  |
| Boosting using decision tree | sklearn.ensemble.  AdaBoostClassifier (base\_estimator=  sklearn.tree.  DecisionTreeClassifier) | *min\_samples\_leaf* | *25* | *learning\_rate* | *1* | *n\_estimators* | *5* |
| Artificial neural network | sklearn.neural\_network.  MLPClassifier | *hidden\_layer\_size* | *(5, )* | *alpha* | *0.0001* | *partial\_fit* | *N/A* |
| k-nearest neighbors | sklearn.neighbors.  KNeighborsClassifier | *n\_neighbors* | *5* | *algorithm* | *’auto’* |  |  |
| Support vector machines | sklearn.svm.SVC | *C* | *1* | *kernel* | *’rbf’* | *max\_iter* | *-1* |

Then, a model complexity analysis is performed to the two datasets using the said five algorithms. In the model complexity analysis, while keeping all the other hyperparameters at default, one of the hyperparameters listed in **Table 1** are changed, and the resulted classification performance score, for both the training set and cross-validation set, is plotted against the changing hyperparameter. For each hyperparameter, 50 different values are tested on the training and the validation set to find the optimal hyperparameter setting. The value of the hyperparameter that grant the maximum cross-validation score is adopted as optimal. Similar to the learning curve study, the cross-validation is performed for the model complexity analysis using 10-section shuffle-split.

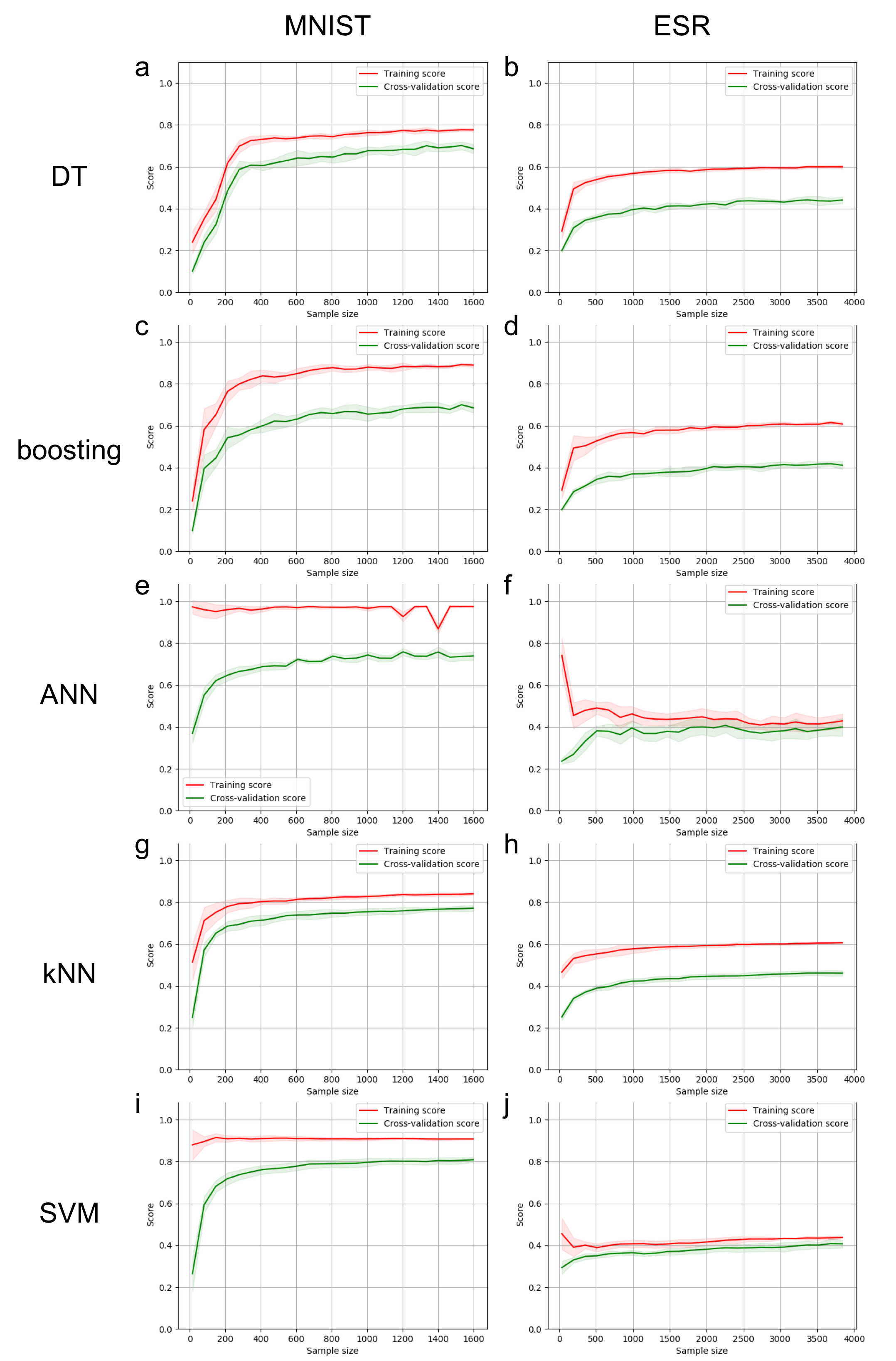
Using the optimal hyperparameters, another set of learning curve study is performed on the datasets using the five algorithms. The training score and cross-validation score is plotted against 50 different sample sizes (from 2% to 100% of the size of the training set), same as the first set of learning curve study. The learning curve plot generated before and after the model complexity analysis is compared, to find the effectiveness of the hyperparameter optimizing on the performance of the respective algorithm. Finally, the running time and the performance score of the five algorithms (using the optimal hyperparameters) on the testing dataset is measured to investigate inter-algorithm difference.

**3. Results**

3.1. Learning curve using default hyperparameter

The learning curve plot using said algorithms on datasets MNIST and ESR against sample size and against iteration are shown in **Figure 1** and **Figure 2**, respectively**.** As both datasets are multilabel classification problems, accuracy score is plotted on the y-axis to describe the performance of the algorithms. It is notable that because the training and validation data are subjected to cross-validation and treated using ShuffleSplit, when the training size is very small, the over-fitting (which is due to the that model fit training set well while not generalizing to validation set) are avoided for the “simpler” algorithms (e.g. decision tree, kNN, and boosting over decision tree).

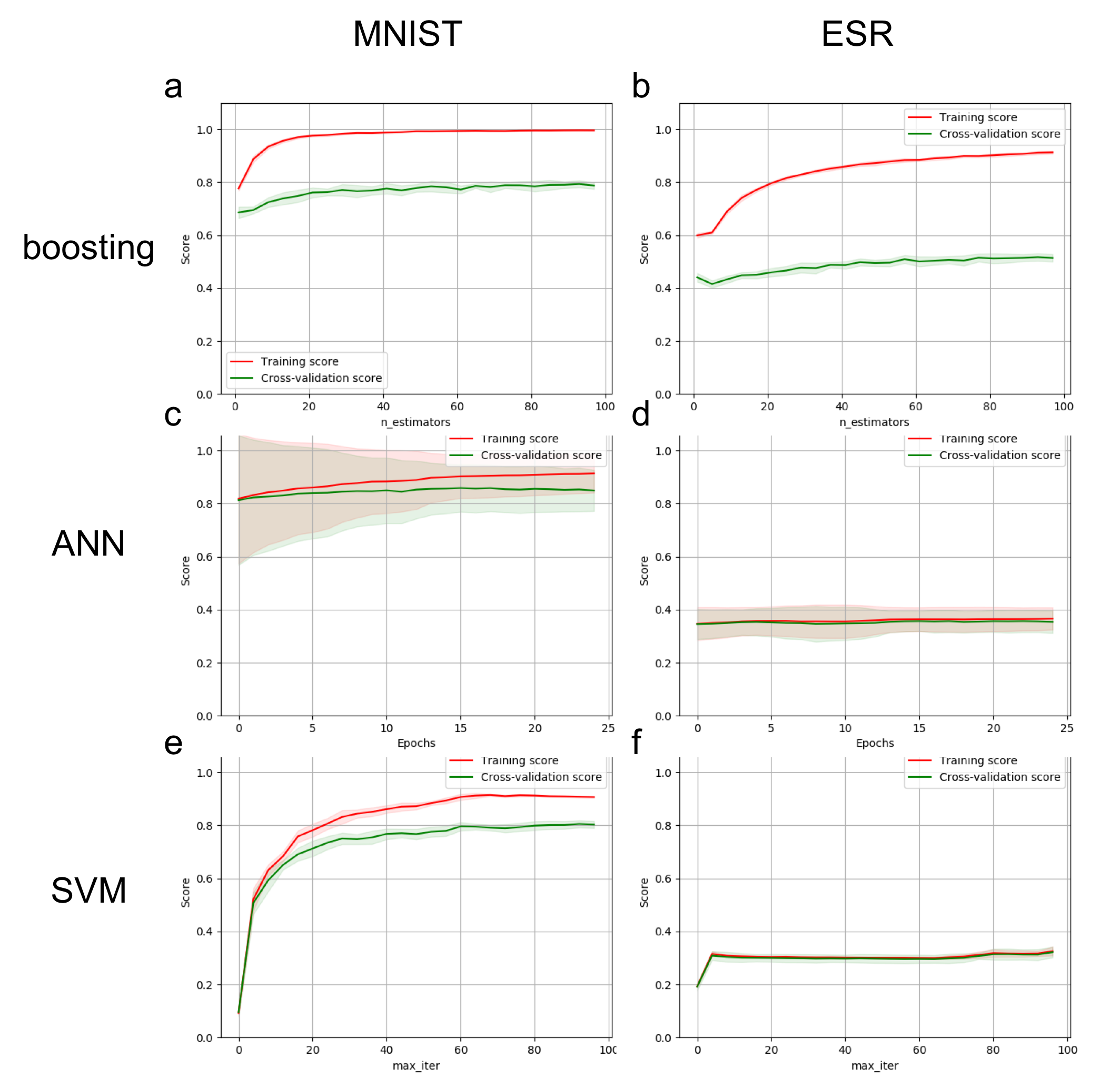
Using decision tree and the boosting-decision tree algorithms, the training and cross-validation score for both MNIST and ESR datasets plateaus alone the increase of sample size (**Fig. 1a-d)**, indicating that the addition of more datasets will not significantly improve their performance. As the training/cross-validation curve does not show the trait to converging over the growth of sample size, the model could be suffering from a high variance problem. Also, as the training curve is far from 1.0, the models could be also suffering from a high bias problem. As it is premature to judge which side should we move in the bias-variance trade-off, it would be reasonable to tune the min\_samples\_leaf over a range between 1 and 200 to show if the pre-pruning of the decision tree by increasing or decreasing the complexity could improve their performance. Additionally, when the model takes more than 40 iterations (i.e. when n\_estimators >= 40, shown in **Fig. 2a-b**), the large train/validation gap and good train set performance indicate that the boosting model suffers from a high-variance and low-bias problem. Decreasing model complexity (e.g. increase leaf size and/or decrease learning rate) could improve the performance thereof.



**Figure 1:** Learning curve over sample size using default hyperparameter

Interestingly, using ANN and SVM, the training and cross-validation score for MNIST and ESR performs differently. On MNIST, the ANN and the SVM algorithms show a typical low bias and high variance problem, features by the large gap between the train/validation curves and high performance on the training curve (**Fig. 1e and i**). Quite on the opposite side however, for ESR, the ANN and the SVM algorithms show a typical high bias and low variance problem, which is features by the small gap between the train/validation curves and poor performance on the training curve (**Fig. 1f and j**). Along the increase of sample size, the training and cross-validation score for both MNIST and ESR datasets plateaus, indicating that the addition of more datasets will not significantly improve their performance. Additionally, along the increase of iteration times, both ANN and SVM show low-bias and low-variance for MNIST, and high-bias and high-variance for ESR, indicating that the increase of iterations (when epochs >= 15 for ANN and max\_iter >=60 for SVM) can only improve the performance on both datasets in a limited manner. Thus, I will target to bring the model complexity down for the MNIST dataset, and up for the ESR.

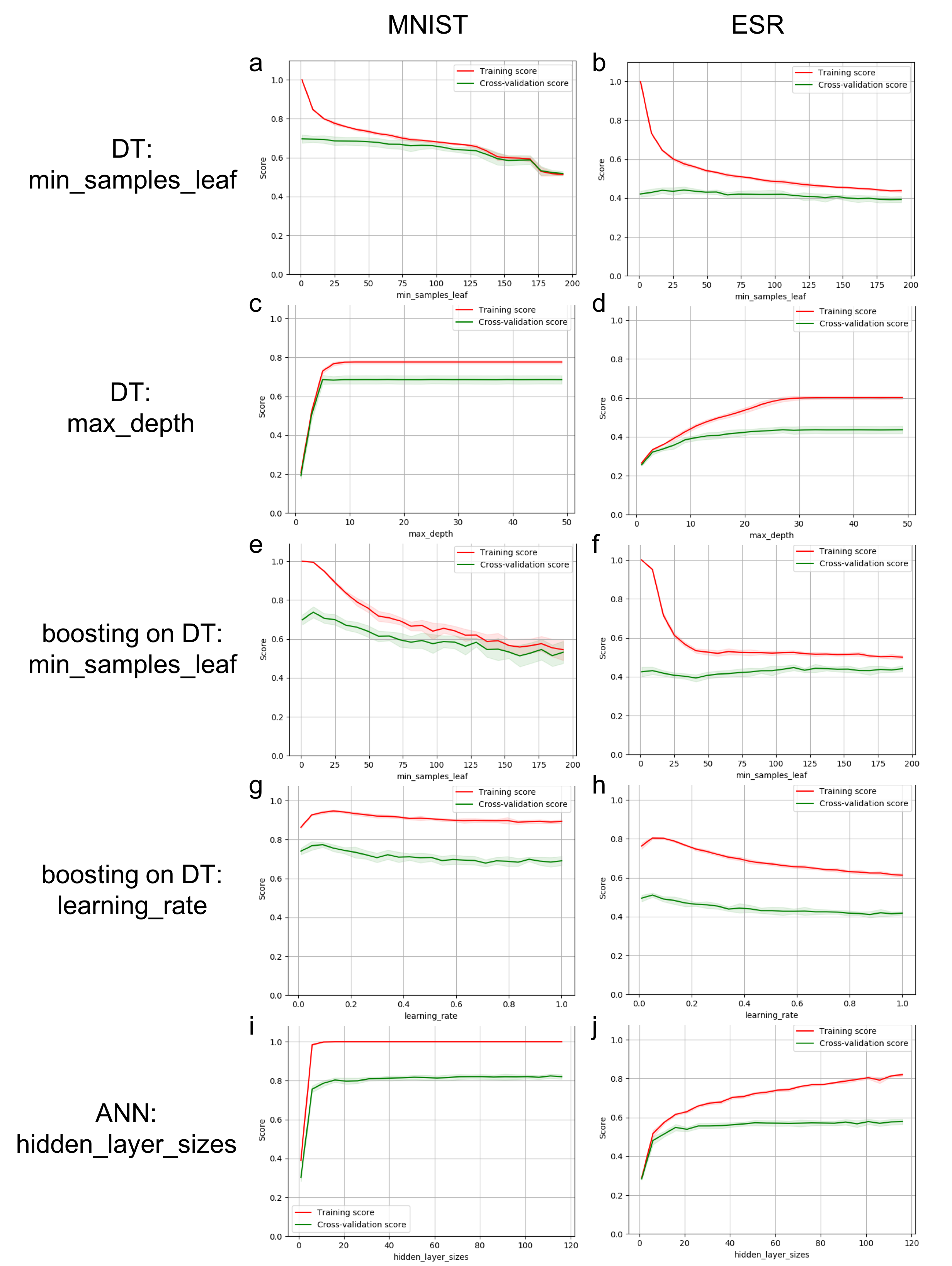
For both MNIST and ESR, the learning curve using kNN looks very similar to that using decision tree (**Fig. 1g-h**). That is, the training and cross-validation score for both datasets plateaus along the increase of sample size, and it is hard to decide if the model is suffering from high-bias or high-variance problem. Thus, I will tune the n\_neighbors to investigate for a better performance on both directions from the default value of n\_neighbors = 5.

****

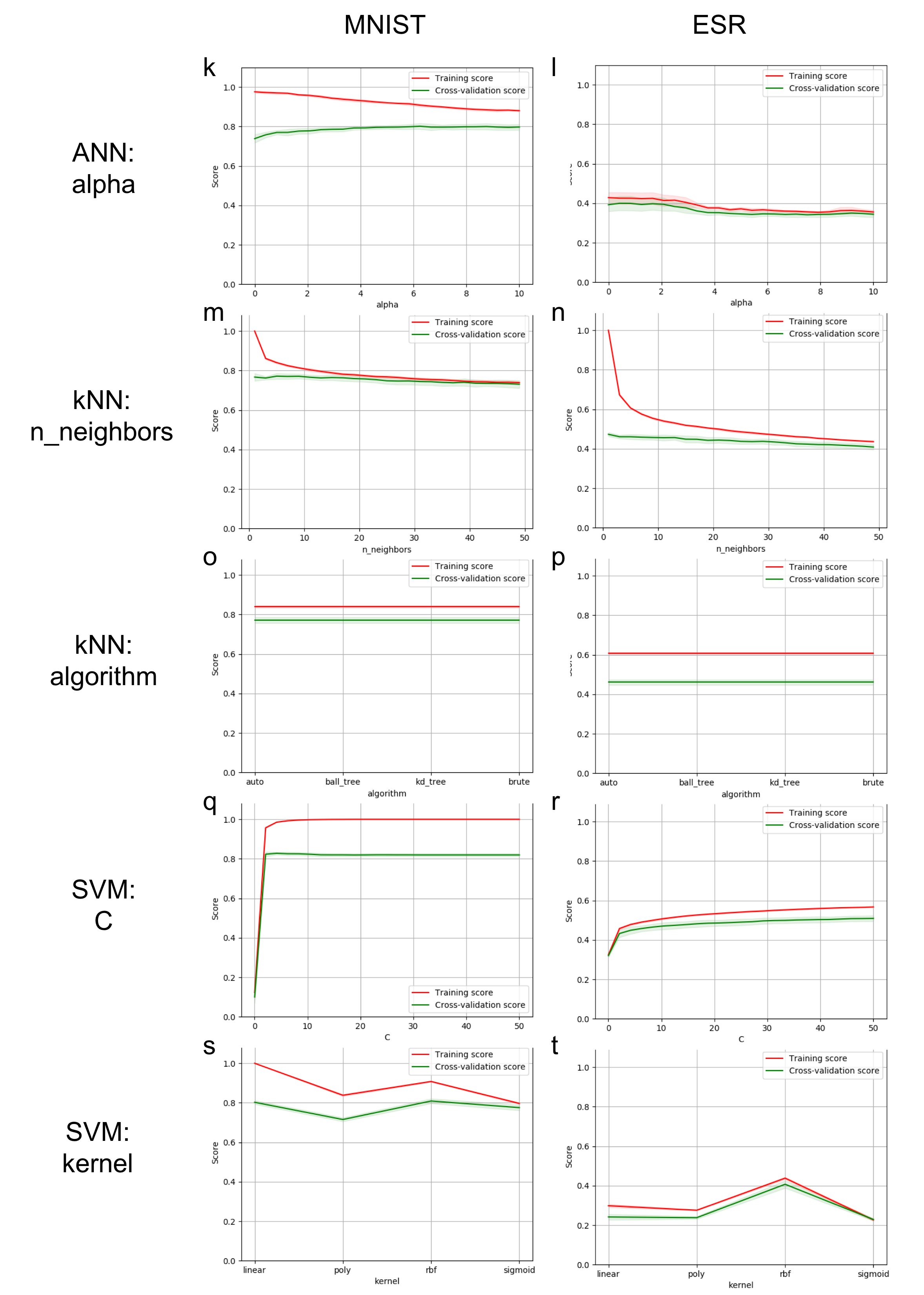
**Figure 2:** Learning curve over iteration using default hyperparameter

3.2. Model complexity analysis

As discussed in **Sections 2.2** and **3.1**, while keeping all the other hyperparameters (or iterators, if applicable) at default, one of the hyperparameters listed in **Table 1** is changed and plotted against the changing hyperparameter. **Figure 3** demonstrate the plotted validation curve for all the five algorithms using MNIST and ESR datasets.



**Figure 3:** Model complexity analysis

****

**Figure 3 (continued):** Model complexity analysis

As shown in **Figure 3a-b**, pre-pruning of decision tree using min\_samples\_leaf on both of the datasets, increasing min\_samples\_leaf decrease the decision tree model’s variance and increase the bias thereof. This is reasonable, as increasing the size of leaf will make the model less complex, thus making the model to take a more generalized point of view (i.e. to be more biased). Additionally, the performance of the model on MNIST decreases alone the increase of leaf size, but the performance of the model on ESR peaks when min\_samples\_leaf = 33. On the other hand, the pre-pruning of decision tree using max\_depth show that higher value of max\_depth will lead to better model performance (**Fig. 3c-d**). Thus, I will take min\_samples\_leaf = 1 and max\_depth = None for MNIST, and min\_samples\_leaf = 33 and max\_depth = None for ESR as the optimal.

Boosting using decision tree show a similar trend when adjusting on min\_samples\_leaf on both of the datasets (**Fig. 3e-f**). Again, increasing leaf size decrease the decision tree model’s variance and increase the bias. [discussion] It also is interesting to find that the ESR dataset always performs better using larger leaf size compared with the MNIST dataset, using both decision tree model and the boosted decision tree model. This is quite counterintuitive, as we would consider the MNIST, as an image-recognizing dataset, would favor a more “general”, or more biased model, just like we consider human use their ‘feelings’ on certain features to recognize photos. However, we see that the image-based MNIST dataset require a more complex model to perform well.

Additionally, the performance of the model on MNIST and on ESR peaks when their min\_samples\_leaf = 9 and = 113, respectively. Adjusting on learning\_rate shows a trend similar to adjusting the min\_samples\_leaf. Alone the increase of learning\_rate, the performance of the model firstly briefly increases, then decreases (**Fig. 3 g-h**). As a result, the optimal parameter learning\_rate can be found at 0.0925 and 0.05125, for the MNIST and the ESR datasets, respectively. When the number of iterations of the boosting model stays constant, a small learning rate will make the model to underfit, or has higher bias [ref]. The high performance using low learning rate indicate that the boosting model on both datasets prefer to bring up the bias, which agrees with our hypothesis in **Section 3.1**.

From the learning curve study, we mentioned that the ANN and the SVM algorithms for MNIST show a low bias and high variance problem, while for ESR, a high bias and low variance problem. Thus, here we strive to bring the bias up for the MNIST dataset, and to bring the bias down for the ESR. This can be achieved by adjusting the hyperparameter alpha from ANN, and C from SVM. Lower alpha values tend to overfit, or be low on bias but high on variance, and *vice versa* . On the opposite, lower C value tend to underfit, or be low on variance buy high on bias, and *vice versa*, which is similar to hyperparameter learning\_rate used in the boosting model. As shown in **Figure 3k-l and q-r**, MNIST prefer higher alpha in ANN but lower C in SVM, while ESR prefer lower alpha in ANN buy higher C in SVM. Again, this observation agrees with our hypothesis (that we need to bring the bias up for MNIST, and down for ESR) here. While in the measurable range (up to C = 1000), the performance of SVM on ESR monotonically increase with higher C (Fig. 3r). However, choosing higher C will also bring up the calculation time. Thus, for ESR, C = 50 is used here as optimal. Similarly, for both datasets on ANN, higher hidden\_layer\_size will result in better performance but longer time consumption (**Fig. 3i-j**). Thus, hidden\_layer\_size = (50, ) is used.

Finally,

, but the performance of the model on ESR peaks when min\_samples\_leaf = 33.

Thus, if there is a trade-off dilemma between choosing low bias and low variance (while choosing neither can significantly improve the model performance), I will prefer the low variance over low bias, as the combination of low variance/high bias will perform better on un-seen testing data, which is what I am going to test these models on at the end of this report.

**4. Discussion**

[Wikipedia] The bias-variance tradeoff is a central problem in supervised learning. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit but may underfit their training data, failing to capture important regularities.

Models with high variance are usually more complex (e.g. higher-order regression polynomials), enabling them to represent the training set more accurately. In the process, however, they may also represent a large noise component in the training set, making their predictions less accurate – despite their added complexity. In contrast, models with higher bias tend to be relatively simple (low-order or even linear regression polynomials) but may produce lower variance predictions when applied beyond the training set.

The second dataset is the Bank Marketing Data Set from UCI Machine Learning Repository [ref]. This dataset is collected as client marketing information of a bank. Each row of the dataset describes 16 different kinds of information (i.e. X1-X16) of a client, including age, job (such as “blue-collar” or “management”), marital status, education, etc, and a binary column that describes whether the client has subscribed a deposit.

Prior to other handling steps, the data in Istanbul.csv was randomly shuffled once (random seed is 123) for improved analysis result. Then, a Decision Tree Learner (DTL) was trained using the first 60% of the data in Istanbul.csv and tested using the other 40% of the data.

Among the *n* columns of the data, the last column is referred as the *Y* column, which represents the result that to be decided by the rest of the columns (i.e. the *X1, X2 … Xn-1* columns). The *Y* column that was generated by passing the *X* columns through the DTL was compared with their original counterparts before DTL training, and then used to calculate the in-sample (using the training data) and out-of-sample (using the testing data) Root Mean Square Error (RMSE) values.

The in- and out-of-sample RMSE values were calculated (by code) when the DTL use a range of leaf size values (1-100). The RMSE values were then plotted against the leaf size vales.

* 1. Result

The plotted data is demonstrated in **Figure 1**. In general, when the value of leaf size is larger than 5, both in- and out-of-sample RMSE show an ascending trend when the leaf size increases. However, when the leaf size is lower than 5, the out-of-sample RMSE decreases when the leaf size increases, as shown in **Figure1a**. On the other hand, the in-sample RMSE increases in this range. When leaf size is smaller than 20, the in-sample RMSE is lower than the out-of-sample RMSE. When leaf size is larger than 80, the in-sample RMSE is higher than the out-of-sample RMSE. When leaf size is in between 20 and 80, the in-sample RMSE and the out-of-sample RMSE have similar value.

* 1. Discussion

In this experiment, the data was randomly shuffled once before it was used to train the DTL. However, while handling financial data, this ‘shuffle before training’ method in in general being suggested against. This is because that the data used for training should always be generated earlier than the data used for testing, to avoid the ‘predict the future’ effect.

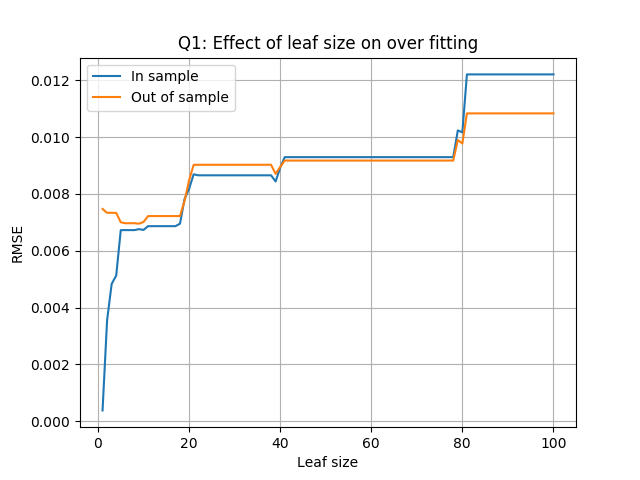
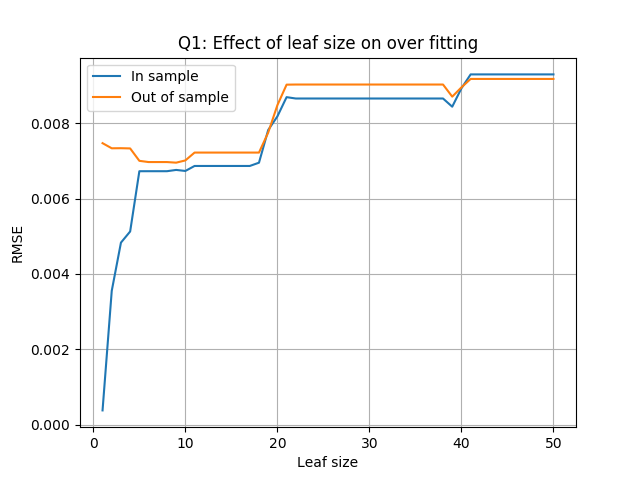
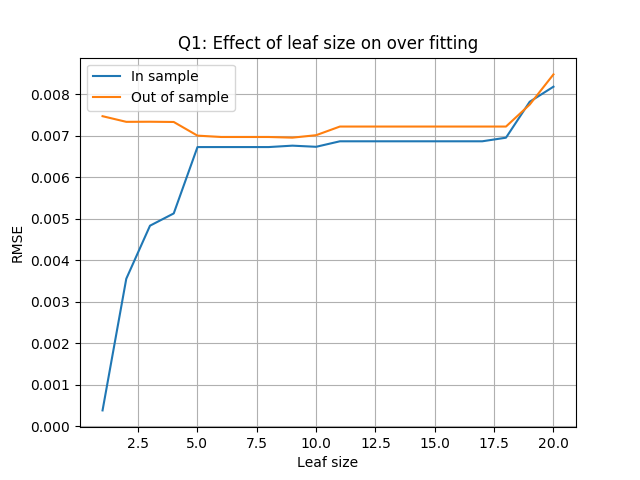
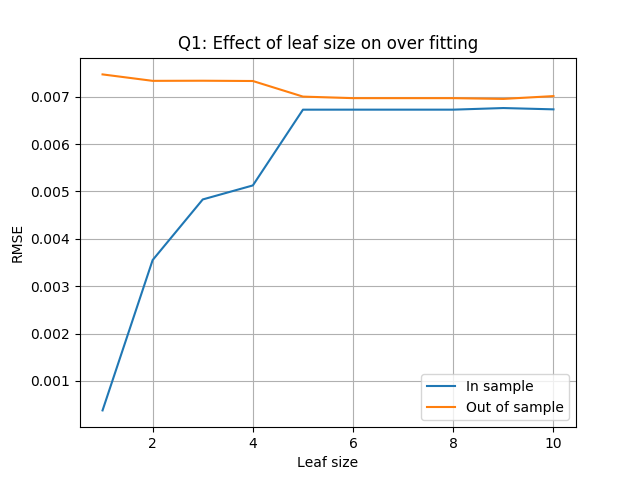
In general, the RMSE value (for both in- and out-of-sample) increases when the leaf size increases. This is expected, as for a certain data set, training using larger leaf size will result in a ‘rougher’ decision tree, and thus more inaccurate prediction. The trend deviation between the in- and out-of-sample RMSE values suggest that when the leaf size is too small (<5 in this case), the DTL describe the samples in an overly detailed manner. Thus, the DTL overfits the data when leaf size < 5.

It is also noticeable that when the leaf size is using a larger value (> 80 in this experiment, as demonstrated in **Figure 1d**) the RMSE of out-of-sample become lower compared with the in-sample. Although this case is not expected, as the training data should fit better with the DTL compared with other data. However, it is reasonable to consider that a certain number of leaves should be needed to generate enough ‘cases’ to the learner to study, and thus to keep the DTL reliable. As we have discussed above, if the leaf size increases while the size of the training data cannot change accordingly (and this is a common scenario, as data is precious), the generated number of leaves will decrease accordingly. For an extreme case, if the leaf size equal to the size of the data, the trained DTL will only have one leaf, and is not necessarily useful anymore. Thus, there should be a trade-off between leaf size and leaf numbers exist, which defines the upper-limit of the leaf size.

* 1. Conclusion

When the leaf size is too small (<5 in this experiment), over fitting does occur in the DTL.

**Figure 1:** Plotting in sample and out of sample RMSE values versus the leaf size using DTL. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100.



a

c

d

b

1. **Question 2**: Can bagging reduce or eliminate overfitting with respect to leaf\_size? Again use the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf\_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.
   1. Experimental Methods

Prior to other handling steps, the data in Istanbul.csv was randomly shuffled once as described in the section 1.1. Then, a Bagging Learner (BL) that contains 20 DTL was trained using the first 60% of the data in Istanbul.csv and tested using the other 40% of the data. To obtain smoother lines and analyze the data more accurately, another set of experiment is performed, which performs 20 iterations under each leaf size.

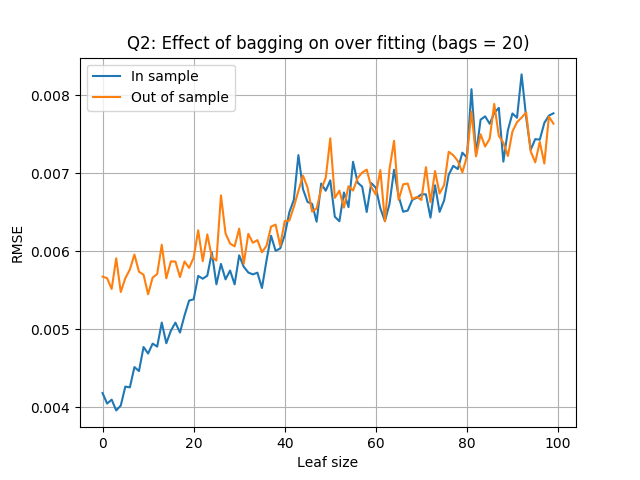
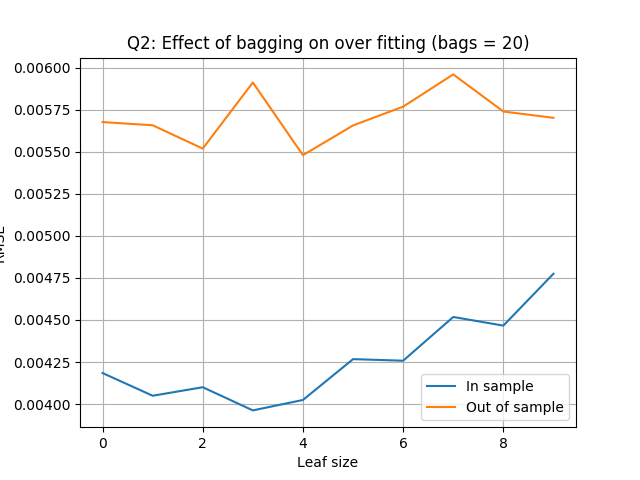
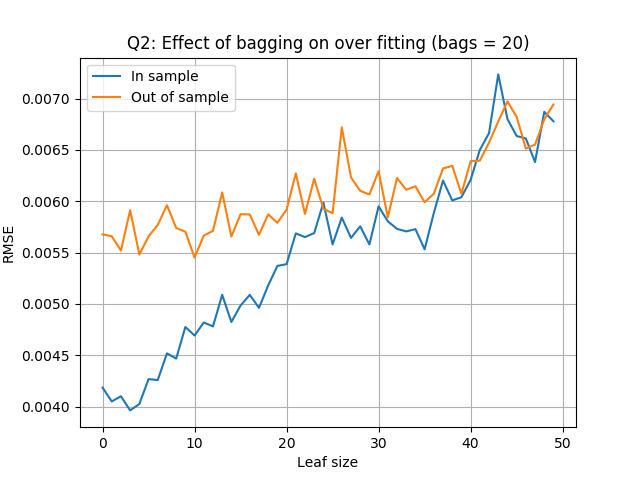
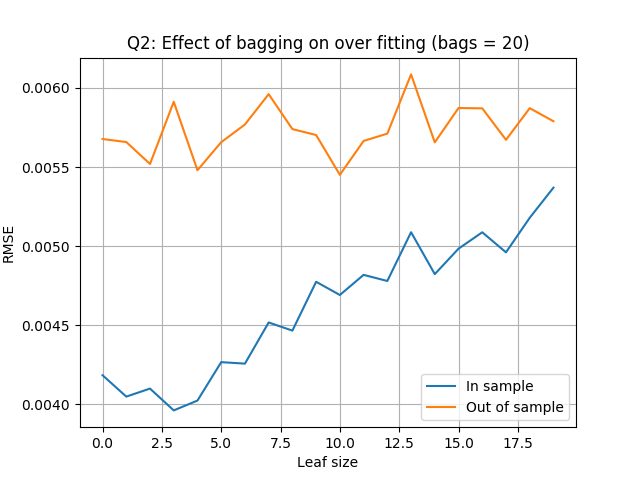
Similar as in section 1.1, for both experiments (i.e. one with single iteration per leaf size, another with 20 iterations per leaf size), the in- and out-of-sample RMSE values under a range of leaf size (1-100) were calculated, which were then plotted versus the leaf size.

* 1. Result

The in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL, with single iteration per leaf size is shown in **Figure 2**, and the figure with 20 iterations per leaf size is shown in **Figure 3**.

Both **Figure 2** and **Figure 3**, suggest that the in- and out-of-sample RMSE have an ascending trend when the leaf size increases. Unlike the DTL shown in **Figure 1**, no over fitting is observed.

**Figure 2:** Plotting in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100.



a

c

d

b

Compared with the DTL, the RMSE values in the BL is in general (when the leaf size is larger than 1) lower. When leaf size = 100, the in- and out-of-sample RMSE generated by the BL are between 0.007 and 0.008, while the counterpart that were generated by the DTL are above 0.01. However, the in-sample RMSE generated by DTL is lower than 0.001, which is lower than the counterpart generated by the BL.

* 1. Discussion

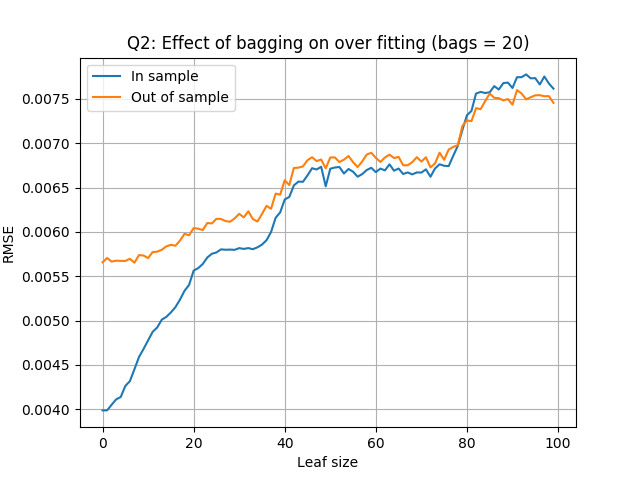
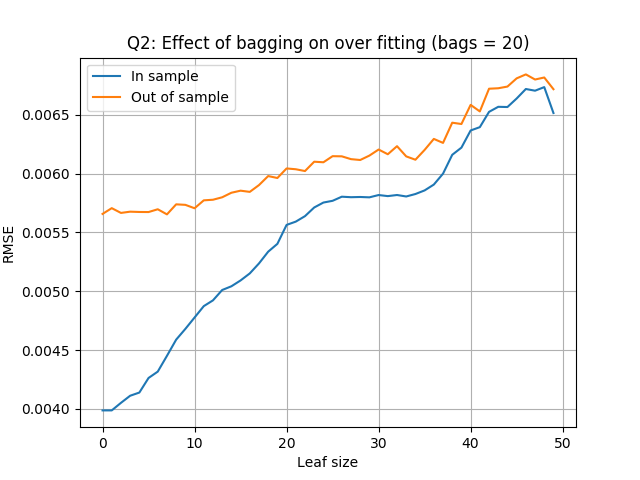
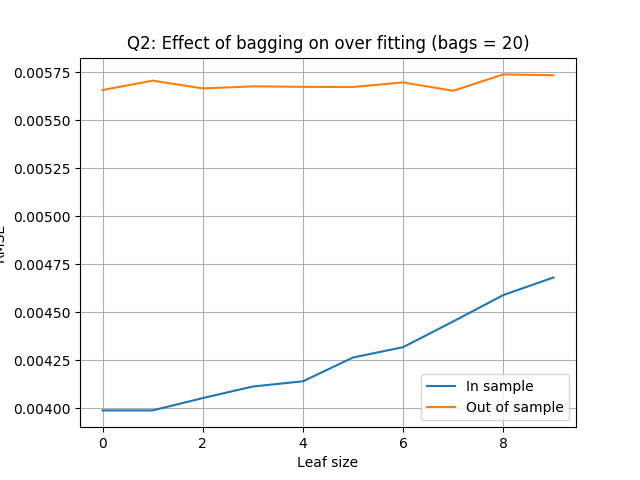
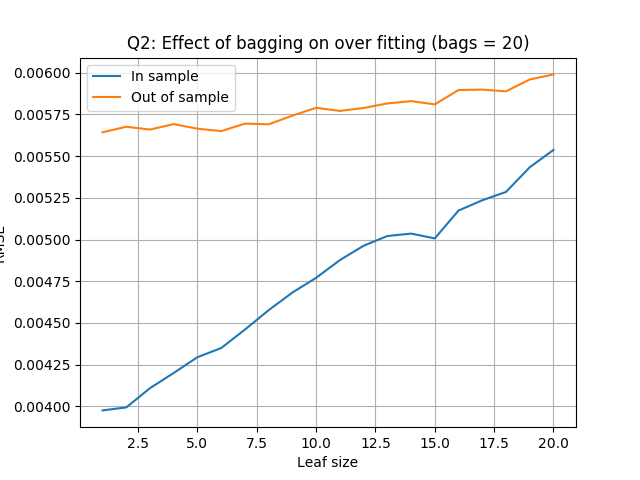
Similar to the DTL experiment, larger leaf size in the BL will result in a higher RMSE value, for both the in- and out-of-sample data. Compared with the DTL, using the same set of data, the BL nearly always grant a lower RMSE. The only scenario that the DTL has a lower RMSE value happens when the leaf size is 1, for the in-sample data. However, this scenario could be considered as useless, as it basically means to ‘predicting’ the data while the data analyzer already has it.

In both experiment performed using the BL (i.e. one iteration per leaf size and 20 iterations per leaf size), no overfitting could be observed.

* 1. Conclusion

Bagging does greatly reduce overfitting, if not completely eliminates it.

**Figure 3:** Plotting in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL, with 20 iterations for each leaf size. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100



a

c

d

b

1. **Question 3**: Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, do not use two measures for accuracy.) Note for this part of the report you must conduct new experiments, don't use the results of the experiments above for this.
   1. Experimental Methods

Prior to other handling steps, the data in Istanbul.csv was randomly shuffled once as described in the section 1.1. the DTL and Random Tree Learner (RTL) were trained using the first 60% of the data in Istanbul.csv and tested using the other 40% of the data.

For both DTL and RTL, the in- and out-of-sample Mean Absolute Error (MSE) and learner running time were calculated under a range of leaf size (1-100), which were then plotted versus the leaf size. MAE is a measure of difference between two continuous variables, which is calculated using the equation below:

Where *Ypred* and *Ygiven* stand for the predicted *Y* values and the given *Y* values, respectively, and *n* is the total numbers of *Y* values. The MAE and learner running time were then respectively plotted versus the leaf size.

* 1. Result

Overfitting happens when using DTL with a smaller leaf size (leaf size < 5, **Figure 4a-b**), while it does not happen when using RTL (**Figure 4c-d**).

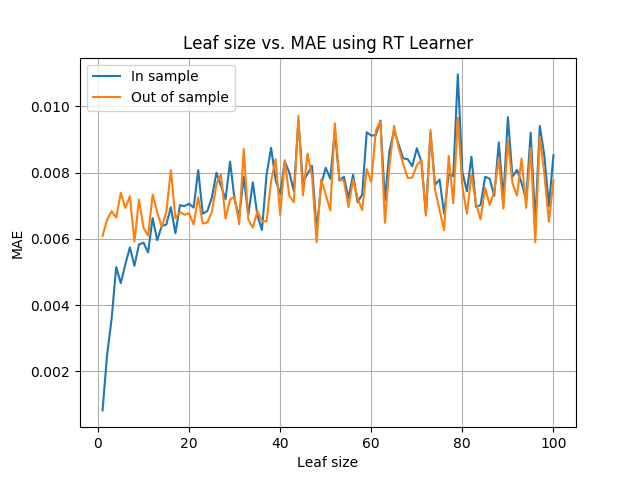
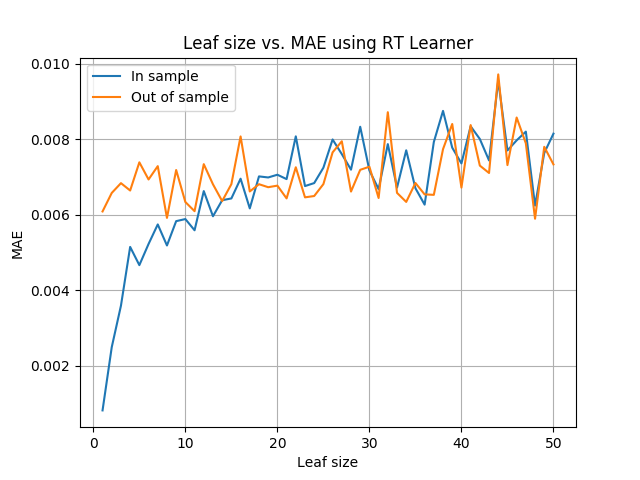
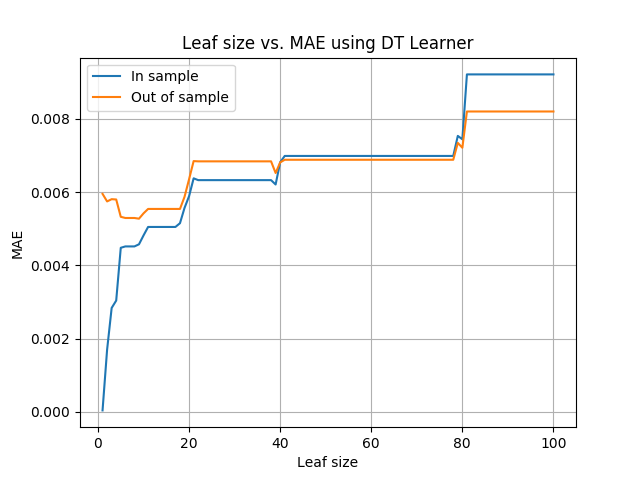
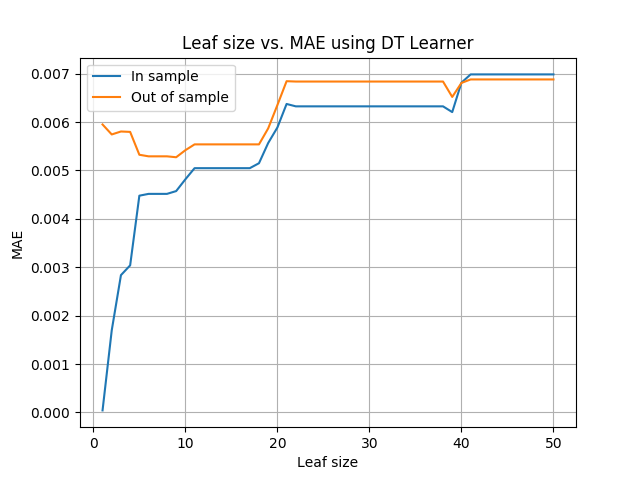
The MAE value generated by DTL is smaller compared with the ones generated by RTL. For example, at leaf size = 50, the in- and out-of-sample MAE generated by DTL are about 0.007, and the MAE generated by RTL are about 0.008. Moreover, the MAE generated by the RTL is more unstable (noisy) compared with their DTL counterparts.

As shown in **Figure 5**, for the same data set and all leaf sizes, RTL requires less running time compared with DTL. Especially, when the leaf size is small (<20 in this experiment), the time efficiency advantage of the RTL is more predominant. For example, DTL need 0.05 seconds to perform one run, but the time need by RTL is less than 0.01 seconds.

* 1. Discussion

Similar to the DTL, larger leaf size in the RTL will sacrifice accuracy (result in a higher MAE value), for both the in- and out-of-sample data. Compared with the DTL, using the same set of data, the RTL nearly always result a (about 10 -20%) higher MAE, or, is less accurate. Moreover, the MAE calculated from RTL predictions are less stable, which is originated from the internal randomness of the RTL.

Sacrificing the accuracy and prediction stability, RTL obtained an advantage of not overfitting, and performs better on running time compared with DTL. This advantage would be especially beneficial when performing machine learning over big data. However, in the cases where the data source is limited, the accuracy advantage of DTL will grant it more preference compared with RTL.



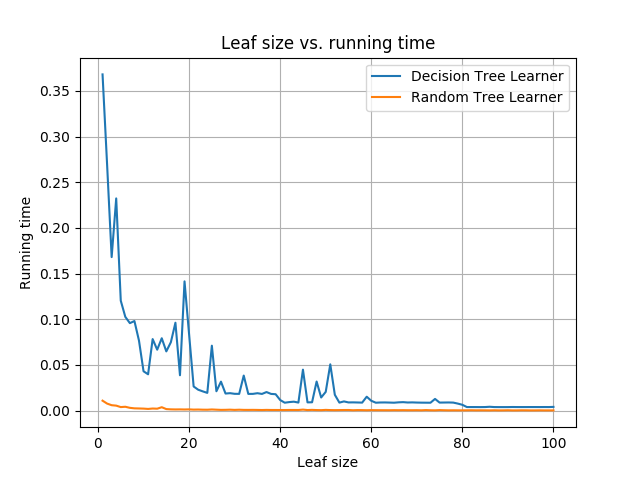
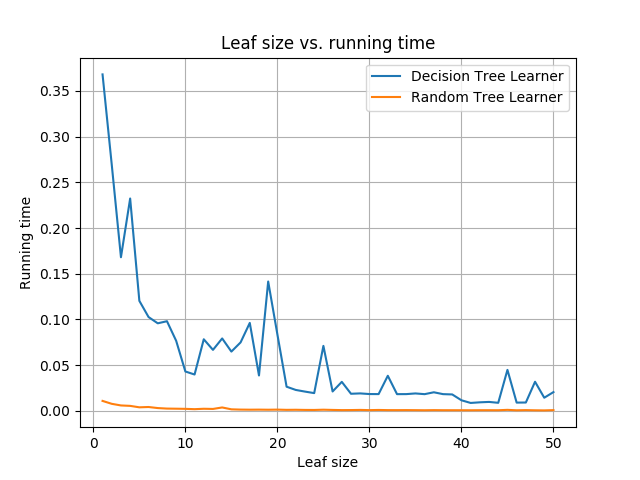
a

c

d

b

**Figure 4:** Plotting in-sample and out-of-sample MAE values versus the leaf size using DTL and RTL. (a). leaf size between 1 and 50 using DTL. (b). leaf size between 1 and 100 using DTL. (c). leaf size between 1 and 50 using RTL. (d). leaf size between 1 and 100 using RTL.



a

b

**Figure 5:** Plotting learner running time versus the leaf size using DTL and RTL. (a). leaf size between 1 and 50 (b). leaf size between 1 and 100.

* 1. Conclusion

Both RTL and DTL have their advantages. RTL is advantageous for it does not overfit, and quicker than DTL in running time. DTL is advantageous as it grants a more accurate prediction compared RTL (using same dataset and leaf size). The data analyzer need to apply them depend on other requirements (e.g. size of the dataset, running time requirement, leaf size requirement, and accuracy requirement).