ITAI Question 1

Anonymous

Student ID: 2215127

***Abstract*—This report will explore the predictions of multiple machine learning algorithms for hourly electronic power generation by power generating companies based on different environments. A suitable performance matrix will be selected to help choose the optimal hyperparameters. The optimal trained model will be predicted for unseen data.**

I. Introduction

In order to find the minimum prediction error, four algorithmic methods will be used. These are ridge regression, decision tree regression, kernel ridge regression, and support vector regression. In addition, polynomial feature processing is used to process the original data features and the processed data is fed into the model for training. The effective hyperparameters for ridge regression are alpha. The hyperparameters for decision tree regression are depth, and the hyperparameters for kernel ridge regression and support vector regression are both alpha and beta. The advantages and disadvantages of the treated versus untreated models are compared.

For this dataset, regression is used to predict an exact value, so for this dataset and our purposes, using a regression algorithm is the most appropriate choice. For the prediction of an accurate value, the mean squared error(MSE) is used as the performance matrix to assess the performance of the model. Based on the mean squared error, the values of the hyperparameters are adjusted to bring the current model to the best possible state. Ultimately the best model is applied to the unseen test set data.

II. Method

In this section, five types of regression models are described. These are the baseline model, the ridge regression model, the decision tree regression model, the kernel ridge regression model and the support vector regression model. It explains how to select hyperparameters, the differences between hyperparameters, the performance matrix for measuring hyperparameters, and which hyperparameters are best and why.

The selection of parameters for the following models is concluded from cross-validation only, and the models are then applied to the test set after the optimal model has been obtained.

A. Baseline model

In general, more complex models will likely output better results. To better show the superiority of complex algorithms, a simple regression algorithm is chosen as a baseline. ”Sklearn.Dummy”, a regression algorithm specifically designed to provide a baseline in Sklearn, is based on a simple strategy that does not focus on the content of the training set, but rather on the simple ”mean”, ”median”, ”quantile”. ”constant” are used to predict unknown data. In my implementation, all four strategies are applied, and the one with the lowest mean square error is used as the baseline. The mean squared errors for validation of the baseline algorithms for the four strategies in the current dataset are shown in Figure 1. The train dataset MSE show as table 1.

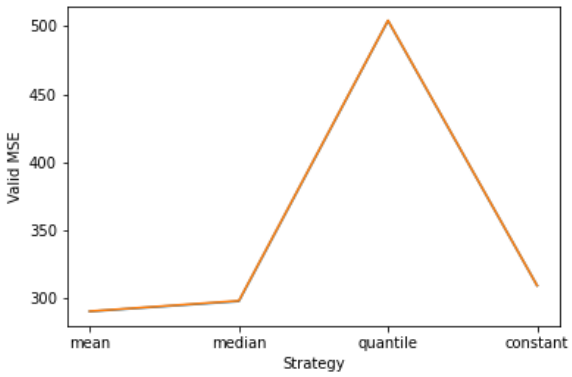

Fig. 1. Valid MSE of baseline model(Sklearn.Dummy) in different strategies.

TABLE I
Train and Validation data MSE

	mean	median	quantile	constant
Train dataset	290.171673	297.676031	504.160785	309.194418
Valid dataset	290.321175	298.025777	504.335722	309.194390

The output shows that the ”mean” strategy is the best strategy for the current data set. However, the MSE of the results for all four strategies is very large. The best ’mean’ strategy still has an MSE of 290 and a difference of 17 between the prediction error and the true value(RMSE), which is a very unacceptable amount of error.

B. Ridge Regression model

Ridge regression is a method based on and improved from ordinary least squares. This method improves on the ordinary least squares method by using Lagrange multipliers with unit diagonal matrices to process the least squares matrix, thereby reducing excessive diagonal values. Therefore, the hyperparameters of ridge regression are lambda (Lagrangian multipliers). Ridge regression performs better than ordinary

least squares when dealing with data with a large number of multicollinearity. Cross-validation is always used to ensure model robustness when dealing with data related to multicollinearity. In addition, to improve the differentiation of the data features, polynomial features will be applied to the original data. This is defined in a function that will be controlled by a variable that turns on the polynomial feature abstraction to extract the original data.

After a large number of cross-validations, the mean square error results for the hyperparameters and the validation set are shown in Figure 2. The intersection of the green vertically oriented line with the validation set is the location of the smallest mean square error, and the number above the intersection is the coordinate of the point.

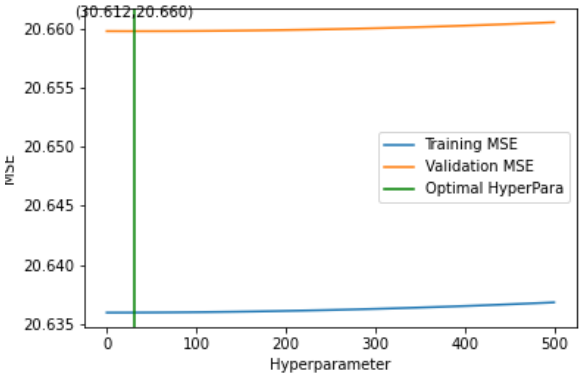

Fig. 2. Valid & Train dataset MSE by hyperparameters.

The optimal hyperparameter value is 30.612 and the minimum MSE is 20.660. This is an acceptable value of error.

When the data is fed into Ridge regression after polynomial feature extraction has been applied, the warning LinAlgWarning: Ill-conditioned matrix is prompted. But the MSE of the validation set has still been reduced and the improvement is significant. the MSE is shown in Figure 3.

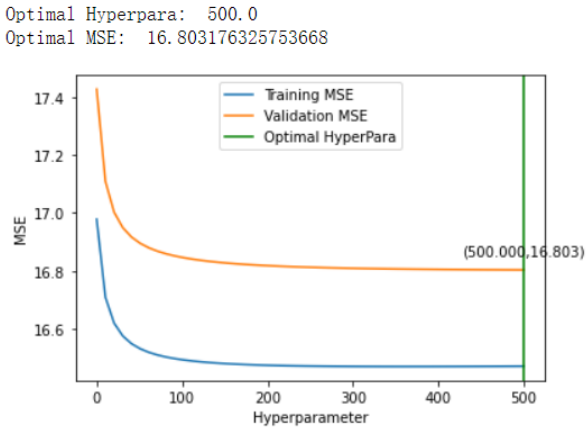

Fig. 3. Valid & Train dataset MSE after PF process

As the figure shown, the optimal MSE just 16.8, which hava a great improvement than before. However, the results are not necessarily reliable, as the predictions are based on ill-conditions

C. Decision Tree Regression model

Decision tree regression divides the feature data into multiple spaces (hyperplane partitioning) based on the original data and divides the feature data into spaces [1]. The depth of the decision tree determines how detailed the feature data is partitioned; the deeper the decision tree, the smaller each data space will be and the more likely it is to be over-fitted.

In fact, the decision tree can only generate data that is also available in the training set. If there is data in the test set that has not been seen in the training set, then the decision tree will return a value that is closest to the true value. The nature of decision tree regression is classification, so it is more accurate for non-linear, monotonically variable, non-time series and more irregular data. It does not work well for data with a stable trend and a single monotonicity. For decision trees, polynomial feature extraction may be a better option for improving accuracy. Polynomial features will allow the original data features to be abstracted to multiple dimensions and will benefit the hyperplane partitioning effect of the decision tree. When the degree parameter of polynomial feature extraction is set to 6, the depth of the decision tree is considered as a hyperparameter. The MSE of the prediction results according to the hyperparameter change is shown in Figure 4.

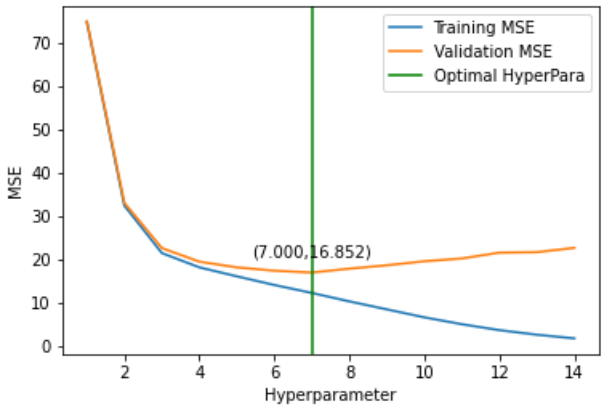

Fig. 4. Valid & Train dataset MSE base on Decision Tree depth

From the results of the validation set, when the polynomial feature has degree=6, the optimal result is obtained when the depth of the decision tree is equal to 7, MSE=16.85.

D. Kernel Ridge Regression model

The main difference between kernel ridge regression and linear ridge regression is that kernel ridge regression has a built-in kernel function matrix. This matrix will make the kernel function more sensitive and more accurate than linear

ridge regression. When the amount of data is small, kernel ridge regression will be less computationally intensive than linear ridge regression, but when the amount of data is large, it will be very computationally intensive [2]. When the kernel function matrix is "rbf", the alpha is in a very small range and the gamma is also in a small range, a good result will be obtained. The results of the cross-validation are shown in Figure 5. Where gamma = 1e-5. The polynomial feature process closed.

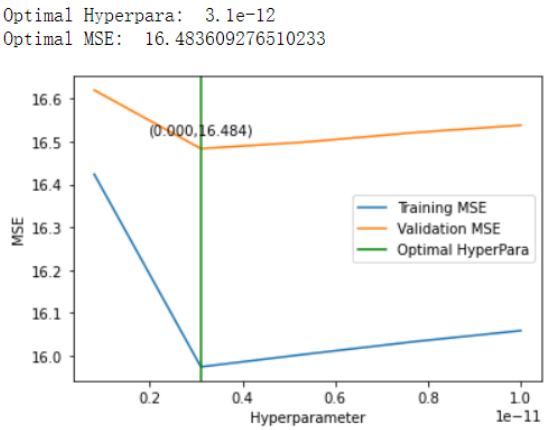

Fig. 5. Valid & Train dataset MSE base on Kernel Ridge Regression

From Figure 5, the optimal mean square error is obtained when the alpha is at its minimum. However, in terms of results, 8e-10 gives the best results, but when the value of alpha is taken to a smaller value, there is a warning and the computation time increases significantly. Thus when Gamma = 1e-5, alpha = 3.1e-12, kernel = 'rbf', the MSE is 16.483.

E. Support Vector Regression model

Support vector machines have the same classification principle as SVMs [3]. Based on the classification interval band of SVM, in SVR, no loss is calculated if the data is in the interval band, and loss is calculated when the difference between the predicted value and the true value is greater than the tolerance deviation. Also, SVR optimises the model by maximising the width of the interval band and minimising the loss function.

Optimisation objectives for SVR,

$$min_{w,b} = \frac{1}{2}\|W\|^2 \tag{1}$$

Loss function for SVR,

$$\sum_{i=1}^{m} l_\epsilon(f(x_i), y_i) \tag{2}$$

The final questions for the SVR :

$$min_{w,b} = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} l_\epsilon(f(x_i), y_i) \tag{3}$$

Sklearn can be called between SVRs, but in fact the SVRs take a very long time to compute, and the results are not very impressive. The results are shown in Figure 6.

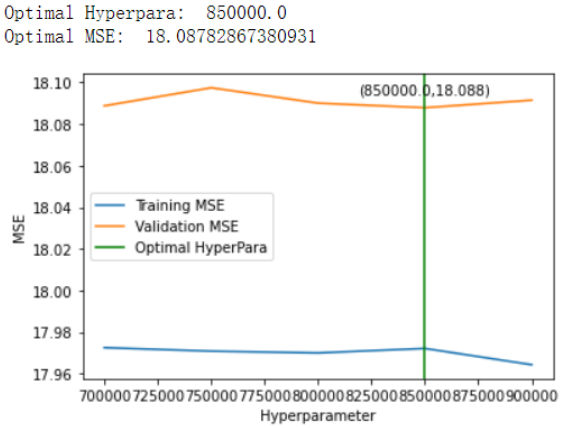

Fig. 6. Valid & Train dataset MSE base on SVR

From the results we can see that the minimum MSE is 18.08 for C=8.5e4 and gamma='scale'. the validation set of the SVR does not have the smallest MSE, but it has the longest computation time.

F. Conclusions of models

In the above tests, a total of five models were used. The baseline model is a very simple model and the strategy for the optimal extreme model is the mean strategy, but the mean squared error is still very large.

The hyperparameter chosen for the ridge regression is alpha. alpha is the only hyperparameter in the ridge regression and the only parameter that distinguishes the difference between ordinary least squares and ridge regression [4]. From the training results, the value of the optimal hyperparameter alpha does not differ much from the MSE when alpha = 0. Therefore for this model, ordinary least squares is also a good choice.

The hyperparameter of decision tree regression is depth. The depth of the decision tree determines how many hyperplanes (leaf nodes) the algorithm will divide the training set into. A proper depth means a proper number of hyperplanes, and when the decision tree is able to accurately divide the predicted data into different planes (branches) when making predictions, then the prediction will be accurate. Of all the algorithms, combining the accuracy of the validation set with the time taken to train the model (which will be shown later in the RESULT), decision tree regression is the best algorithm for the current data set.

Kernel ridge regression has three optional hyperparameters, alpha, gamma, and kernel, and the current training set is too large for kernel ridge regression to be computationally intensive. Therefore, alpha is used as the most important hyperparameter in the project, and the performance of the model is improved by adjusting the value of alpha. In fact, if the gridSearch method is used to train the alpha, gamma, and kernel iteratively, the time cost will be much greater than the value of the performance gain. Therefore, finding the optimal solution locally is the best option.

The optional hyperparameters for support vector regression are C, gamma, kernel. the situation for SVR is similar to that of KRR in that the time cost of training the data is too high. For support vector regression C is the most important hyperparameter, and the best MSE value is found by varying C.

For these four complex models, even the simplest ridge regression model has an MSE of only 20.66, which is an error of only +/- 4.5 for a true value of about 450. The smallest MSE is from KRR with an MSE of only 16.48, followed by decision tree regression with an MSE = 16.85.

III. Results

In the previous section, the optimal model for each algorithm has been determined by the selection of hyperparameters. Applying the optimal model to the test set data will give the MSE of the test set, for which both the time spent training the model and the amount of MSE optimisation should be used as a measure of a model's performance. In the following section, I will discuss the validation set MSE, the test set MSE, and the time spent on each model.

Referring to Table 2, the first column of the table shows the validation MSE of each model, the second column shows the MSE of the model on the test set, and the third column shows the time spent on the tuning process for each model. The fourth column is the number of different hyperparameters tried in the search for the optimal hyperparameter.

TABLE II
VALIDMSE, TEST MSE, TIME ELAPSED, HYPERPARA NUMBER.

	Valid_MSEs	Test_MSEs	Time_elapsed(s)	Hy-para num
Base	290.321175	295.513986	0.020219	4.0
RR	16.803176	18.144375	3.136228	50.0
SVR	18.087829	18.969162	509.484329	5.0
KRR	16.483609	17.783594	86.645421	5.0
DTR	16.934197	17.943310	47.478333	15.0

For the current dataset, the MSE of the predictions from the test set of the baseline model is very high. The MSEs for the other models were much lower than the baseline model. The MSE for ridge regression on the validation set was 16.8 and the result on the verification set was 18.144, with the training process taking only 3.13s with 50 hyperparameters tested. this is a very good result.

For the SVR, the SVR performs the worst on the current dataset, with the highest MSE on both the validation and test sets in addition to the baseline. And with a time spent of 509s for just five training sessions, the performance is very poor in terms of time spent.

For KRR, both the validation and test sets performed very well, with an MSE of 16.48 for the validation set and 17.78 for the test set, but the model took slightly longer to train five times. The time required to train the model five times was 86.6s, but overall, the performance of KRR was very good.

For DTR, the validation set MSE is only 16.93 and the test set MSE is only 17.94. The performance of DTR in terms of MSE is similar to that of KRR. However, the DTR model takes only 47.5s to train 15 times, which is much better than KRR, and the computational effort of KRR increases exponentially as the amount of data increases. Therefore DTR is a very good performance model.

For all four models, the difference between the MSE on the validation set and the MSE on the test set is very small. This indicates that the performance of the models is highly robust. Also, it shows that this way of selecting parameters and the performance matrix is very effective in the face of unseen data. Combining these four models, decision trees and ridge regression are the optimal choices for the current data and problem processing.

IV. References

References

[1] Microstrong. "Regression tree". zhihu.com. https://zhuanlan.zhihu.com/p/82054400.

[2] N. Paradigm. "Support Vector Machine Regression for Machine Learning (Machine Learning Techniques)". CSDN.net. https://blog.csdn.net/qq_34993631/article/details/79367175.

[3] S. Sayad. "Support Vector Machine - Regression (SVR)". saedsayad.com. http://www.saedsayad.com/support_vector_machine_reg.htm.

[4] M. Taboga. "Ridge regression". statlect.com. https://www.statlect.com/fundamentals-of-statistics/ridge-regression