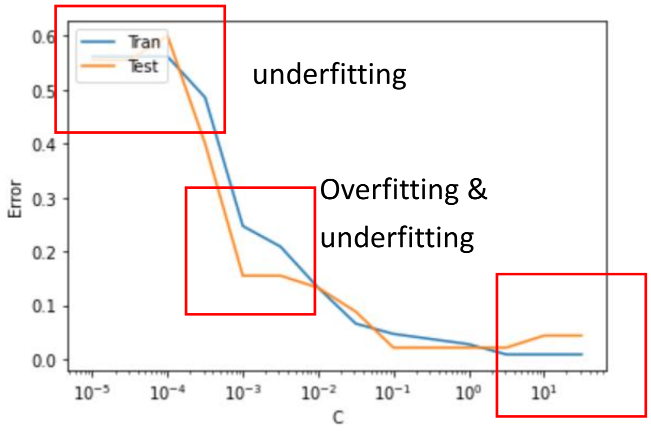
Logistic Regression:

Iris dataset:

I changed the max_iter from 100 (default) to 1000 in order to allow larger C



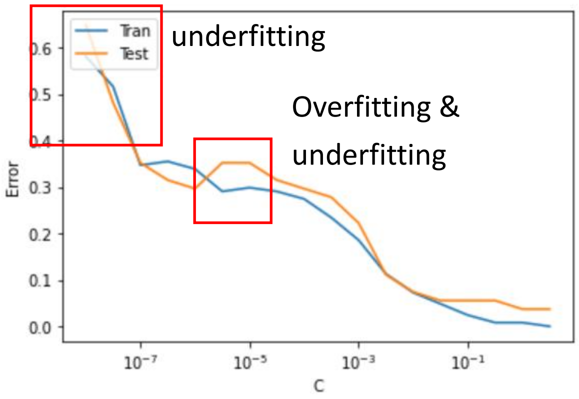
underfitting

Overfitting & underfitting

overfitting

Wine dataset

I changed the max_iter from 100 (default) to 4000 in order to allow larger C

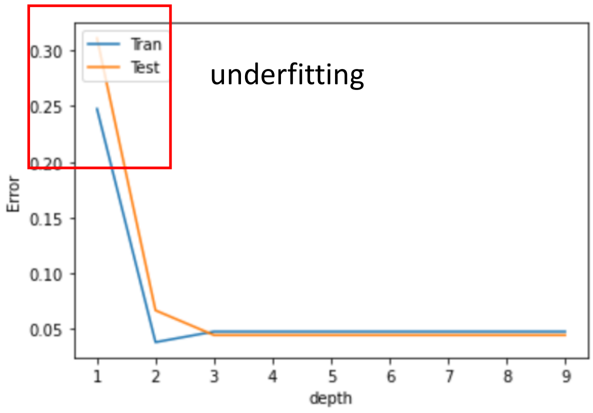


underfitting

Overfitting & underfitting

Random Forest:

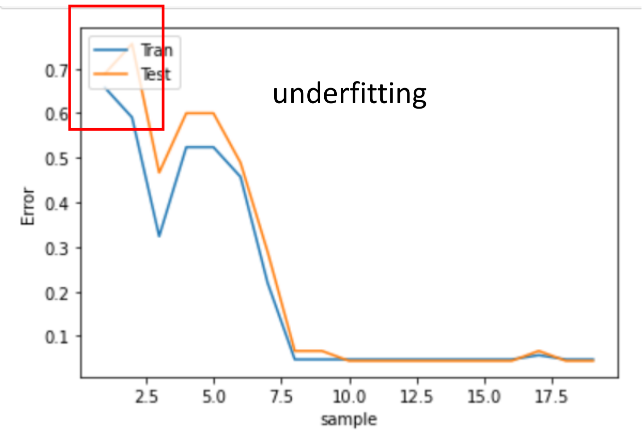
Iris dataset: Tunning max_depth from 1 to 10 with max_sample = 10



underfitting

When max_depth decreased to 1, the training and testing error increased to over 20%

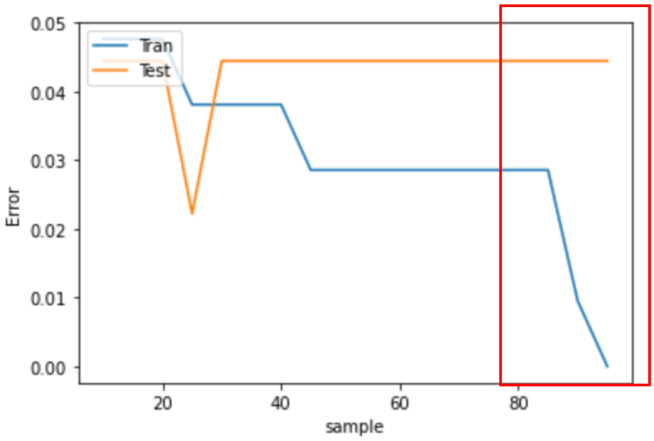
Iris dataset: Tunning max_sample from 1 to 20 with max_depth = 10 (underfitting)



underfitting

when max_sample < 6, both training error and testing error are large

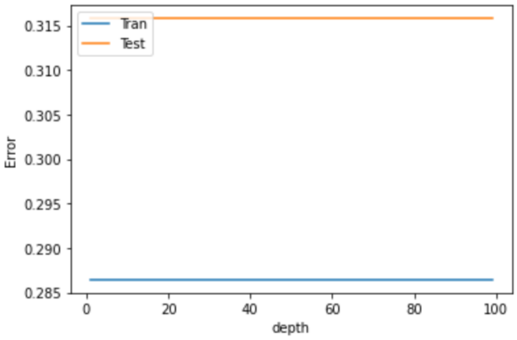
Iris dataset: Tunning max_sample from 10 to 100 (overfitting)



overfitting

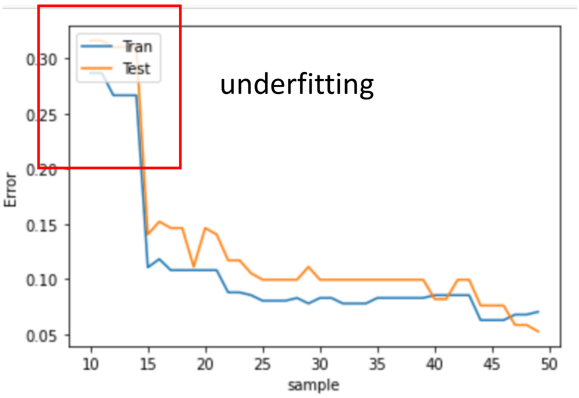
When max_sample greater than 80, the training error drop to near 0, while the testing error remain above 4%

Breast Cancer dataset: tuning max_depth from 1 to 100 with max_sample = 10



training and testing error maintain large, the model is too simple, which is underfitting

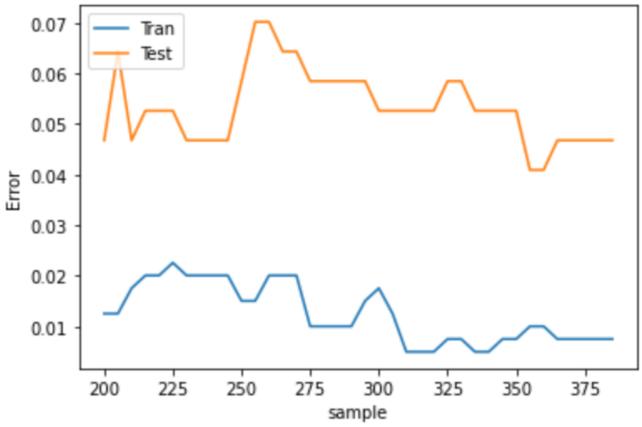
Breast Cancer dataset: tuning max_sample (underfitting)



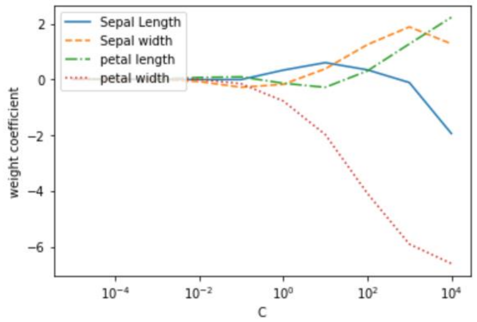
underfitting

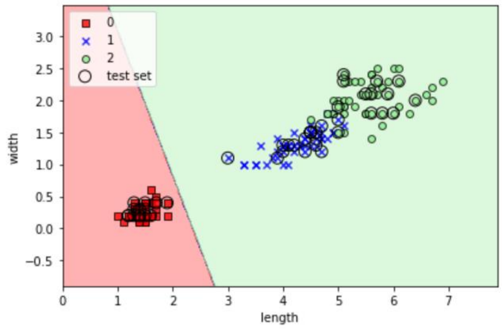
When the max_sample below 15, training error and testing error increase dramatically

Breast Cancer dataset: tuning max_sample (overfitting)



key reasons behind underfitting/overfitting:
for logistic regression, with smaller inverse regularization parameter C, the regulation strength is stronger, therefore the model is becoming less complex where each weight coefficients are similar and trend to be 0, there will be underfitting since the model cannot capture the pattern in the training data





With larger C, the weights varies to classifie all training set, when C is too high, the variance between testing error and training error will be high, as the model try to fit well with all training data but loss on generalize, which the testing error is much higher than training error, resulting in overfitting

For random forest, the max_depth decide the depth of decision tree, the deeper the decision tree, the more complex the decision boundary becomes, from the iris dataset, we can see that when the depth is decreased to 1, the testing error and training error increase a lot, showing under fitting as the model is too simple. While having large max_depth and keep increasing it, in the two dataset, the error remain constant.

For the max_sample, the smaller the max_sample, the higher randomness of the random forest, when max_sample is too small, there will exist underfitting as the result is highly random, when the max_sample is too large, the model fit well in training data with almost 0 error but higher testing error as the randomness is low.