6 Problem 6

Part a: Theoretical Explanation

- The total and within variance for the tree is given by the formula:

$$Var_{\Theta;\boldsymbol{Z}}T(x;\Theta(\boldsymbol{Z})) = Var_{\boldsymbol{Z}}E_{\Theta|\boldsymbol{Z}}T(x;\Theta(\boldsymbol{Z})) + E_{\boldsymbol{Z}}Var_{\Theta|\boldsymbol{Z}}T(x;\Theta(\boldsymbol{Z}))$$

$$\text{Total Variance} = Var_{\boldsymbol{Z}}\hat{f}_{rf}(x) + \text{within-Z Variance}$$

- The following demonstrations are based on a simulation model

$$Y = \frac{1}{\sqrt{50}}\sum_{j=1}^{50} X_j + \epsilon,$$

  with all the X_j and ϵ iid Gaussian. We use 500 training sets of size 100, and a test set of size 600.

- For getting the horizontal line on in the right figure I picked randomly a random number of trees (say n) from each trained random forest on a particular, and then I picked n random indices of trees from this random forest, computed the bias squared for them and then averaged over the number of those trees and then added this number to an array and then I repeadted this process for the X test i and then finally averaged the results !!!! which is very similar to sampling from the distribution Z and theta Trees and predicting for each X test i. We could have done it the other way by just saving these models in an array of array and pick randomly for each test point X i. And this can be generalized to many scnearios in the code that instead of saving the models and after that sampling fro the distribution of Θ and Z, we can randomly at each iteration the size of how many trees we want, which for large number of iterations is somewhat the same as sampling from the distribution of Θ and Z.

- In the text it wasn't specified how many trees we should grow in each random forest, so I picked a standard number of trees and that is 100.

- For the correlation between trees i.e. figure 15.9 we have used the formula: $\rho(x)$ is the *sampling* correlation between any pair of trees used in the averaging:

$$\rho(x) = \text{corr}[T(x;\Theta_i(\mathbf{Z})), T(x;\Theta_j(\mathbf{Z}))]$$

  where $\Theta_1(\mathbf{Z})$ and $\Theta_2(\mathbf{Z})$ are a randomly drawn pair of random forest trees grown to the randomly sampled $\mathbf{Z}$;

- We can see a general trend that as m increases the correlation between pair of trees increases as we can expects since the trees are more similar to each other.

- The Bias formula is given by:

$$\begin{aligned}\text{Bias}(x) &= \mu(x) - E_Z \hat{f}_{\text{rf}}(x) \\ &= \mu(x) - E_Z E_{\Theta|Z} T(x; \Theta(\mathbf{Z})).\end{aligned}$$

- But the thing is with the figure on the left it is not working correctly I revised the code so many times and I still don't know where is the mistake. I honestly think that I am calculating the total variance correctly but it is not the case.
- In figure 15.10 the left one we can see that the within variance is in the right shape and on the right scale what is off is the total variance.

Part b: Graphs and Plots

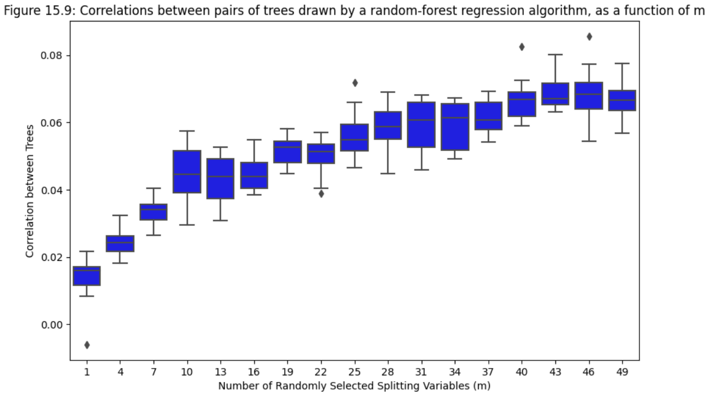

Figure 15.9: The correlation between trees in a random forest, as a function of the number of trees.

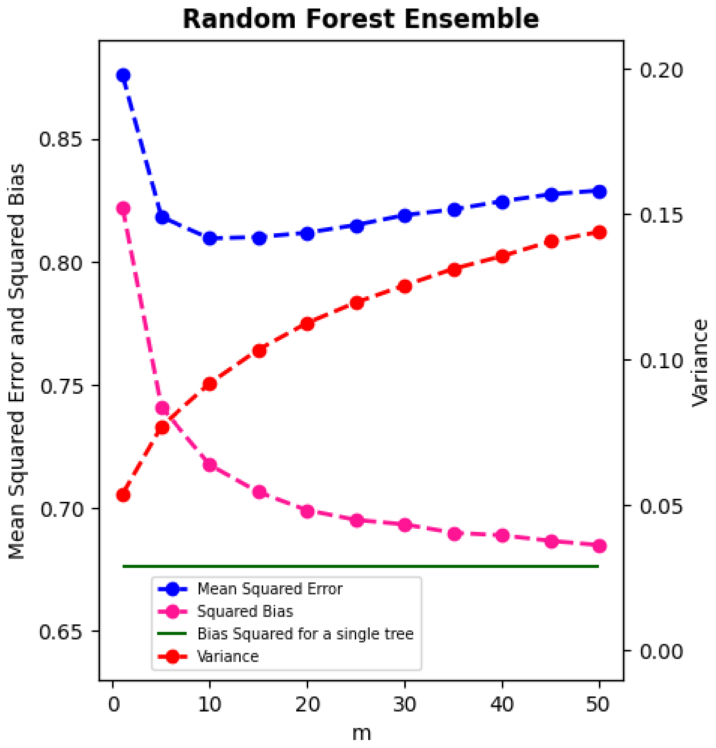

Figure 15.10 right: Variance, Bias and MSE for the random forest model for different number of trees in the forest.

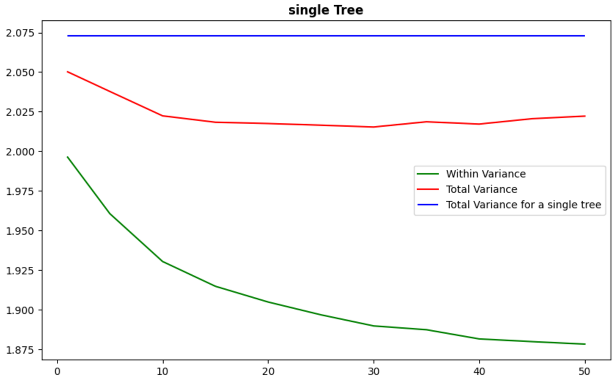

Figure 15.10 left: Variance for one Tree.