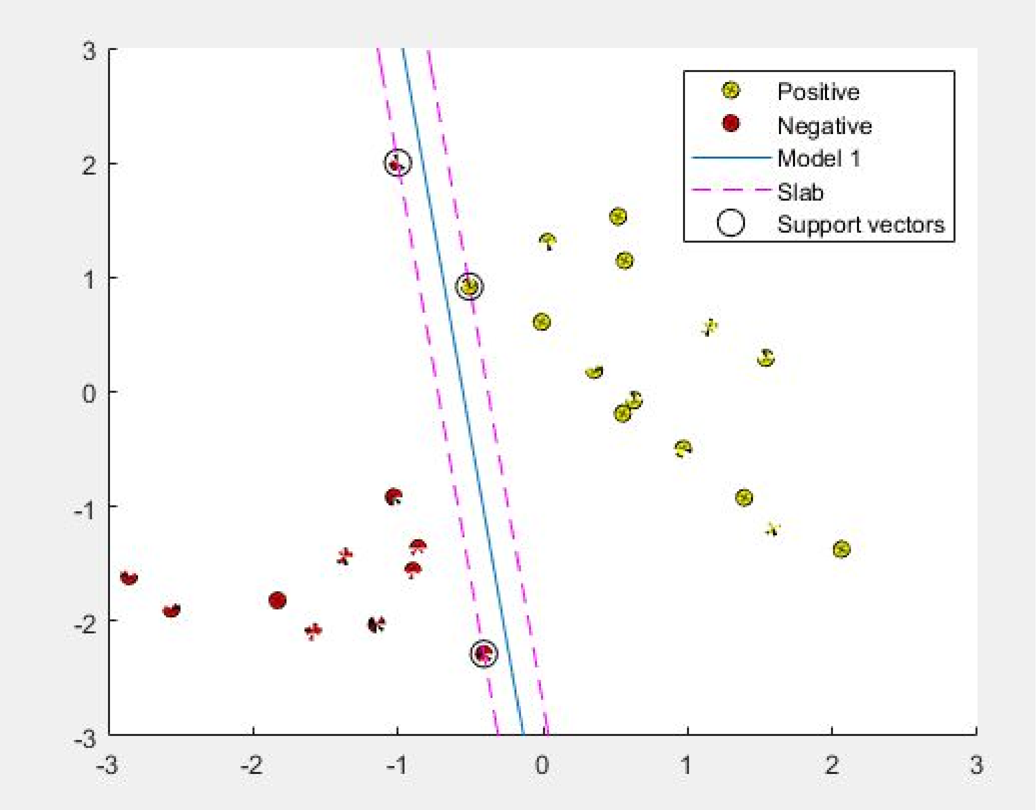
1.4 Yes, the result showing that deeper tree is worse than the original tree just as i expected. Because even though by growing a deeper tree we obtain the decision boundary which fits the training data 100% and has no training error, it is obviously overfitting the data. So when we test the models using k-fold cross validation the original one is better in most cases .

2. In order to import titanic.csv I had to exclude some part of data which i consider irrelevant. Which are Passenger ID, Name and Ticket because I considered they have no relation to whether the passenger die or lives. I also excluded column Cabin because most of cabin data is lost, since i couldn’t exclude all data without cabin i excluded cabin. Then I excluded all rows that doesn’t have the data I selected.

After preprocess the data, I’m left with 714 data sets.I decided to use 600 of them as the training data and rest as test data. In order to compute the best pruning alpha I first grow a tree using training data, then i used cvloss method to obtain the best pruning level which then help me get the best pruning alpha. And i grow a tree called TreeP with the best pruning alpha. In other hand to obtain best MaxNumSplit I used method called hyperparameters and from that MaxNumSplit I grow a tree.

By compare those trees I see that the tree grown with best pruning alpha is much smaller than the one grow with best MaxNumSplit.

3.b There is some problem with the plot but you can see that when add extra x and y the hyperplane changes because that point would have violate the role and landed at positive side of the hyperplane.



5. In first try I used 1000 training data with linear, gaussian and polynomial as Kernel function to train the model and the result wasn’t great, all of them predicted result with accuracy less than 10%. (stuck here for a week...)

And at the last day before deadline I realized that I was always using the first 1000 data in the whole training set and so no matter what I did i couldn’t get a better result, also validation set is too big so the training is very slow(I have 10000 and I changed it to 100 now it’s better). Then I got some help from my classmate. Seem that in order to get better results at the accuracy. I have use all the training data and use randomly picked BoxConstrain and KernelScale from the range of 0.001-1000(with linespace(0.001,1000,100)) recombined them in every possible combinations. So 10000 pairs in total. But using 10000 pairs of BoxConstrain and KernelScale are too time consuming so i had to first use small set of training data to select some pairs and increase the training data size and repeat. So first I only used first 1-100 training data, then 100 to 1000, 1000 to 10000, 10000 to 50000 etc. Sadly the resulting accuracy rate is only 11% so I increased to linespace(0.001,1000,1000). I don’t expect this to finish in time since i only have 40 min left.

So i will finish the report just here. And do as much of the task and send it in.