Extra Credit Report

* For ripple.csv

I plot bag number ~ rmse with bagging method when boost is used and boost is not used.



We can see that boost actually appear to increasingly overfit with fluctuation as number of bags increases, while bagging without boost do not appear to overfit as bag number increases. Yet some fluctuation does happen.

Conclusion: overfit does not happen when boost is not applied, while overfit happens when boost is applied for ripple.csv.

* For 3\_groups.csv

I plot bag number ~ rmse with bagging method when boost is used and boost is not used.



We can see that the rmse with boosting applied fluctuate dramatically as number of bags increases, while the rmse without boosting applied continue to decrease as number of bags increases.

Conclusion:

Overfitting does not happen for 3\_groups either when boost is applied or not.

* For My\_data.csv

I plot bag number ~ rmse with bagging method when boost is used and boost is not used.



Conclusion:

We can clearly see that when boost is used, the overfit happens.

My\_data.csv generating code:

def worst4boost(length):

data=[]

for x in np.random.choice(np.array([0,1]),size=length):

data.append([x,x,x])

idx=np.random.choice(length,size=length)

print idx

data=np.array(data)

return data[idx]

I simply choose from 0 and 1 randomly.