1-5 ok (last 2 are in notebook) get to that in a sec

6-7 ok

My slides

8-13 ok

14-17 skip

18-21 intro talk about notebook

22-28 skip

29-30 obama Clinton

31-35 gini

36-38 skip

39 – 40 wrap up of trees

Bagging

Bootstrap aggregation. Trees have high variance, split in 2, will look different. A low variance model like lin reg will return similar results

Bagging general purpose method to reducing th variance of a statistical elarning method

Averaging a set of observations reduces variance >increase prediction accuracy

What if we had more datasets, would that inc or dec variance. Not practical

Bagging is: Bootstrap samples, fit model on each, then run predict on each, then average the predictions.

W DT, construct B reg trees, using B bootstrapped traing sets, avg predictions. The trees are grown deep and not pruned. So each tree has high variance. Increase accuracy, by combining hundreds or thousands of trees.

Classfication, majority vote

Each bagged tree makes use of around 2/3 of observations, rest of 1/3 is OOB observations, test set

Difficult to interpret the results, so we look at feature importance

. Record total amount that RSS is dereased due to splits averaged across all B trees, a large value indicates an important predictor. Same w GINI add up

RF decorrelate the tree, problem i

Each time a split is considered, a random sample of m predictors is chosen as split candiates from p

M = sqrt p, . so the algo is NOT EVEN ALLOWED TO CONSIDER a majority of available predictrs

Reason if there is one strong predictor it will always be chosen on top

Baggin m=p, RF is m=sqtp

Recall that bagging involves creating multiple copies of the original training

data set using the bootstrap, fitting a separate decision tree to each

copy, and then combining all of the trees in order to create a single predictive

model.

Boosting. Trees are grown sequentially. Each tree is based on infor from previous tree. Each tree is fit on a modified version of the original data set.

Bossting is a slow learner. Fit the 2nd tree on the residuals from the model. Rather than outcome Y as the target. Then add to the fitted function to update residuals.

Short trees. Slowly improve prediction in areas where it does not perform well.

Shrinkage parameter, lamda slows the process down even futher.01Learning rate. Typical values are 0.01 or 0.01

Therefore in boosting unlike bagging, the construction of each tree depends strongly on the trees created before it.

# of splits in each tree, often d=1 works well. Stump (one split)

This highlights one difference between boosting

and random forests: in boosting, because the growth of a particular tree

takes into account the other trees that have already been grown, smaller

trees are typically sufficient.

Using smaller trees can aid in interpretability

as well; for instance, using stumps leads to an additive model.

Ensemble

41-42 ok

43 skip

44- 50ok

go back to notebook

wrap up RF