The Bagging Althorithm Theorem and Realization

Jianqi Huang & Junda Chen

School Of Management and Engieering,CUFE

2022/10/24

Before Bagging

Resampling

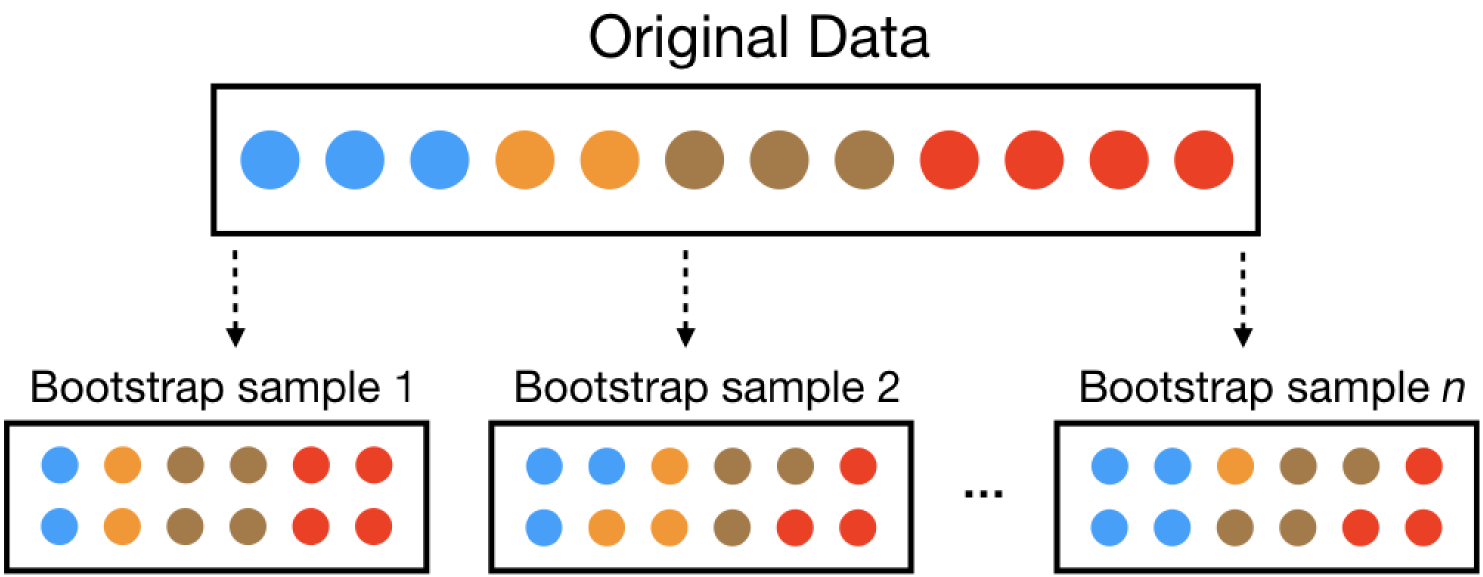
- Evaluate the learning effect from data: the generalization performance.
- Just based on data: good performance in this data but not in others.
- **Validation Approach**: splitting training set into training set and validation set(or hold set).**it always causes a positive estimate**.
- **Re-sampling methods** provide an alternative approach by allowing us to repeatedly fit a model of interest to parts of the training data and test its performance on other parts.

The Advantage of Bootstrap

- Since observations are replicated in bootstrapping, there tends to be less variability in the error measure compared with k-fold CV(cross validation).

Bootstrap

- A bootstrap sample is a random sample of the data taken with replacement
- A bootstrap sample is the same size as the original data set from which it was constructed.



Bootstrap in R

R Code Plot

```
library(rpart)
library(MASS)
data(Pima.tr) ## load data
Diabetes <- Pima.tr[,8] ## response
X <- Pima.tr[,-8] ## predictor
tree <- rpart(Diabetes ~ ., data=X,
control=rpart.control(xval=10)) ## 10-fold CV
n <- nrow(X)
subsample <- sample(1:n, n , replace=TRUE)
sort(subsample)
tree_boot <- rpart(Diabetes ~ ., data=X, subset=subsample,
control=rpart.control(xval=10)) ## 10-fold CV
```

Bagging(*B*ootstrap *Ag*gregation)

Definition

- Bootstrap aggregating (bagging) prediction models is a general method for fitting multiple versions of a prediction model and then combining (or ensembling) them into an aggregated prediction
- Bagging is a fairly straight forward algorithm in which b Bootstrap copies of the original training data are created
- New predictions are made by averaging the predictions together from the individual base learners.

$$f\widehat{(X_{bag})} = \widehat{f_1(X)} + \widehat{f_2(X)} + \cdots + \widehat{f_b(X)}$$

- the $f\widehat{(X_{bag})}$ is bagged prediction.
- The $\widehat{f_1(X)}, \widehat{f_2(X)}, \cdots, \widehat{f_b(X)}$ are the predictions from the individual base learners.
- Bagging does not always improve upon an individual base learner.
- Bagging works especially well for unstable, high variance base learners

Algorithm

Bagging Tree has the following algorithm. Let $\hat{Y}$ be a tree(or other predictor) based on samples $(X_1, Y_1), \cdots, (X_n, Y_n)$

- Draw indices $(j_1, \cdots, j_n)$ from the set $\{1, \cdots, n\}$ with replacement. Fit the tree $\hat{Y}^*$ based on samples

$$(X_{j1}, Y_{j1}), \cdots, (X_{jn}, Y_{jn})$$

- Repeat first step B times to obtain

$$\hat{Y}^{*,1}, \cdots, \hat{Y}^{*,B}$$

- Bagged estimator is

$$\hat{Y}_{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}^{*,b}$$

The Thought in Bagging

$$\hat{Y}_{Bag} = \frac{1}{B}\sum_{i=1}^{B}\hat{Y}^{*,i}$$

- for $B \to \infty$(many bootstrap samples)

$$\overline{Y}_{Bag} \to E(\hat{Y})$$

- the aggregation of information in large diverse groups results in decisions that are often better than could have been made by any single member of the group.

Using $\widetilde{Y}_{Bag} \rightarrow E(\hat{Y})$ for $B \rightarrow \infty$

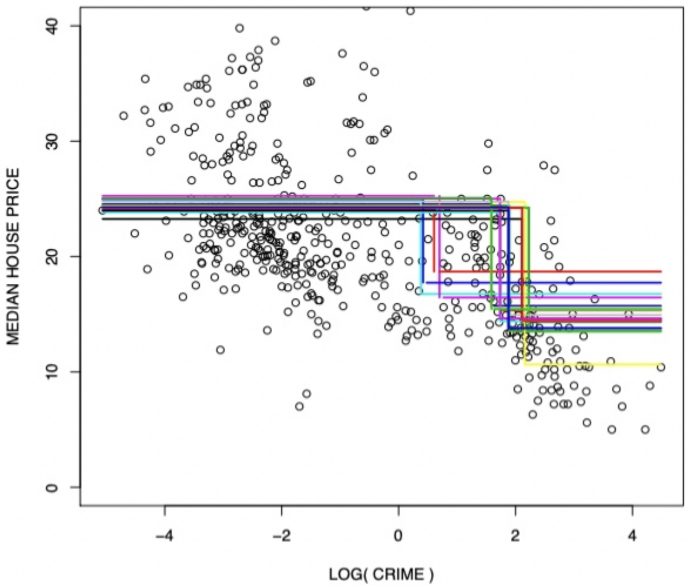
$$
\begin{aligned}
E((\hat{Y} - E[\hat{Y}])^2) &= E[(Y - \widetilde{Y}_{bag} + \widetilde{Y}_{bag} - \hat{Y})^2] \\
&= E[(Y - \widetilde{Y}_{Bag})]^2 + E[(\widetilde{Y}_{Bag} - \hat{Y})^2] \\
&\geq E((\hat{Y}_{Bag} - E[\hat{Y_{Bag}}])^2)
\end{aligned}
$$

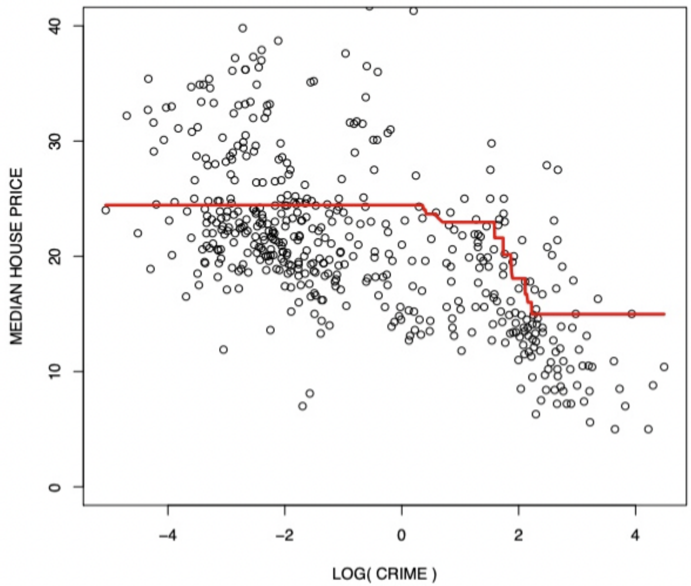
the population bagging estimatoe $\widetilde{Y}_{Bag}$ thus reduced the squared error loss by eliminating the variance of $\hat{Y}$ around its mean $E(\hat{Y})$

- For trees, this means that bagging has a very beneficial effect on trees with a large size

the comparision with different depth

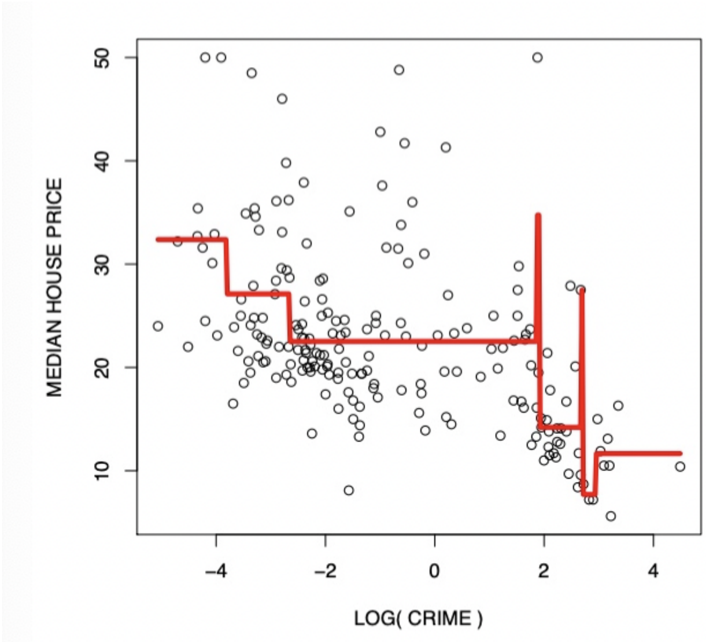
Bagged stumps $\hat{Y}^{*,b}, b = 1, 2, \cdots, 10$

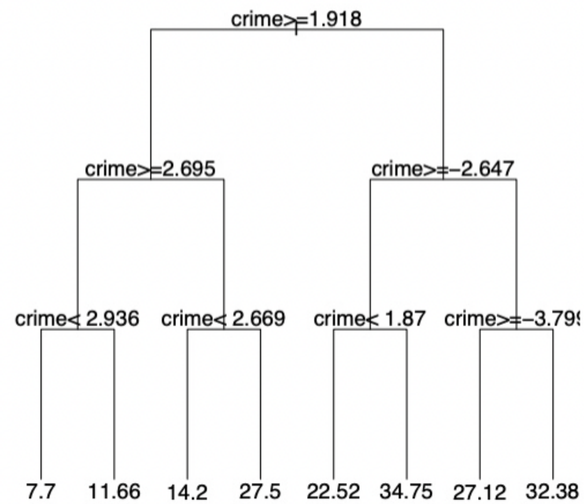




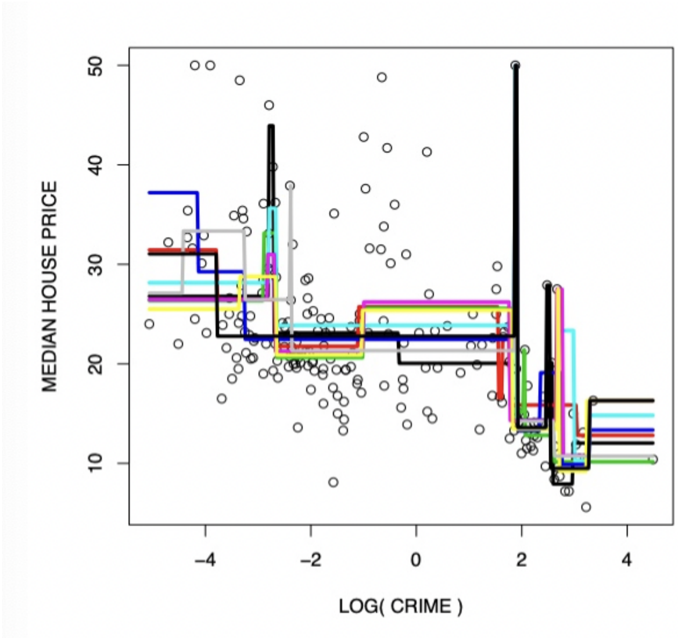
Bagging leads to a small but not a dramatic improvement.

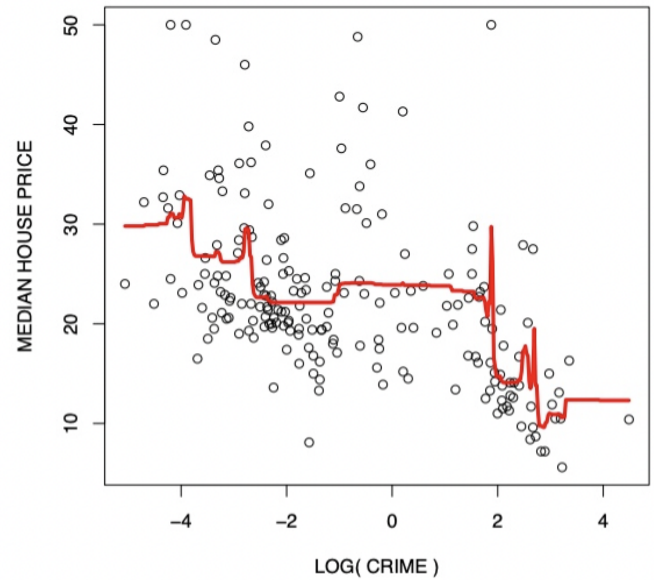
the fit with depth $d = 3$ have a poor performance





$\hat{Y}$ has a high variance(and low bias), bagging leads to a large improvement.





Out of Bag test error estimation

$$R_{test} := E(L(Y, \hat{Y}_{Bag}))$$

$$\hat{R}_{test} = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{Y}_i^{oob})$$

Bagging Realization with R

The Core Code

R Code   Plot

```
n <- nrow(Boston)
X <- Boston[,-14]
Y <- Boston[,14]
maxdepth<- 10 # plot the depth d = 3 and d = 5
tree <- rpart(Y ~.,data = X,
              control = rpart.control(maxdepth = maxdepth,minsplit =
plot(tree,margin=.1,uniform=TRUE);text(tree,cex=1.3)
```

R Code   Plot

```
B <- 100
prediction_oob <- rep(0,length(Y)) ## vector with oob predictions
numbertrees_oob <- rep(0,length(Y)) ## how many oob trees
for (b in 1:B){ ## loop over bootstrap samples
  subsample <- sample(1:n,n,replace=TRUE) ## "in-bag" samples
  outofbag <- (1:n)[-subsample] ## "out-of-bag" samples ## fit tree o
  treeboot <- rpart(Y ~ ., data=X, subset=subsample,
                    control=rpart.control(maxdepth=maxdepth,minsplit=
  ## predict on oob-samples
  prediction_oob[outofbag] <- prediction_oob[outofbag] +
    predict(treeboot, newdata=X[outofbag,])
  numbertrees_oob[outofbag] <- numbertrees_oob[outofbag] + 1
}
## final oob-prediction is average across all "out-of-bag" trees
prediction_oob <- prediction_oob / numbertrees_oob
plot(prediction_oob, Y, xlab="PREDICTED", ylab="ACTUAL")
df<-as.data.frame(cbind(prediction_oob,Y))
ggplot(data=df,aes(prediction_oob,Y))+
  geom_point(aes(prediction_oob,Y))+
  geom_smooth(method = 'lm',formula = y ~ x, se = F)
```

References

- Efron, Bradley, and Robert Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." Statistical Science. JSTOR, 54–75.

- Therneau, Terry M, Elizabeth J Atkinson, and others. 1997. "An Introduction to Recursive Partitioning Using the RPART Routines." Mayo Foundation.

All models are wrong, but some are useful。

——George E. P. Box (1987)

Thanks!