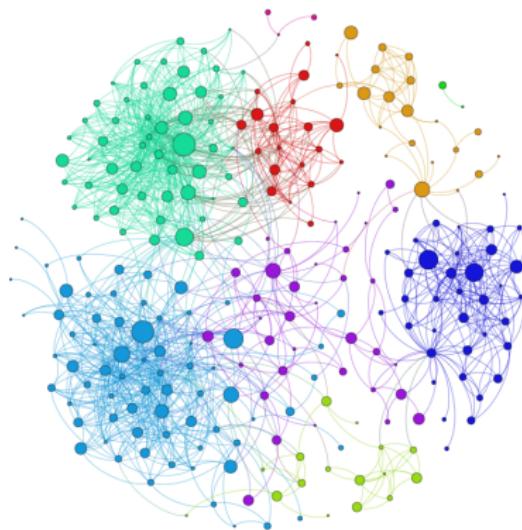


# Tree-based Methods

Jiaming Mao

Xiamen University



Copyright © 2017–2019, by Jiaming Mao

This version: Spring 2019

Contact: [jmao@xmu.edu.cn](mailto:jmao@xmu.edu.cn)

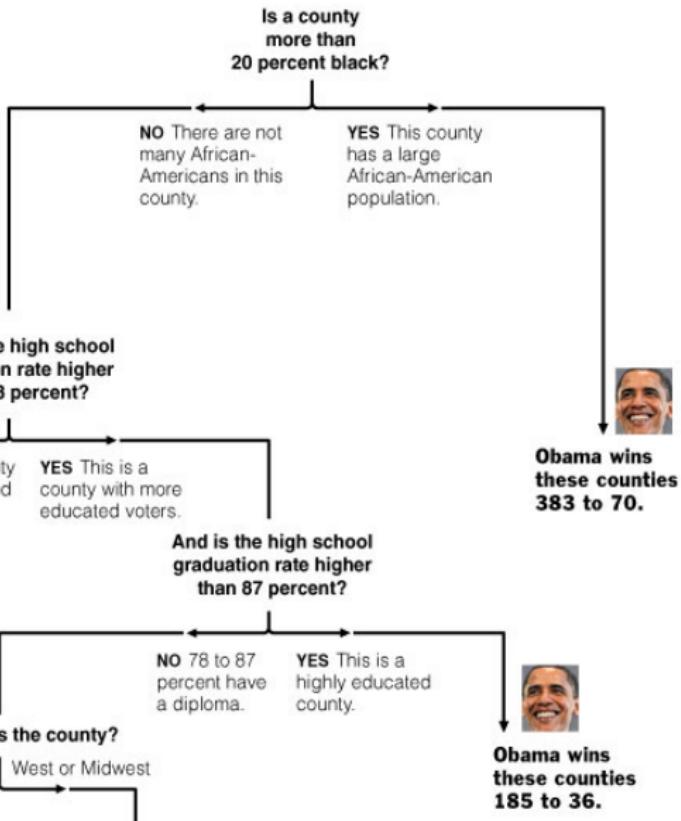
Course homepage: [jiamingmao.github.io/data-analysis](https://jiamingmao.github.io/data-analysis)



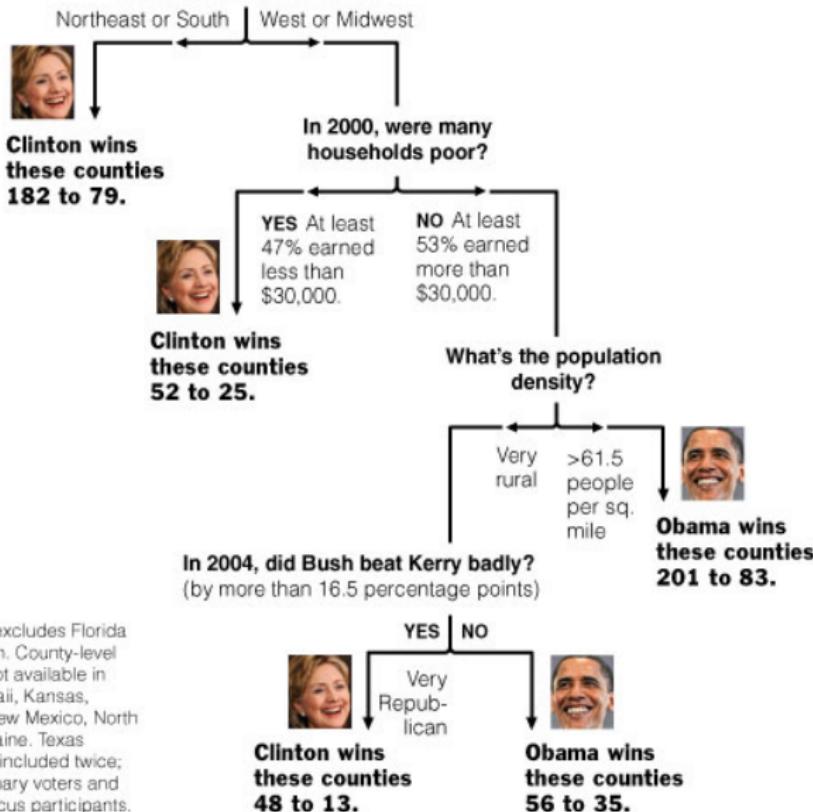
All materials are licensed under the [Creative Commons Attribution-NonCommercial 4.0 International License](#).

# Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



### And where is the county?



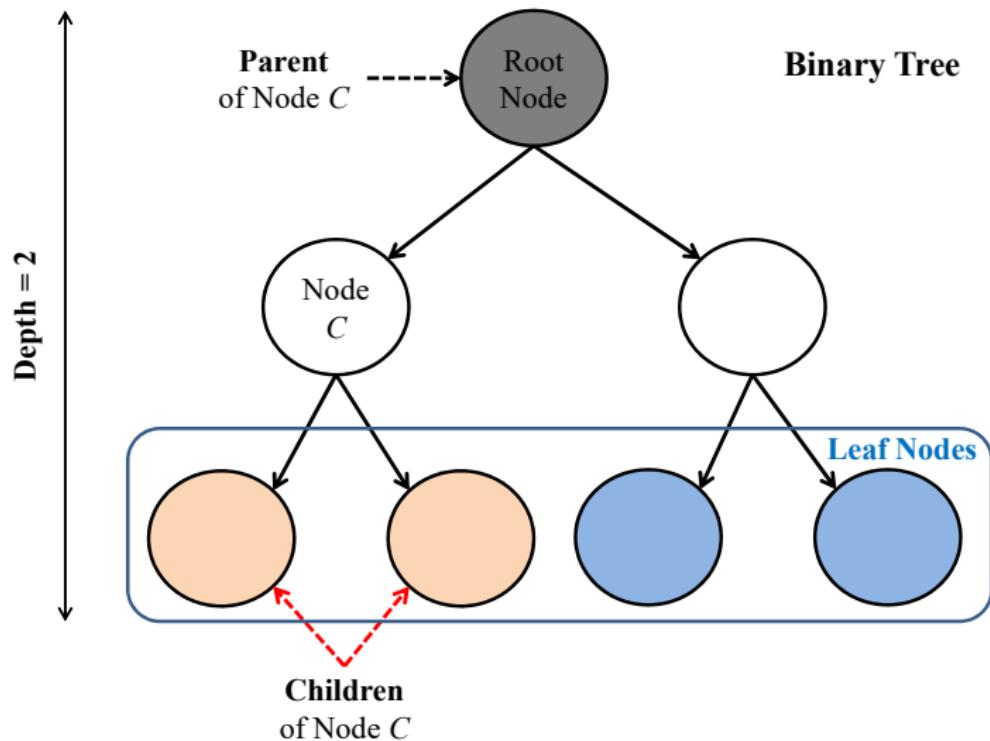
Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/  
THE NEW YORK TIMES

# Decision Trees

- Tree-based methods divide the predictor space into different regions, and then fit a model in each region.
- The predictor space can be divided by making successive binary splits on the predictor variables  $x_1, \dots, x_p$ : we choose a variable  $x_j$ ,  $j = 1, \dots, p$ , divide up the predictor space according to  $x_j \leq c$  and  $x_j > c$ , and then proceed on each half.
- A decision tree thus constructed consists of **internal nodes** and **terminal nodes (leaves)**. Each internal node is a decision to split the predictor space, and the leaves are final predictions. The **size** of a tree is the number of its leaves.
- Decision trees can be applied to both regression and classification problems.

# Decision Trees



## Regression Tree

Let  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  be the training data. A regression tree of  $M$  leaves can be thought of as dividing the predictor space into  $M$  regions  $R_1, \dots, R_M$ , each corresponding to a leaf of the tree.

For each  $R_m$ ,  $m = 1, \dots, M$ , let

$$\bar{y}_m = \frac{1}{n_m} \sum_{x_i \in R_m} y_i \quad (1)$$

, where  $n_m$  is the number of observations in  $R_m$ .

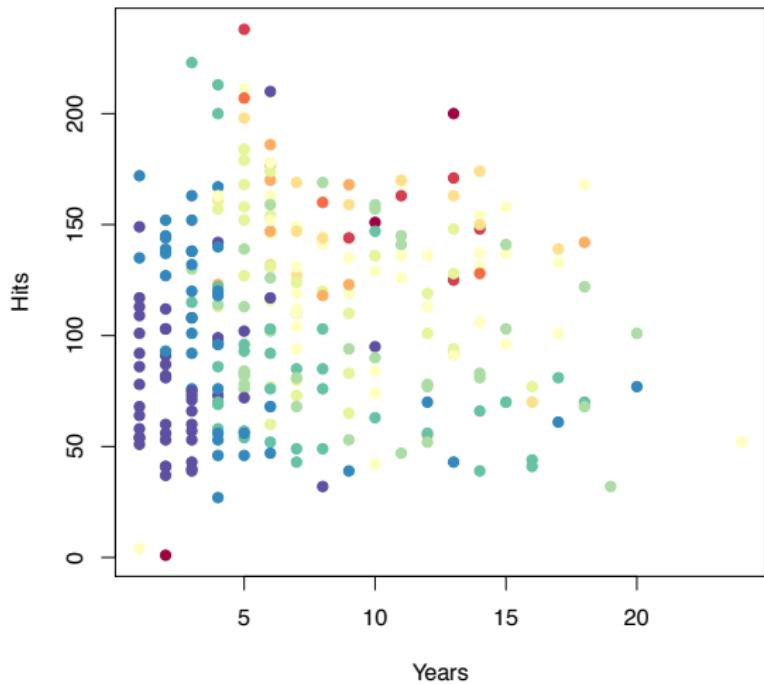
# Regression Tree

The estimated regression function is:

$$\hat{f}(x) = \sum_{m=1}^M \bar{y}_m \cdot \mathcal{I}\{x \in R_m\} \quad (2)$$

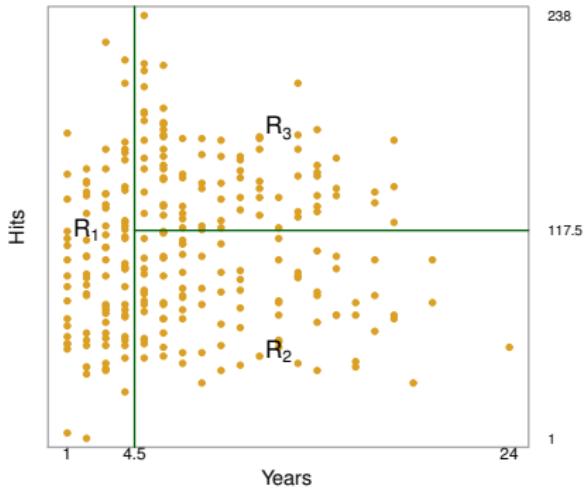
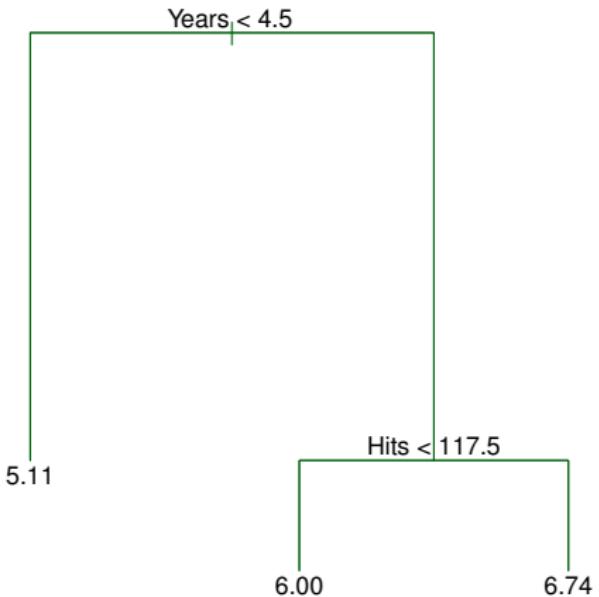
- For every observation that falls into the region  $R_m$ , we make the same prediction, which is simply the mean of the response values for the training observations in  $R_m$ .
- The result is a **piecewise constant model**.

# Baseball Salary



Color-coded from low (blue, green) to high (yellow, red)

# Baseball Salary



The results suggest that years is the most important factor in determining salary. Among players who have played for more than 4.5 years, the number of hits made (in the previous year) affects salary.

## Building a Regression Tree

- The goal is to find  $R_1, \dots, R_M$  that minimize  $E_{in}$ <sup>1</sup>. In the regression setting, using the L2 loss, we minimize the RSS:

$$\text{RSS} = \sum_{m=1}^M \sum_{x_i \in R_m} (y_i - \bar{y}_m)^2$$

- Unfortunately, it is computationally infeasible to consider every possible partition of the predictor space.

---

<sup>1</sup>Here we minimize  $E_{in}$  first and regularize it later.

# The CART Algorithm

The CART algorithm adopts a **topdown, greedy** approach and divides the predictor space by making successive **binary** splits. At each step, it chooses the split to achieve the biggest drop in RSS.

- The algorithm begins by dividing the predictor space into two regions. Then it divides one of the two regions<sup>2</sup> further into two regions ... The process continues until a stopping criterion is reached<sup>3</sup>.
- This is a **greedy** approach because at each step of the tree-building process, the optimal split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.

---

<sup>2</sup>Depending on splitting which would result in the biggest drop in RSS.

<sup>3</sup>For instance, we may continue until no region contains more than 10 observations.

# The CART Algorithm

Given a region  $R \subseteq \mathbb{R}^p$ , the optimal split is choosing an  $x_j$  and a split point  $s$ , so that the two sub-regions

$$R_{\text{left}} = \{x \in R : x_j < s\} \text{ and } R_{\text{right}} = \{x \in R : x_j \geq s\}$$

are as **homogenous** in response  $y$  as possible.

For regression trees, this means choosing  $(j, s)$  to minimize:

$$\sum_{x_i \in R_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{x_i \in R_{\text{right}}} (y_i - \bar{y}_{\text{right}})^2$$

# The CART Algorithm

The process generates a large tree  $T_0$ . We then **prune** the tree by choosing, for each value of  $\alpha \geq 0$ , a subtree  $T \subset T_0$  that minimizes

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \bar{y}_m)^2 + \alpha |T|$$

, where  $|T|$  is the size of  $T$  and  $\alpha$  is a **tuning** parameter that controls the tradeoff between the subtree's complexity and its fit to the training data.

- $\alpha = 0 \Rightarrow T = T_0$
- $\alpha \uparrow \Rightarrow |T| \downarrow$
- The optimal  $\alpha$  can be selected using cross-validation.

# The CART Algorithm

- Intuitively, a small tree might be too simple, while a large tree might overfit the data.
- The strategy of growing a large tree and then pruning it is used because any stopping rule may be short-sighted, in that a split may look bad but it may lead to a good split below it.

# The CART Algorithm

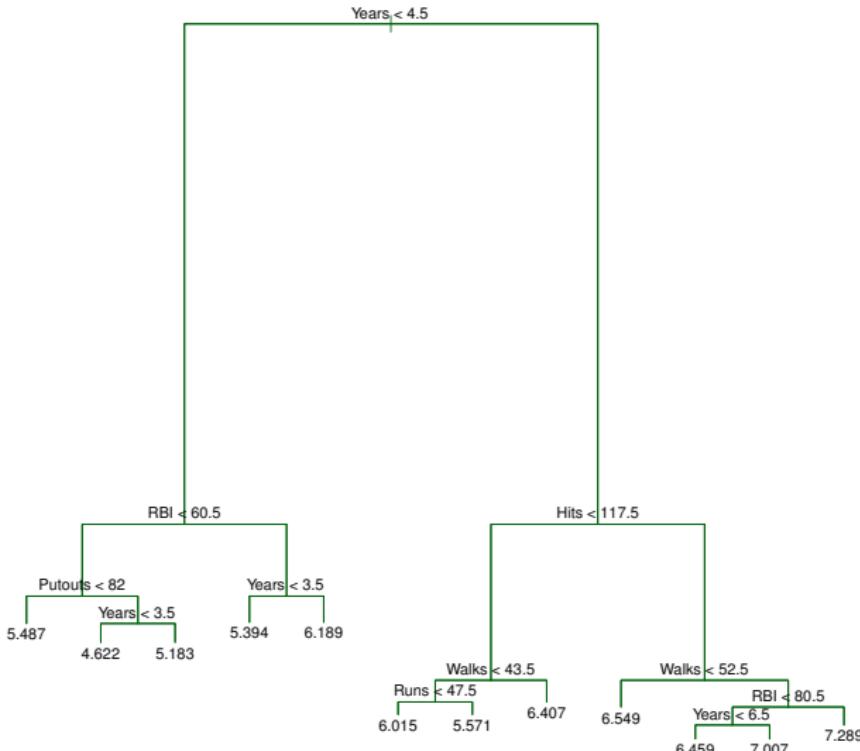
---

## Algorithm 8.1 Building a Regression Tree

---

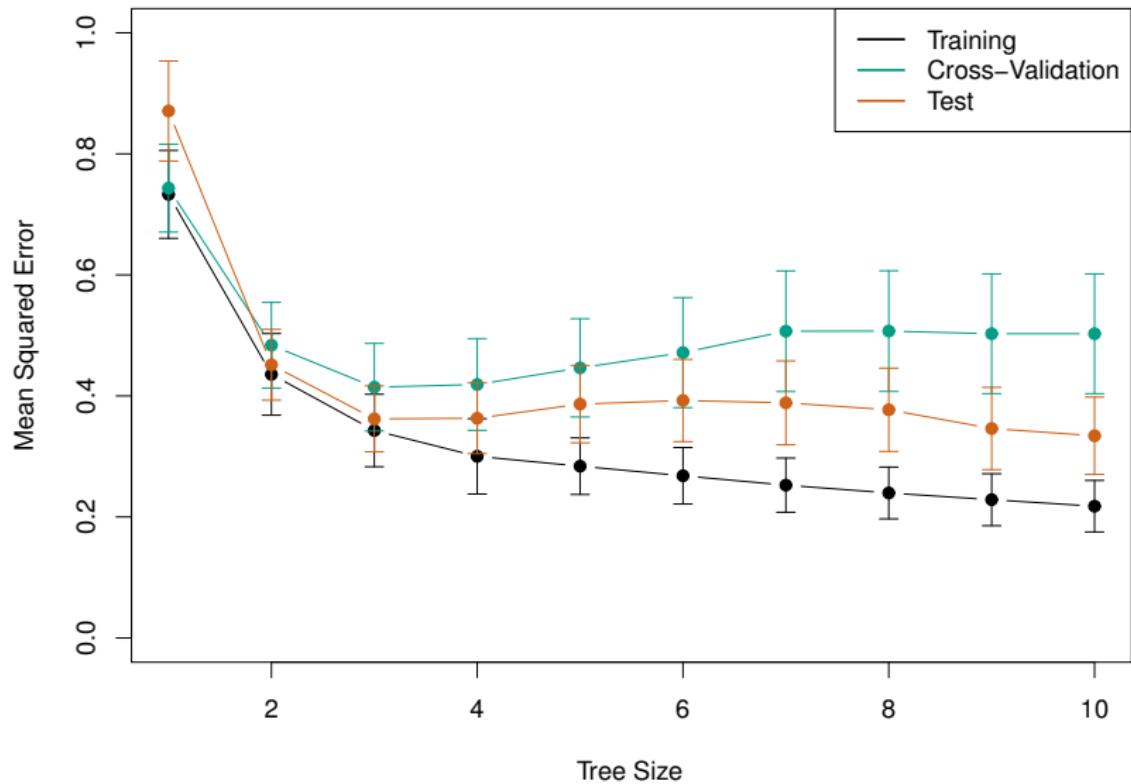
1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
  2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of  $\alpha$ .
  3. Use K-fold cross-validation to choose  $\alpha$ . That is, divide the training observations into  $K$  folds. For each  $k = 1, \dots, K$ :
    - (a) Repeat Steps 1 and 2 on all but the  $k$ th fold of the training data.
    - (b) Evaluate the mean squared prediction error on the data in the left-out  $k$ th fold, as a function of  $\alpha$ .Average the results for each value of  $\alpha$ , and pick  $\alpha$  to minimize the average error.
  4. Return the subtree from Step 2 that corresponds to the chosen value of  $\alpha$ .
-

# Baseball Salary



The unpruned tree

# Baseball Salary



# Classification Tree

For classification trees, the leaves are class predictions.

Let  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  be the training data, where  $y \in \{1, \dots, J\}$ . For each  $R_m, m = 1, \dots, M$ , we assign a class label  $c_m$ .

We then classify a new point  $x \in \mathbb{R}^p$  by

$$\hat{f}(x) = \sum_{m=1}^M c_m \cdot \mathcal{I}\{x \in R_m\} \quad (3)$$

# Classification Tree

Let  $p_j^m \equiv \Pr(y = j | x \in R_m)$  be the class probabilities in  $R_m$ .  $p_j^m$  can be estimated by

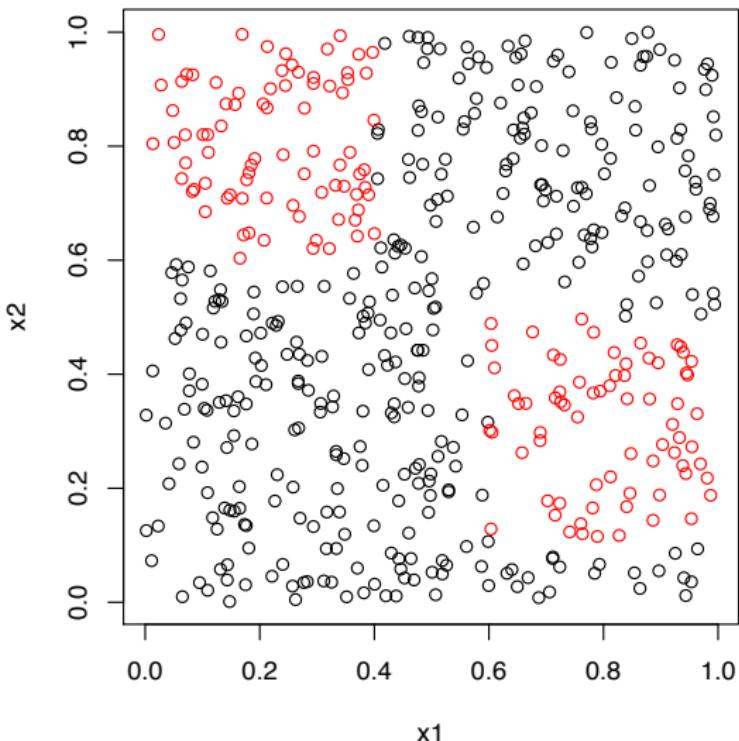
$$\hat{p}_j^m = \frac{1}{n_m} \sum_{x_i \in R_m} \mathcal{I}\{y_i = j\} \quad (4)$$

Then

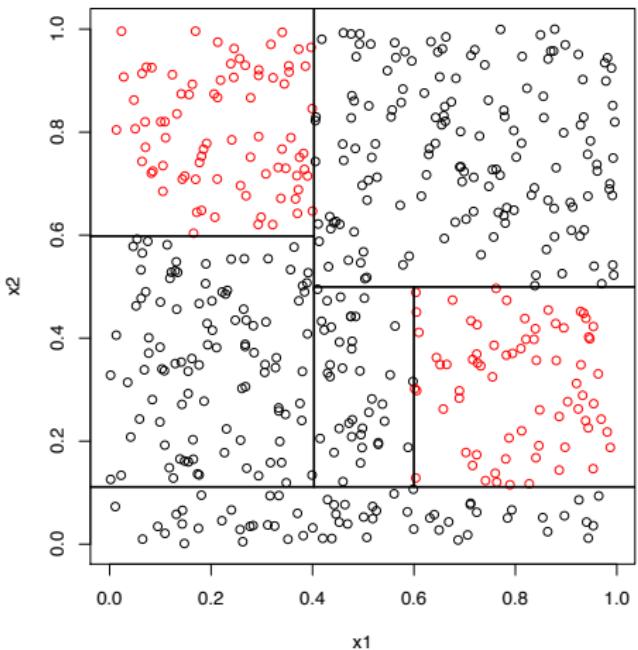
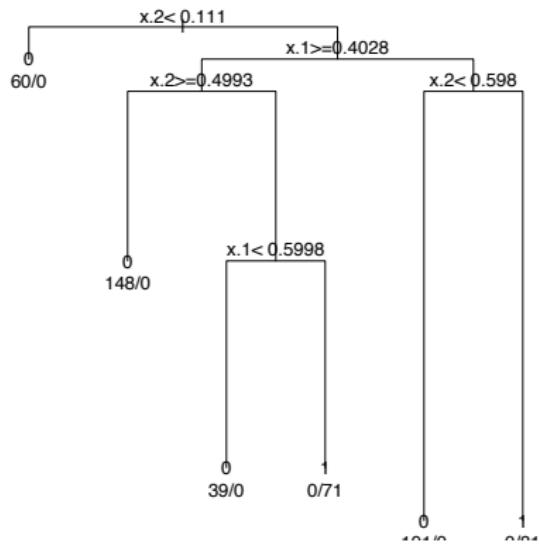
$$c_m = \arg \max_j \hat{p}_j^m \quad (5)$$

, i.e.,  $c_m$  is the most common occurring class in  $R_m$ .

# Classification Tree



# Classification Tree



## Building a Classification Tree

As in building regression trees, we first find  $R_1, \dots, R_M$  that minimize  $E_{in}$ . Let  $Q_m = E_{in}(R_m)$ . Then at each step, the CART algorithm chooses the optimal split to minimize:

$$n_{\text{left}} Q_{\text{left}} + n_{\text{right}} Q_{\text{right}}$$

For classification trees,  $Q_m$  is also called the **node impurity** measure.

# Node Impurity Measures

Commonly used node impurity measures are:

**Misclassification rate:**

$$Q_m = 1 - \hat{p}_{c_m}^m$$

, where  $\hat{p}_{c_m}^m$  is the proportion of observations in  $R_m$  that belong to class  $c_m$ .

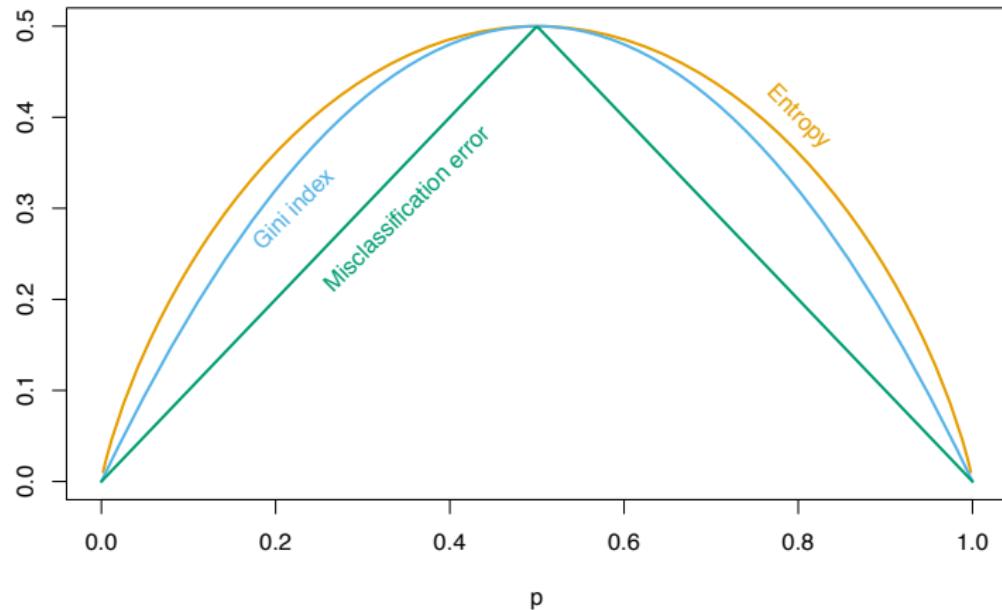
**Gini index:**

$$Q_m = \sum_{j=1}^J \hat{p}_j^m (1 - \hat{p}_j^m)$$

**Cross-entropy:**

$$Q_m = - \sum_{j=1}^J \hat{p}_j^m \log \hat{p}_j^m$$

# Node Impurity Measures



Node impurity measures for binary classification (cross-entropy has been scaled to pass through  $(0.5, 0.5)$ ). Note that the Gini index and cross-entropy are differentiable and thus easier to optimize numerically.

## NBC Shows

Data from NBC on 40 TV pilots. Data on each show include its genre, ratings, and viewer demographics. The demographics information contain over 50 variables including viewer location, household size, race, income, occupation, home ownership, etc.

```
nbc <- read.csv("nbc_showdetails.csv")
demo <- read.csv("nbc_demographics.csv")
```

# NBC Shows

There are three genres: {Drama/Adventure, Reality, Situation Comedy}.  
To predict genre based on demographics:

```
#####
# Multinomial Logistic Regression #
#####

require(nnet)
require(AER)
y <- nbc$Genre
demo <- demo[,-1]
logitfit <- multinom(y~.,data=demo)
result <- coeftest(logitfit)
```

# NBC Shows

```
result[1:5,]

##                                     Estimate Std. Error   z value Pr(>|z|) 
## Reality:(Intercept)           0.0143     5.85  0.002447  0.998
## Reality:TERRITORY.EAST.CENTRAL 6.6040  2235.83  0.002954  0.998
## Reality:TERRITORY.NORTHEAST    1.5089  1404.82  0.001074  0.999
## Reality:TERRITORY.PACIFIC      -2.1562  1776.01 -0.001214  0.999
## Reality:TERRITORY.SOUTHEAST     0.6641  1910.97  0.000348  1.000
```

None of the coefficient estimates are significant ...

```
result[result[,4]<=0.05,]

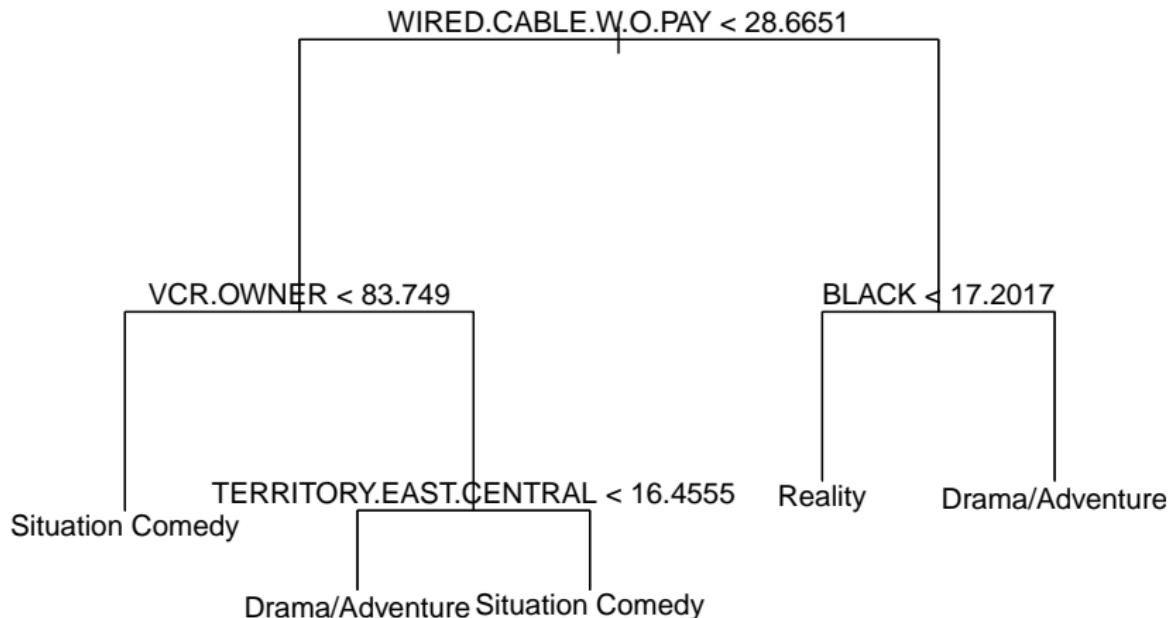
##          Estimate Std. Error z value Pr(>|z|)
```

# NBC Shows

```
#####
# Classification Tree #
#####
require(tree)
t0 <- tree(y~.,data=demo,mincut=1) #mincut: minimum size for child nodes
summary(t0)

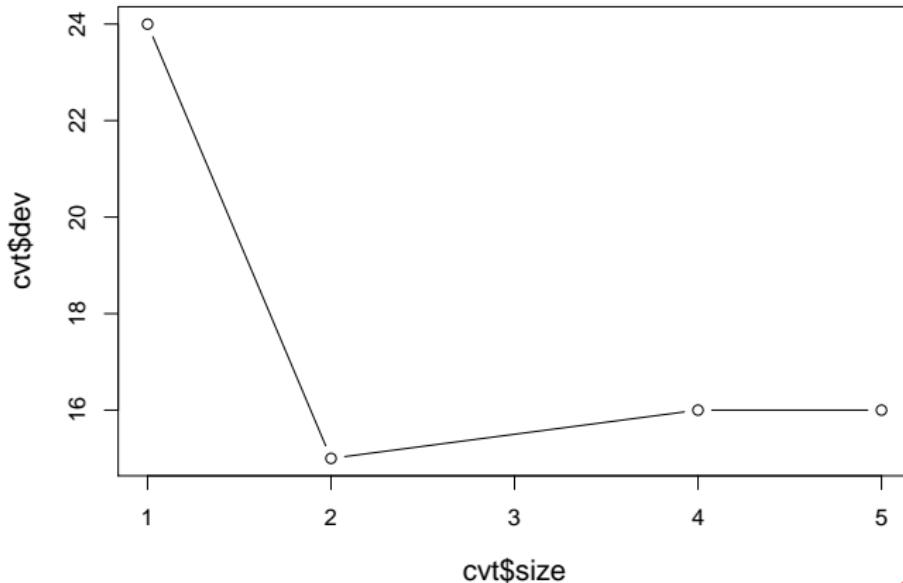
##
## Classification tree:
## tree(formula = y ~ ., data = demo, mincut = 1)
## Variables actually used in tree construction:
## [1] "WIRED.CABLE.W.O.PAY"      "VCR.OWNER"
## [3] "TERRITORY.EAST.CENTRAL"  "BLACK"
## Number of terminal nodes:  5
## Residual mean deviance:  0.192 = 6.73 / 35
## Misclassification error rate: 0.05 = 2 / 40
```

# NBC Shows

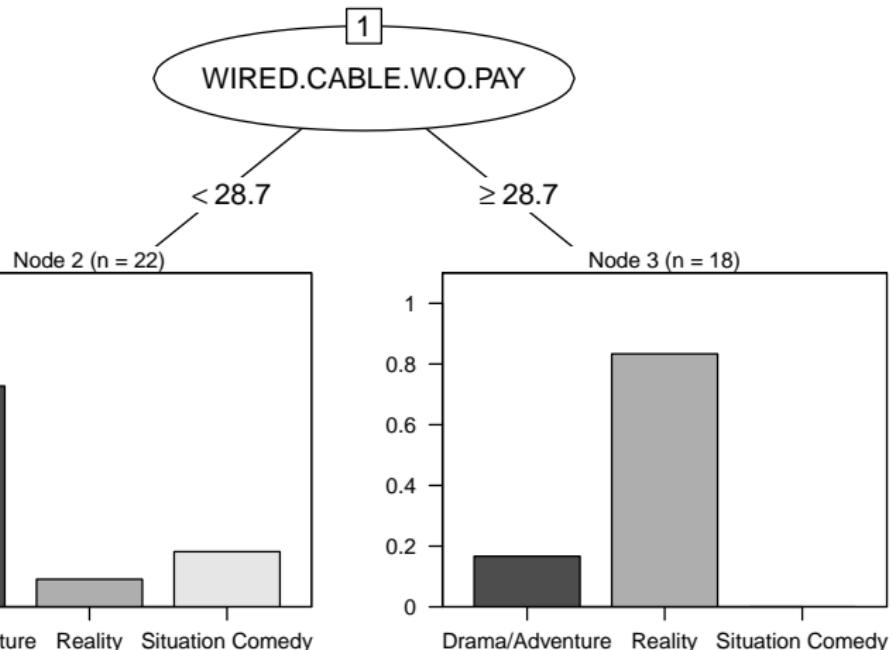


# NBC Shows

```
# To prune the tree, use CV to select the best tuning parameter
cvt <- cv.tree(t0, FUN=prune.misclass) # use misclassification rate
# for pruning
t1 = prune.misclass(t0, best=cvt$size[which.min(cvt$dev)])
```



# NBC Shows



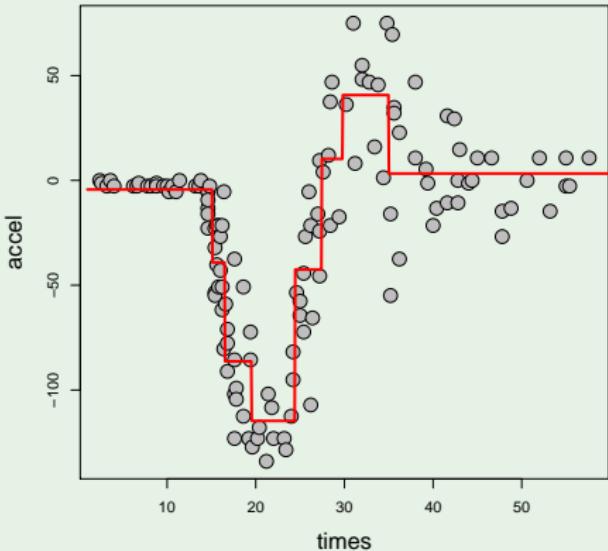
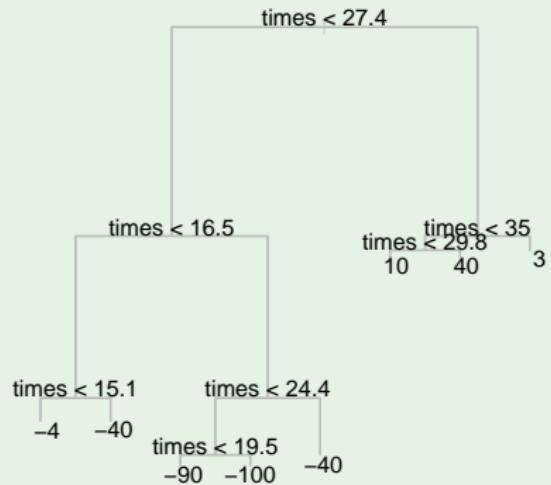
# Advantages and Disadvantages of Trees

## Pros:

- Highly interpretable
- Automatically detecting nonlinear relationships
- Automatically modeling interactions

# Advantages and Disadvantages of Trees

## Motorcycle crash test dummy data



$x$  : time from impact;  $y$  : head acceleration

# Advantages and Disadvantages of Trees

## Cons:

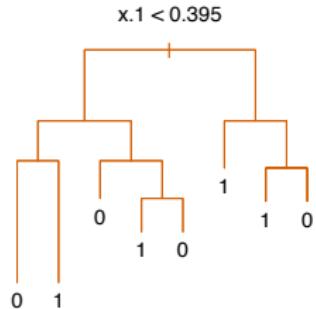
- Relatively poor predictive performance
  - ▶ Trees generally suffer from high variance and are **unstable**: a small change in the observed data often results in a completely different tree.
  - ▶ This is due to their hierarchical nature: once a split is made, all the splits under it are changed as well.
- Difficulty in capturing simple relationships
  - ▶ Trees require a large number of parameters (splits) to capture simple linear and additive relationships.
  - ▶ In other words, the weaknesses of tree-based methods are the strengths of linear models, and vice versa.

# Bagging

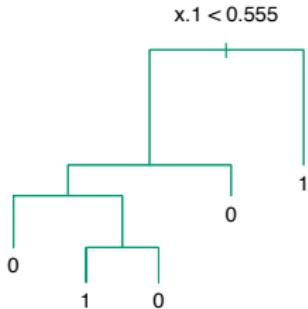
- **Bootstrap aggregation**, or **bagging**, is a general-purpose procedure for reducing the variance of a statistical learning method.
- We know that given a set of  $N$  independent random variables each with variance  $\sigma^2$ , the variance of the mean is  $\frac{\sigma^2}{N}$ .
- Hence if we can average the predictions made on many independent training data sets drawn from the population, then we can reduce the variance and hence increase the prediction accuracy of our method.
- In practice, we do not have access to multiple independent training data sets from the population. Instead, we can draw multiple samples from the **empirical distribution** using bootstrap.

# Bagging

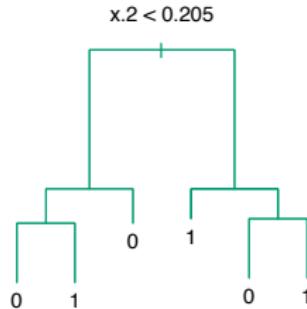
**Original Tree**



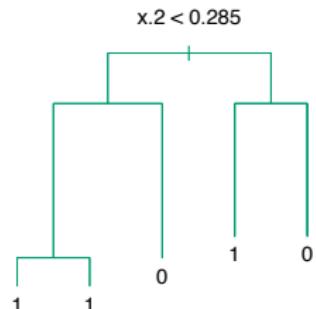
**$b = 1$**



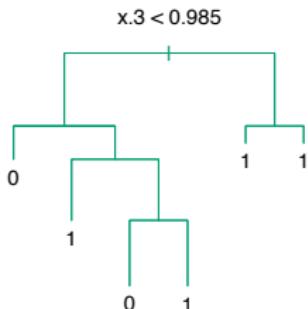
**$b = 2$**



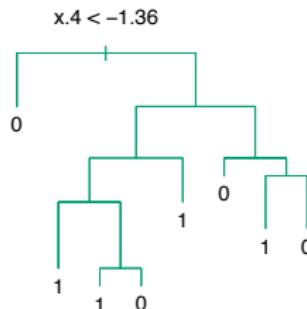
**$b = 3$**



**$b = 4$**



**$b = 5$**



## Bagging for Regression

Let  $b = 1, \dots, B$  index the bootstrap samples drawn from the training data. We fit our model on each bootstrap sample.

For regression problems, let  $\hat{f}^{(b)}(x)$  be the model's prediction trained on sample  $b$ . We then average all the predictions to obtain

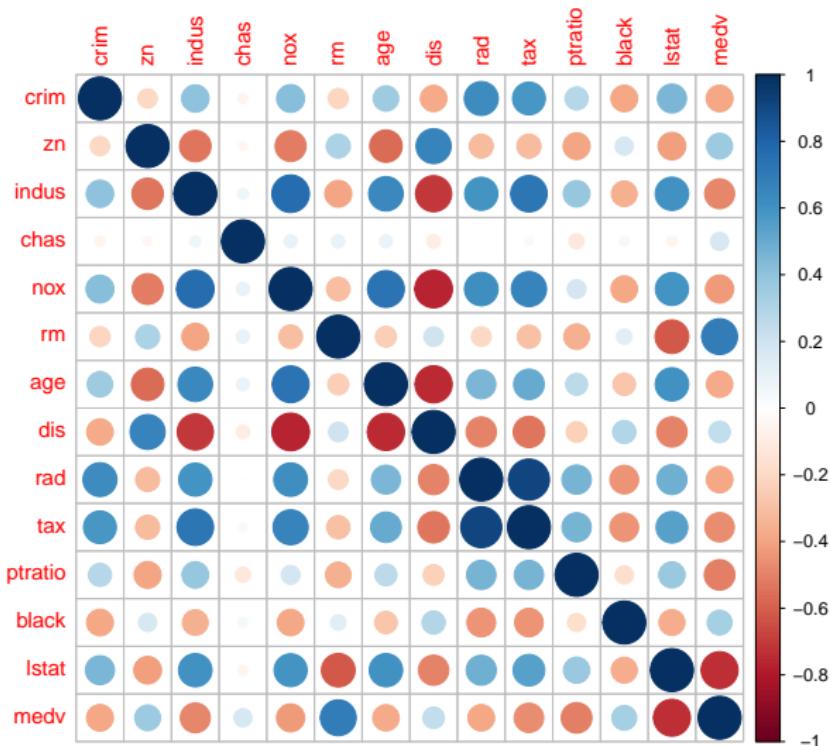
$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{(b)}(x) \quad (6)$$

# Boston Housing Price

Data: median value of owner-occupied homes in 506 census tracts in the Boston area from the 1970 census. Other variables include average rooms per house, distance to employment centers, crime rate, percentage of population with lower socioeconomic status, etc.

```
require(MASS) # contains the data set 'Boston'  
require(rpart)  
head(Boston,3)  
  
##      crim zn indus chas   nox    rm    age   dis rad tax ptratio black lstat  
## 1 0.00632 18  2.31     0 0.538 6.58 65.2 4.09     1 296 15.3  397 4.98  
## 2 0.02731  0  7.07     0 0.469 6.42 78.9 4.97     2 242 17.8  397 9.14  
## 3 0.02729  0  7.07     0 0.469 7.18 61.1 4.97     2 242 17.8  393 4.03  
##      medv  
## 1 24.0  
## 2 21.6  
## 3 34.7
```

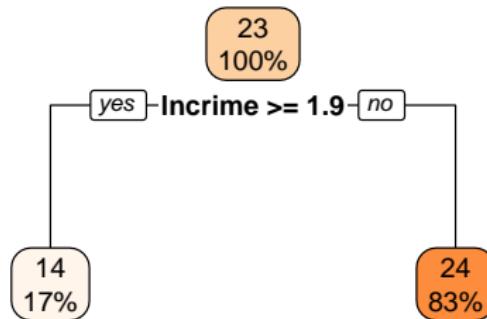
# Boston Housing Price



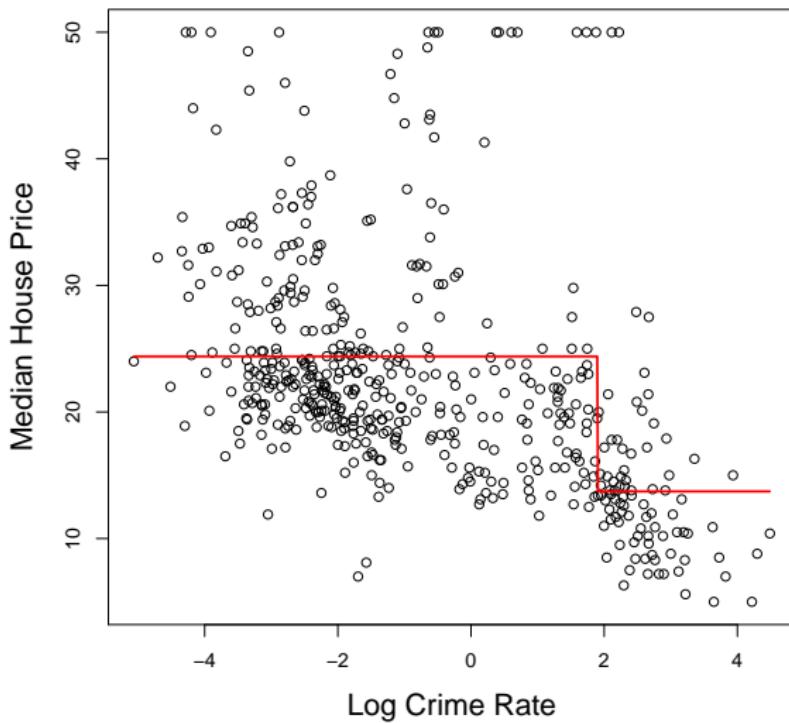
# Boston Housing Price

Let's first predict median house prices based on crime rates only, using a single **decision stump** – the simplest tree with a single split at root.

```
# Fitting a single stump
n = nrow(Boston)
Boston$lncrime = log(Boston$crim)
fit = rpart(medv~lncrime,data=Boston,control=rpart.control(maxdepth=1))
```



# Boston Housing Price

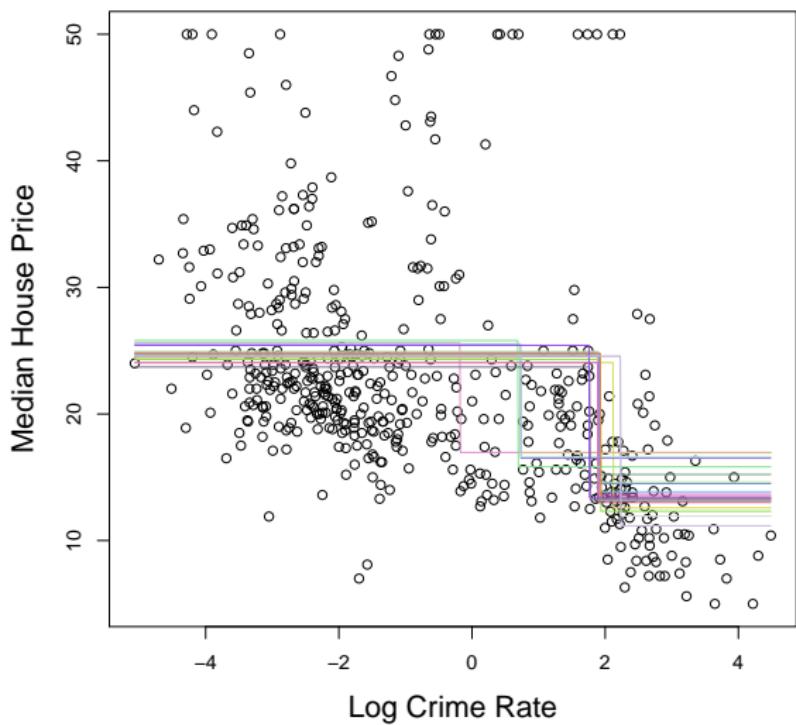


# Boston Housing Price

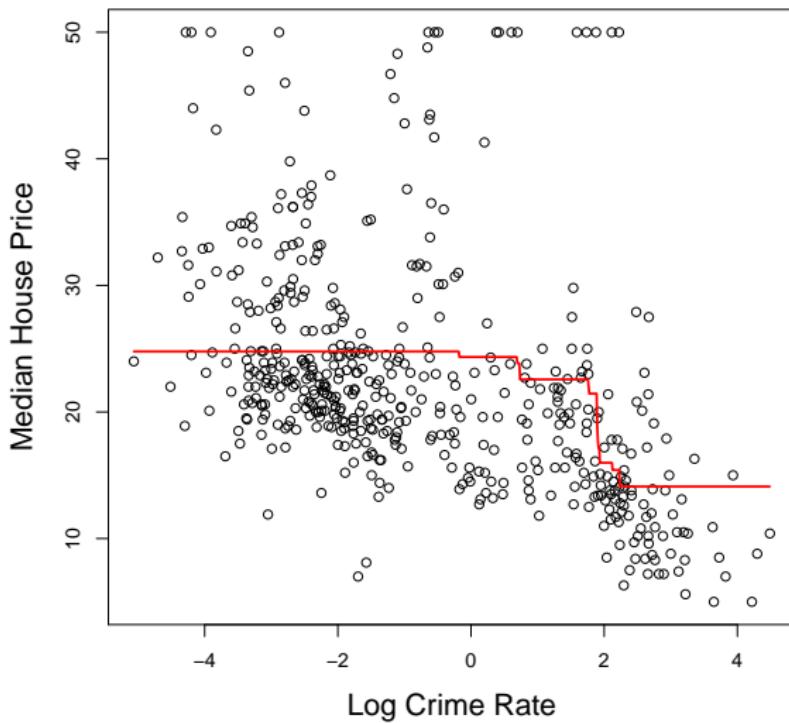
Now let's perform bagging using  $B = 20$  bootstrap samples and make predictions of housing price on a range of crime rate values:

```
crimelims = range(Boston$lncrime)
crime.grid = seq(crimelims[1],crimelims[2],0.01)
#
# Function to fit a single stump
stump = function(index,b){
  fit = rpart(medv~lncrime,data=Boston,subset=indx,
              control=rpart.control(maxdepth=1))
  return(predict(fit,newdata=list(lncrime=crime.grid)))
}
# Bagging
B = 20
yhat_bag = 0
for (b in 1:B){
  bootsample = sample(n,n,replace=T)
  yhat_bag = yhat_bag + stump(bootsample,b)/B
}
```

# Boston Housing Price



# Boston Housing Price



# Bagging for Classification

For classification problems, let  $\hat{p}_j^{(b)}(x)$  be the class probabilities estimated on bootstrap sample  $b$ . We have:

$$\hat{f}(x) = \arg \max_{j \in \{1, \dots, J\}} \hat{p}_j(x) \quad (7)$$

, where

$$\hat{p}_j(x) = \frac{1}{B} \sum_{b=1}^B \hat{p}_j^{(b)}(x)$$

This is the *probability approach*.

# Bagging for Classification

Alternatively, for classification problems, let

$$\hat{f}(x) = \arg \max_{j \in \{1, \dots, J\}} \sum_{b=1}^B \mathcal{I}(\hat{f}^{(b)}(x) = j) \quad (8)$$

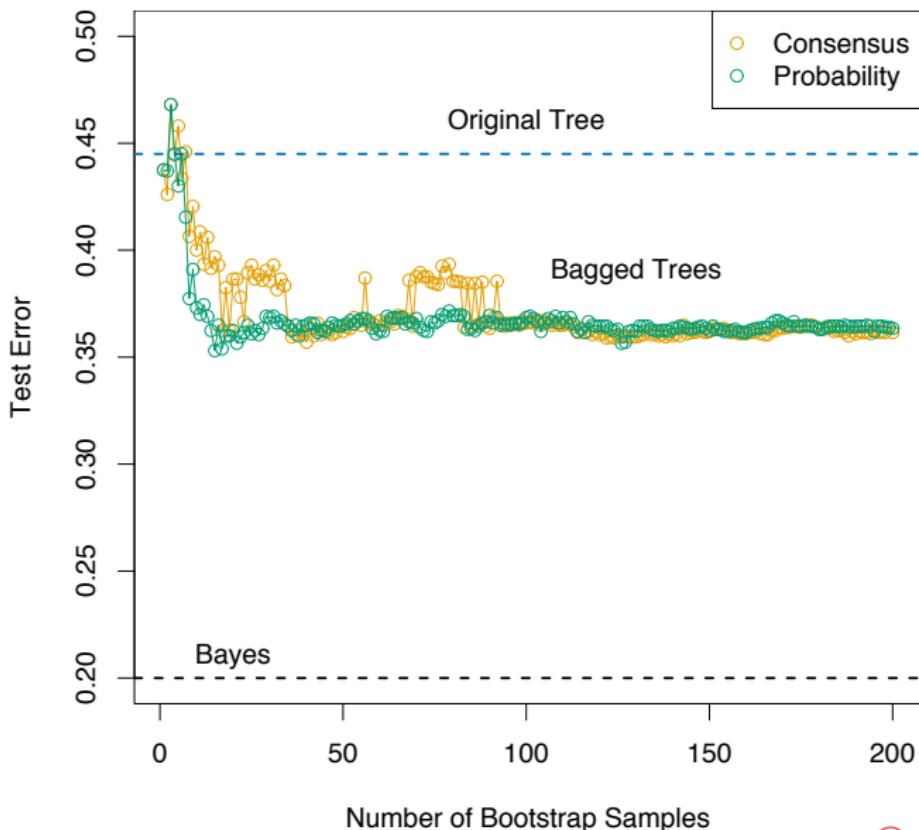
This is the *consensus approach* – decision by “**majority vote**”.

## Bagging for Classification

Simulation:  $N = 30, p = 5, J = 2$ . Each predictor has a standard Gaussian distribution with pairwise correlation 0.95.  $y$  is generated according to  $p(y = 1 | x_1 \leq 0.5) = 0.2$  and  $p(y = 1 | x_1 > 0.5) = 0.8$ .

Bagging is performed by fitting classification trees (with no pruning) to 200 bootstrap samples.

# Bagging for Classification



# The Wisdom of Crowds

Consider a binary classification problem where  $y \in \{F, T\}$ . Suppose that for a given input  $x$ , the true value of  $y$  is  $T$ . We have  $B$  independent classifiers,  $f^{(b)}(x)$ ,  $b = 1, \dots, B$ , and each has a misclassification rate of 0.4, i.e.,  $\Pr(f^{(b)}(x) = F) = 0.4$ .

Let  $\mathbb{V} = \sum_{b=1}^B \mathcal{I}(f^{(b)}(x) = T)$  be the total number of “votes” for  $y = T$  among the  $B$  classifiers. Then

$$\mathbb{V} \sim \text{Binomial}(B, 0.6)$$

The bagged classifier (consensus approach) is

$$f^{\text{bag}}(x) = \arg \max_{j \in \{F, T\}} \sum_{b=1}^B \mathcal{I}(f^{(b)}(x) = j)$$

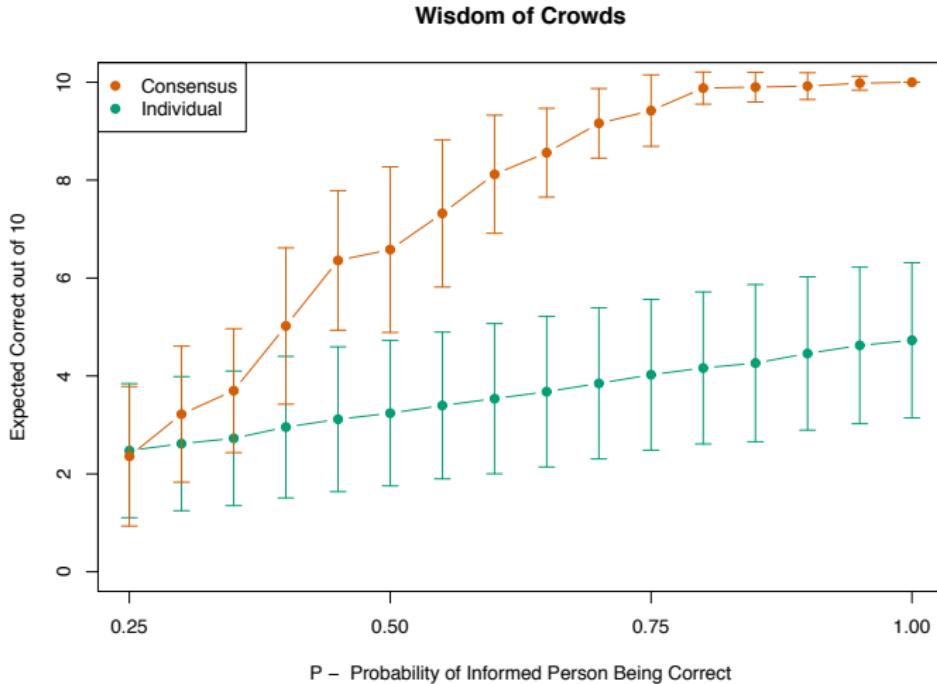
# The Wisdom of Crowds

Hence the misclassification rate of the bagged classifier is

$$\Pr(f^{\text{bag}}(x) = F) = \Pr\left(\frac{\text{votes}}{B} < \frac{B}{2}\right) \rightarrow 0 \quad \text{as } B \rightarrow \infty \quad (9)$$

- (9) is true in this example as long as the misclassification rate of the individual classifier is less than 0.5. If  $\Pr(f^{(b)}(x) = F) > 0.5$ , then  $f^{\text{bag}}(x)$  will become perfectly inaccurate as  $B \rightarrow \infty$ .
- Bagging a good classifier can improve predictive accuracy, but bagging a bad one hurts.

# The Wisdom of Crowds



50 members vote in 10 categories, each with 4 nominations. For each category, only 15 (randomly selected) voters are informed (probability of selecting the “correct” candidate  $> 0.25$ ).

# Bagging

- Bagging can dramatically reduce the variance and **stabilize** the predictions of unstable procedures, and is therefore particularly useful for **high-variance, low-bias** procedures.
- Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias.
- Real structure that persists across data sets shows up in the average. Noisy useless signals will average out to have no effect.
- This technique of **model averaging** is central to many advanced nonparametric learning algorithms.
- Downside? Loss of interpretability. *A bagged tree is no longer a tree.* Bagging improves prediction accuracy at the expense of interpretability.

## Random Forests

- **Random forests** improve on bagging by reducing the correlation between the sampled trees.
- We know that given a set of  $N$  identically distributed random variables with variance  $\sigma^2$  and positive pairwise correlation  $\rho$ , the variance of the mean is  $\rho\sigma^2 + (1 - \rho)\frac{\sigma^2}{N}$ .
- Hence, in the context of decision trees, we can improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much.
- This is achieved in random forests through random selection of the input variables in the tree-growing process.

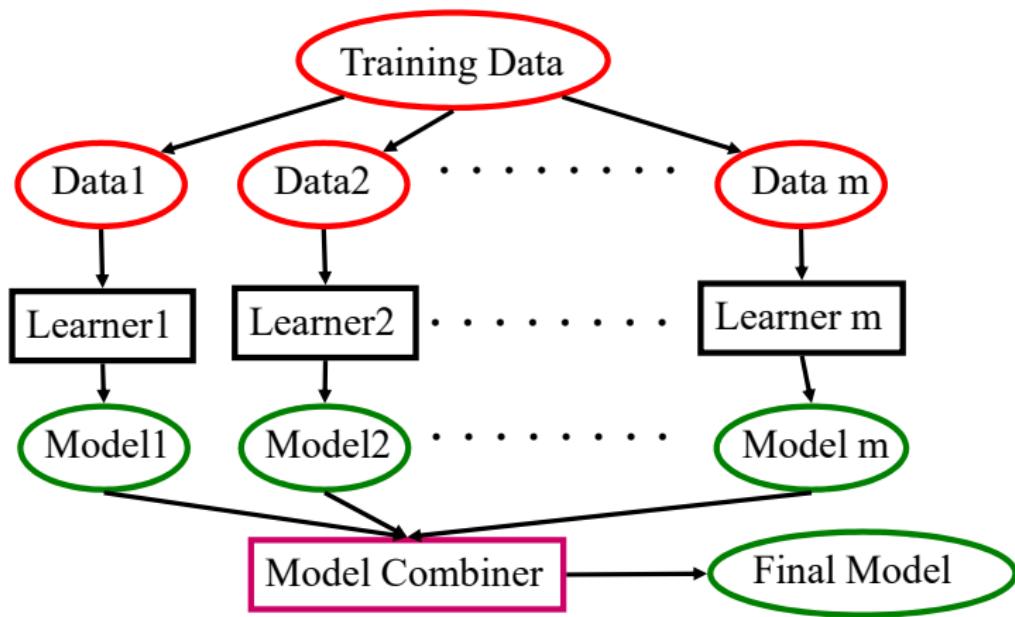
## Random Forests

- As in bagging, random forests build a number of trees on bootstrapped samples.
- When building these trees, each time a split in a tree is considered, a random selection of  $m$  ( $m < p$ ) predictors is chosen as split candidates from the full set of  $p$  predictors. The split is allowed to use only one of these  $m$  predictors.
- A fresh selection of  $m$  predictors is taken at each split.
  - ▶ Typical choice of  $m$ :  $m = \sqrt{p}$ .

# Ensemble Learning

- Bagging and random forests are examples of **ensemble methods**.
- Instead of learning a single hypothesis from a hypothesis set, ensemble methods select a collection (ensemble) of hypotheses and combine their predictions.
- Doing so often improves prediction accuracy at the expense of interpretability.

# Ensemble Learning



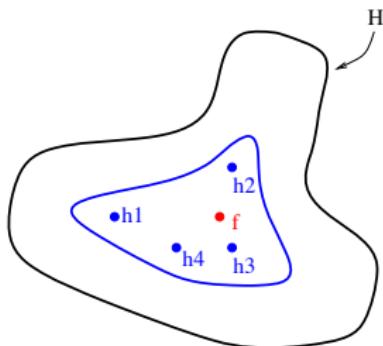
# Ensemble Learning

Three reasons why ensemble methods may work well:

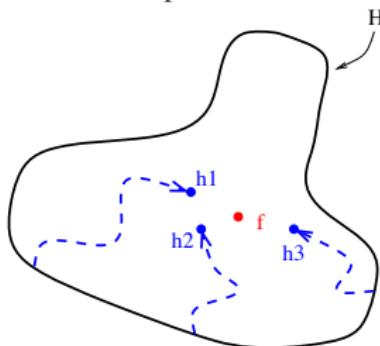
- ① **Statistical**: for a given hypothesis set  $\mathcal{H}$ , when sample size is small, many different hypotheses in  $\mathcal{H}$  can all give the same accuracy on the training data. Combining their predictions helps lower the variance and improve prediction accuracy.
- ② **Computational**: combining local optima produced by local search
- ③ **Representational**: by forming weighted sums of hypotheses drawn from  $\mathcal{H}$ , it might be possible to expand  $\mathcal{H}$  and produce a better approximation to target function  $f$ .

# Ensemble Learning

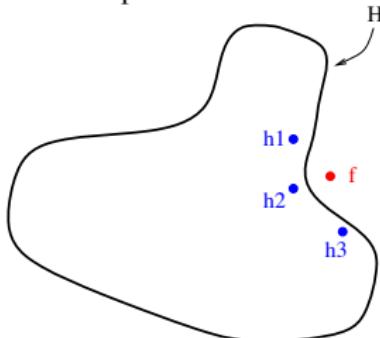
Statistical



Computational



Representational



# Boosting

- Like bagging, boosting is a general and powerful tool that can be applied to many statistical learning methods for regression or classification.
- While bagging fits a model (**base learner**) independently to multiple bootstrap samples, boosting fits a simple model (**weak learner**) to the training data **sequentially** to construct a more complex model (**strong learner**).
- Basic idea: in each round, the training data are **re-weighted** according to the performance of the model in the previous round. Observations that were not accurately predicted see their weights increase. The model is then fit to the re-weighted data. In this way, we **slowly** improve the performance of the model in areas where it does not perform well.

## L2 Boost (Regression)

① Set  $\hat{f}(x) = 0$  and  $r_i = y_i \forall i$

② For  $b = 1, \dots, B$ :

① Fit a regression model  $\hat{f}^{(b)}$  to the training data  $\{(x_1, r_1), \dots, (x_N, r_N)\}$ .

② Update  $\hat{f}$  by adding in a *shrunken* version of  $\hat{f}^{(b)}$ :

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^{(b)}(x)$$

③ Update the residuals:

$$r_i \leftarrow r_i - \lambda \hat{f}^{(b)}(x_i)$$

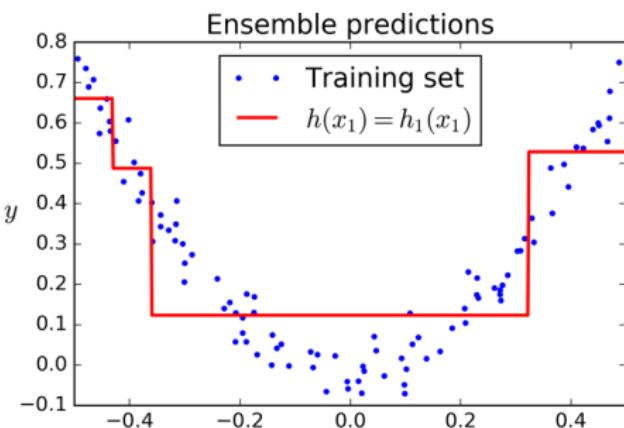
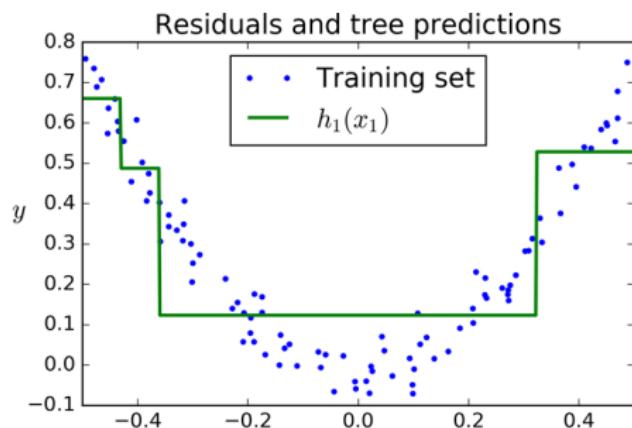
④ Output

$$\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^{(b)}(x)$$

## Boosting for Regression

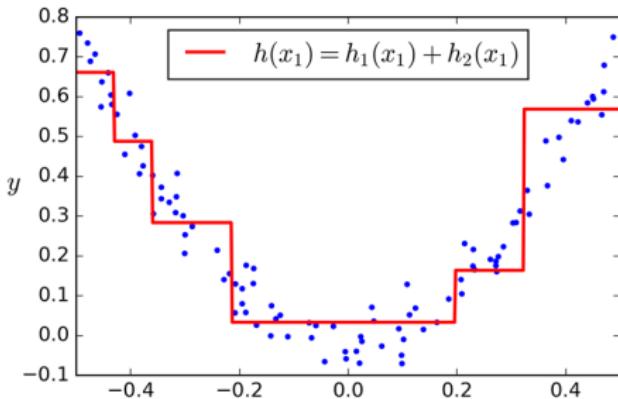
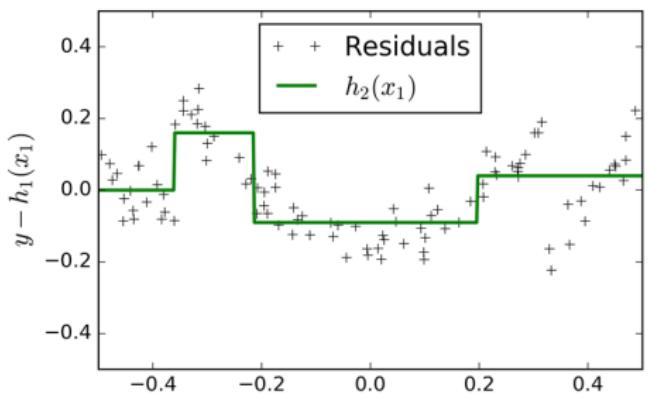
- When applying boosting to regression algorithms, in each round  $b$ , we simply fit the base learner,  $\hat{f}^{(b)}$ , to the residuals of  $\hat{f}^{(b-1)}$ . This is akin to re-weighting the training data: observations that were fit well in the previous round see their importance diminished ( $r_i$  small), while those that were not fit well become more important ( $r_i$  large) in this round.
- The base learner  $\hat{f}^{(b)}$  is usually a *weak learner*, e.g., a stump.
- The shrinkage parameter  $\lambda$  is a small positive number (typical values: 0.01, 0.001, etc.). The boosting approach is to learn *slowly* and  $\lambda$  controls the rate of learning.

# Boosting for Regression



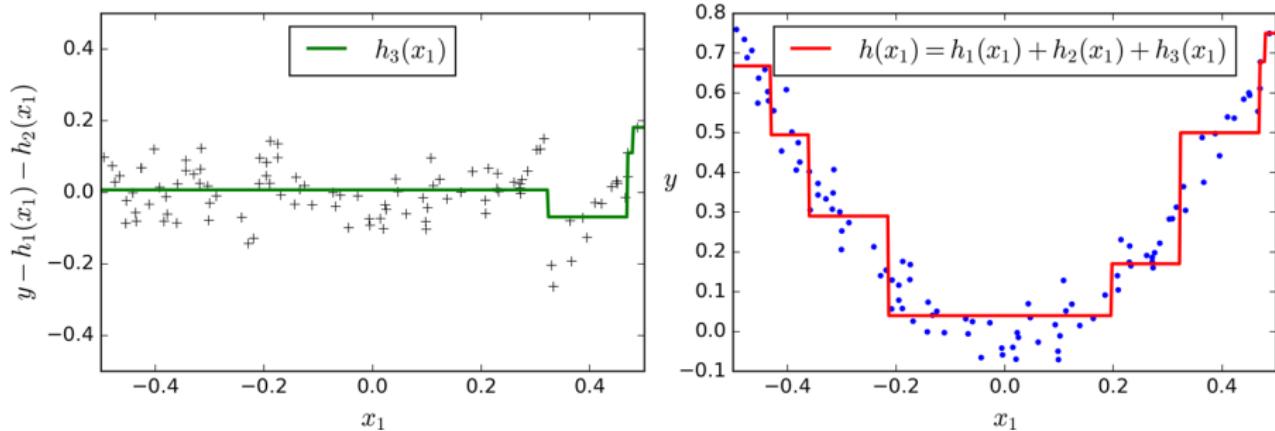
L2 Boost ( $\lambda = 1$ )

# Boosting for Regression



L2 Boost ( $\lambda = 1$ )

# Boosting for Regression



L2 Boost ( $\lambda = 1$ )

## AdaBoost (Binary Classification)

- ① Let  $y \in \{-1, 1\}$ . Initialize  $w_i = \frac{1}{N} \forall i$ .
- ② For  $b = 1, \dots, B$ :
  - ① Fit a classifier  $\hat{f}^{(b)}$  to the training data by minimizing the weighted error

$$\frac{\sum_{i=1}^N w_i \mathcal{I}(y_i \neq f^{(b)}(x_i))}{\sum_{i=1}^N w_i}$$

- ② Let  $\alpha_b = \log((1 - \epsilon_b)/\epsilon_b)$ , where  $\epsilon_b$  is the weighted error of  $\hat{f}^{(b)}$ , and update  $w_i$  as follows:

$$w_i \leftarrow w_i \exp(\alpha_b \mathcal{I}(y_i \neq \hat{f}^{(b)}(x_i)))$$

- ③ Output

$$\hat{f}(x) = \text{sign} \left( \sum_{b=1}^B \alpha_b \hat{f}^{(b)}(x) \right)$$

## Boosting for Classification

- For classification problems, the boosted classifier is a weighted sum of individual classifiers, with weights proportional to each classifier's accuracy on the training set (good classifiers get more weight).
- In AdaBoost, If an individual classifier has accuracy  $< 50\%$  ( $\epsilon_b > 0.5$ ), we flip the sign of its predictions and turn it into a classifier with accuracy  $> 50\%$ . This is achieved by making  $\alpha_b < 0$  so that the classifier enters negatively into the final hypothesis.

# Boosting for Classification

- In each round of the boosting process, we re-weight the observations in the training data: those that were misclassified in the previous round<sup>4</sup> see their weights increase relative to those that were correctly classified. In this way, successive classifiers are forced to place greater emphasis on points that have been misclassified by previous classifiers, and data points that continue to be misclassified by successive classifiers receive ever greater weight<sup>5</sup>.

---

<sup>4</sup> After classifiers with  $\epsilon_b > 0.5$  are “flipped”.

<sup>5</sup> This adaptive data weighting scheme gives the algorithm its name: AdaBoost stands for “Adaptive Boosting”.

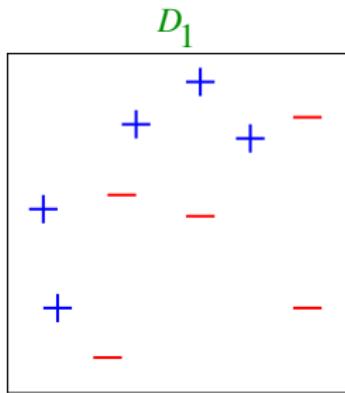
## Boosting for Classification

Now suppose we use decision tree as the base classifier. To fit classification trees to weighted data, we calculate the weighted  $\hat{p}_j^m$  for each region:

$$\hat{p}_j^m = \frac{\sum_{x_i \in R_m} w_i \mathcal{I}(y_i = j)}{\sum_{x_i \in R_m} w_i} \quad (10)$$

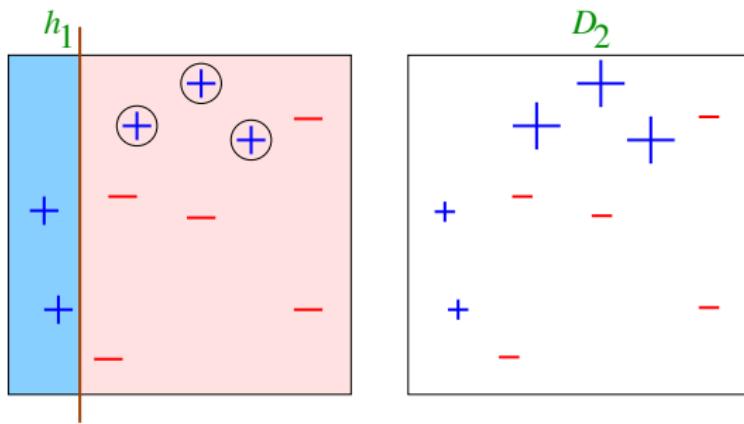
and let (10) replace (4) in the CART algorithm.

# Boosting for Classification



Weak classifiers = vertical or horizontal half-planes

# Boosting for Classification

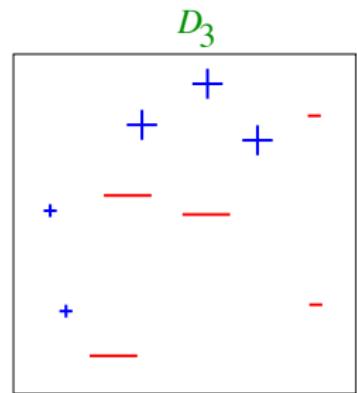
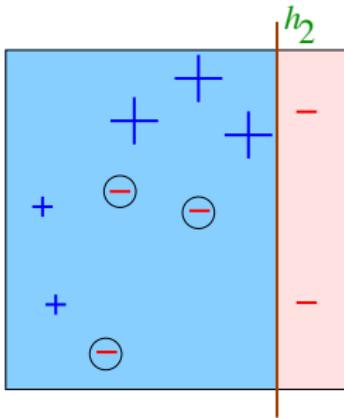
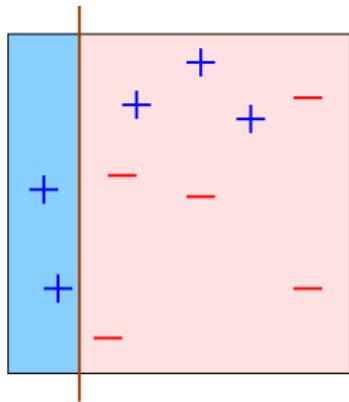


$$\epsilon_1 = 0.30$$

$$\alpha_1 = 0.42$$

Round 1

# Boosting for Classification

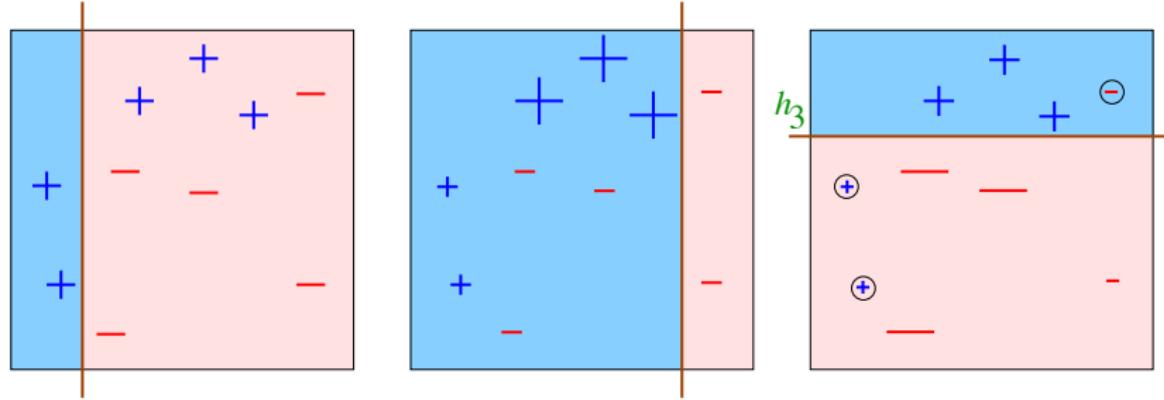


$$\varepsilon_2 = 0.21$$

$$\alpha_2 = 0.65$$

Round 2

# Boosting for Classification

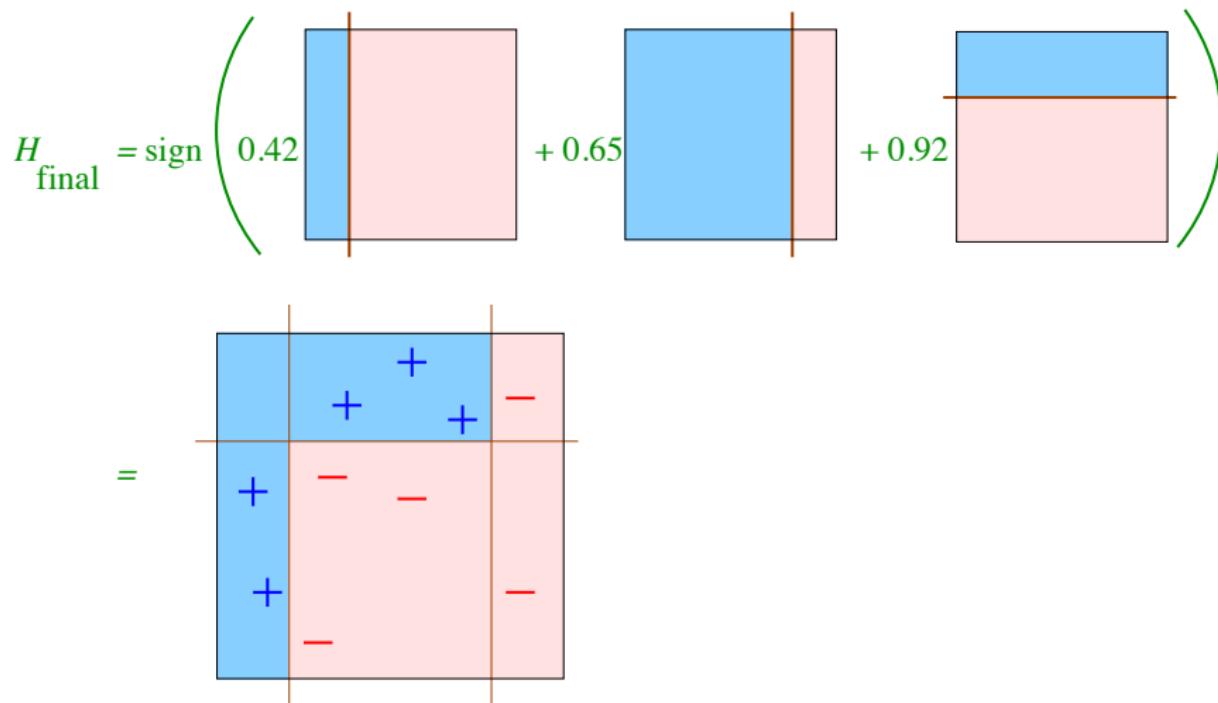


$$\epsilon_3 = 0.14$$

$$\alpha_3 = 0.92$$

Round 3

# Boosting for Classification



Final Classifier

# Boosting

In essence, boosting is a forward stepwise algorithm to fit an adaptive additive model:

$$f(x) = \alpha_1\phi(x; \beta_1) + \cdots + \alpha_M\phi(x; \beta_M) \quad (11)$$

(11) is equivalent in functional form to a linear basis function model. The difference is that here the basis functions are *base learners* (e.g., decision trees) whose functional form are not chosen before seeing the data but are *learned* from the data – they are called **adaptive basis functions**.

# Boosting

Fitting (11) requires minimizing the in-sample error:

$$E_{in}(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(y_i, f(x_i))$$

, where  $\theta = \{\alpha_m, \beta_m\}_{m=1}^M$ .

- L2 loss:  $\ell(y, f(x)) = (y - f(x))^2 \Rightarrow$  L2 Boost
- Exponential loss:  $\ell(y, f(x)) = \exp(-y \cdot f(x)) \Rightarrow$  AdaBoost<sup>6</sup>

---

<sup>6</sup>See Appendix for why AdaBoost corresponds to forward stepwise additive modeling with exponential loss.

# Boosting

Instead of doing a global minimization, the boosting strategy is to follow a forward stepwise procedure by adding basis functions *one by one*<sup>7</sup>:

- Choose the  $\phi(x; \beta_1)$  and  $\alpha_1$  that gives the smallest in-sample error.  
Let  $\hat{f}(x) = \hat{\alpha}_1 \phi(x; \hat{\beta}_1)$ .
- Choose the  $\phi(x; \beta_2)$  and  $\alpha_2$  that gives the largest *additional* reduction in in-sample error:

$$(\hat{\alpha}_2, \hat{\beta}_2) = \arg \min_{\alpha_2, \beta_2} \sum_{i=1}^N \ell(y_i, \hat{f}(x_i) + \alpha_2 \phi(x_i; \beta_2))$$

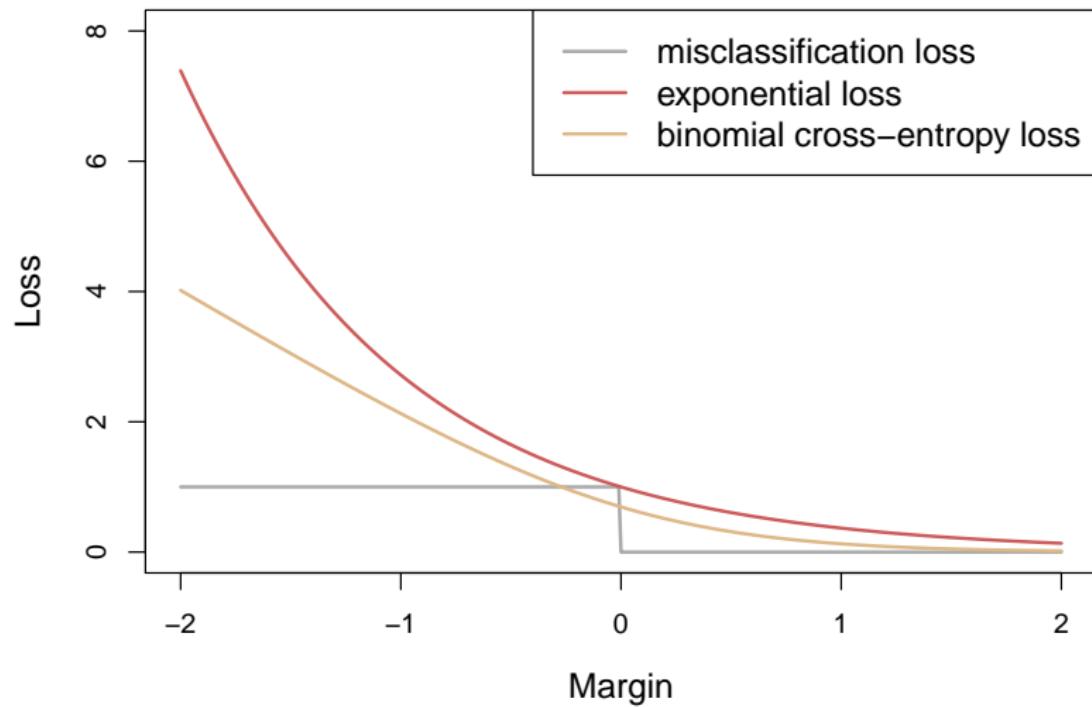
$$\text{Let } \hat{f}(x) = \hat{\alpha}_1 \phi(x; \hat{\beta}_1) + \hat{\alpha}_2 \phi(x; \hat{\beta}_2).$$

- And continue ...

---

<sup>7</sup>Similar to forward stepwise linear regression.

# Boosting



# Boston Housing Price

Now let's predict median house prices based on all variables in the data set:

```
train = sample(nrow(Boston), nrow(Boston)*0.7) # create training & test set

#####
# Linear Regression #
#####

require(AER)
fit = lm(medv~., data=Boston[train,])
#
# test error
y = Boston[-train,"medv"] # true medv on test data
yhat = predict(fit, newdata=Boston[-train,])
err = mean((yhat-y)^2)
err

## [1] 20
```

# Boston Housing Price

```
coeftest(fit)

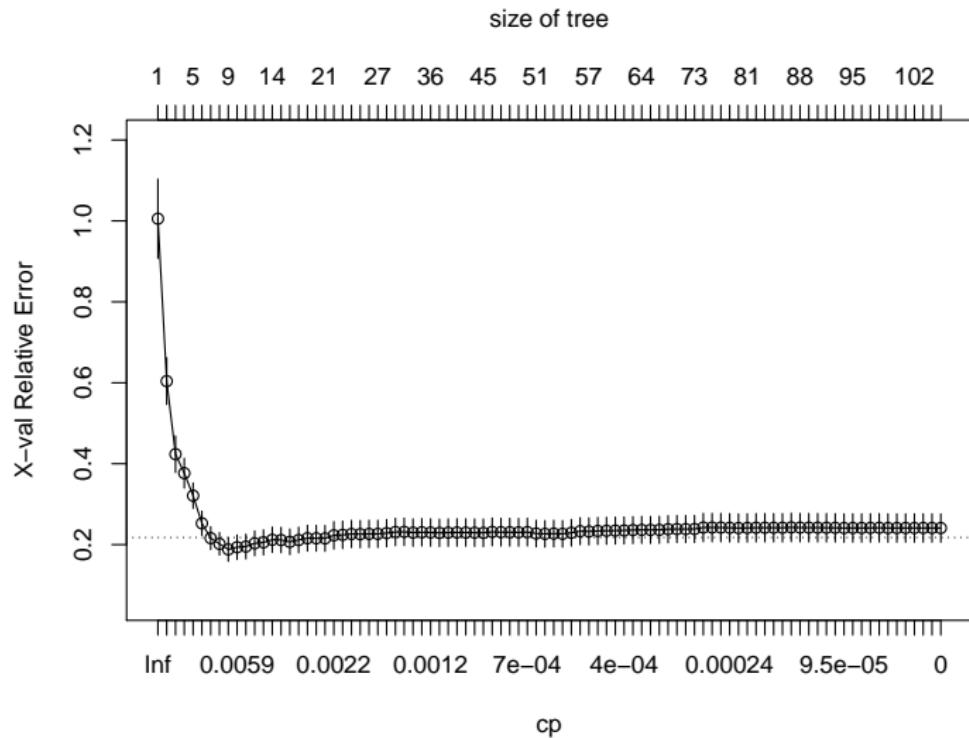
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 36.70285   6.23988   5.88  9.7e-09 *** 
## crim        -0.13702   0.03813  -3.59  0.00037 *** 
## zn          0.04984   0.01848   2.70  0.00735 **  
## indus       -0.00786   0.07909  -0.10  0.92091    
## chas         3.81260   1.05081   3.63  0.00033 *** 
## nox        -19.20027   4.62601  -4.15  4.2e-05 *** 
## rm          3.65357   0.50990   7.17  4.9e-12 *** 
## age         0.00995   0.01705   0.58  0.56002    
## dis        -1.50116   0.25087  -5.98  5.5e-09 *** 
## rad         0.32176   0.08097   3.97  8.6e-05 *** 
## tax        -0.00980   0.00483  -2.03  0.04315 *  
## ptratio     -0.84854   0.16634  -5.10  5.6e-07 *** 
## black       0.00753   0.00344   2.19  0.02951 *  
## lstat      -0.60700   0.06055  -10.02 < 2e-16 *** 
## ---
```

# Boston Housing Price

```
#####
# Decision Tree #
#####
# Step 1: grow tree until stopping criteria:
#           cp: cost complexity parameter; any split must lower in-sample
#                   error by a factor of cp in order to continue grow
#           minbucket: minimum number of observations in any terminal node
fit0 = rpart(medv~.,data=Boston,subset=train,
             control=rpart.control(cp=0,minbucket=2))
#
# Step 2: prune the tree
# select the cost-complexity parameter ("cp") with the smallest
# (relative) cross-validation error ("xerror")
fit = prune(fit0,
            cp=fit0$cptable[which.min(fit0$cptable[, "xerror"]),"CP"])
```

# Boston Housing Price

```
plotcp(fit0)
```



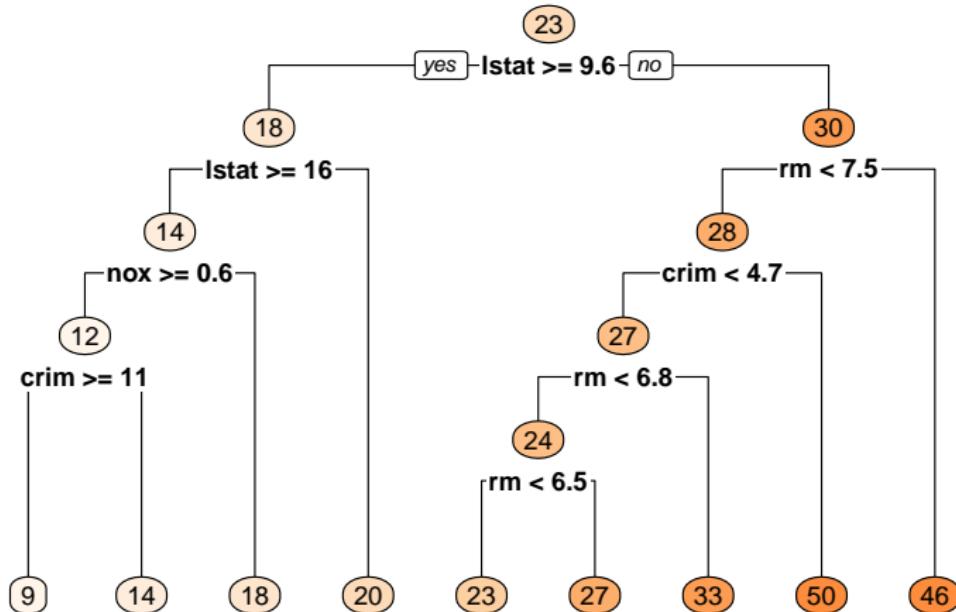
# Boston Housing Price

```
# test error
y = Boston[-train, "medv"]
yhat = predict(fit, newdata=Boston[-train,])
err = mean((yhat-y)^2)
err

## [1] 24.7
```

# Boston Housing Price

pruned tree



# Boston Housing Price

```
#####
# Bagging #
#####
require(randomForest)
# mtry: number of predictors considered for each split
# bagging is a special case of a random forest with mtry = all predictors
fit = randomForest(medv~., data=Boston, subset=train,
                     mtry=13, importance=TRUE)
fit

##
## Call:
##   randomForest(formula = medv ~ ., data = Boston, mtry = 13, importance =
##   Type of random forest: regression
##   Number of trees: 500
##   No. of variables tried at each split: 13
##
##   Mean of squared residuals: 10.2
##   % Var explained: 88.2
```

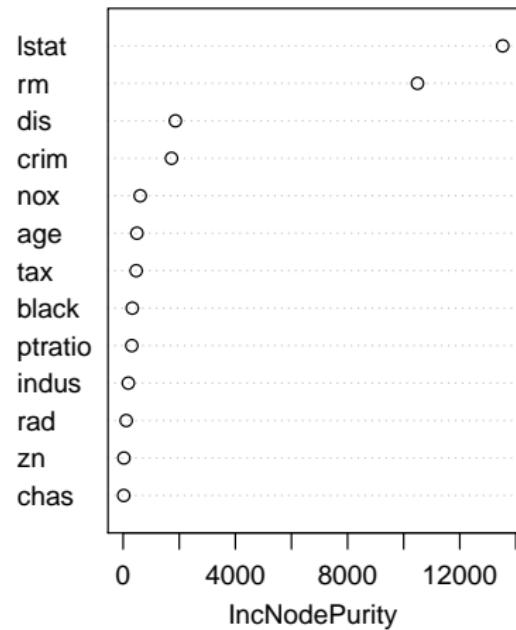
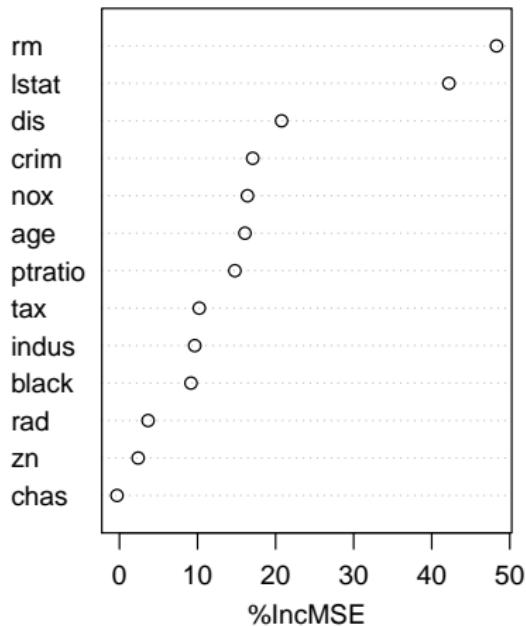
# Boston Housing Price

```
# test error
y = Boston[-train, "medv"]
yhat = predict(fit, newdata=Boston[-train,])
err = mean((yhat-y)^2)
err

## [1] 11.3
```

# Boston Housing Price

Variable Importance Plot



# Boston Housing Price

```
#####
# Random Forest #
#####
fit = randomForest(medv~., data=Boston, subset=train, mtry=5, importance=TRUE)
fit

##
## Call:
##   randomForest(formula = medv ~ ., data = Boston, mtry = 5, importance =
##                 Type of random forest: regression
##                 Number of trees: 500
## No. of variables tried at each split: 5
##
##                 Mean of squared residuals: 11.1
##                 % Var explained: 87.2
```

# Boston Housing Price

```
# test error
y = Boston[-train, "medv"]
yhat = predict(fit, newdata=Boston[-train,])
err = mean((yhat-y)^2)
err

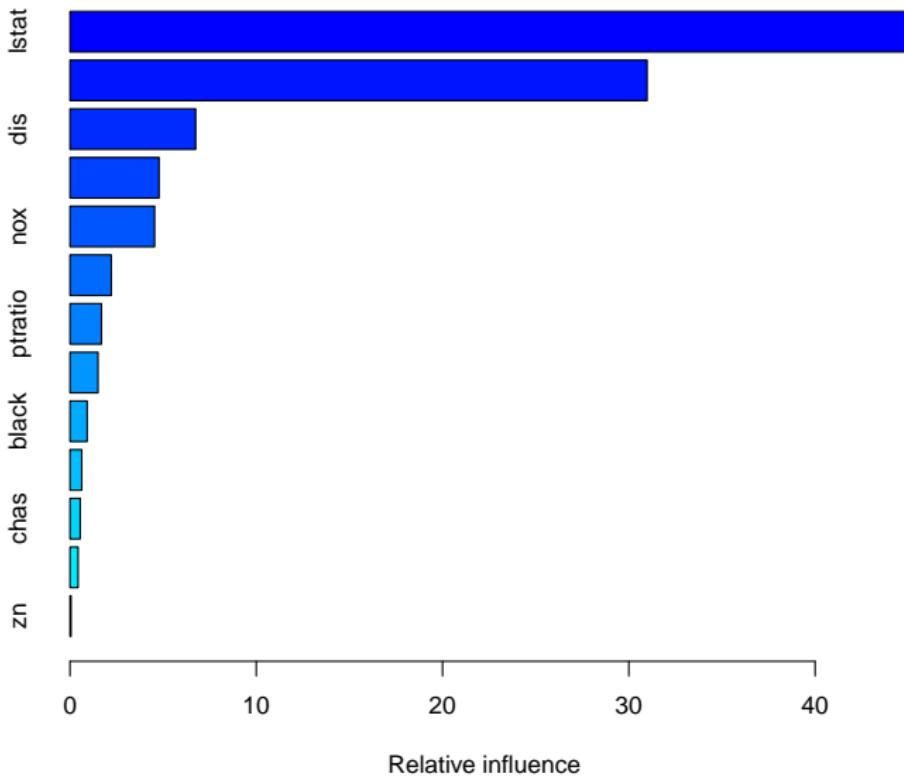
## [1] 7.88
```

# Boston Housing Price

```
#####
# Boosting #
#####
require(gbm)
fit = gbm(medv~., data=Boston[train,], distribution="gaussian",
           n.trees=5000, interaction.depth=5, shrinkage=0.001)
#
# test error
y = Boston[-train, "medv"]
yhat = predict(fit, newdata=Boston[-train,], n.trees=5000)
err = mean((yhat-y)^2)
err

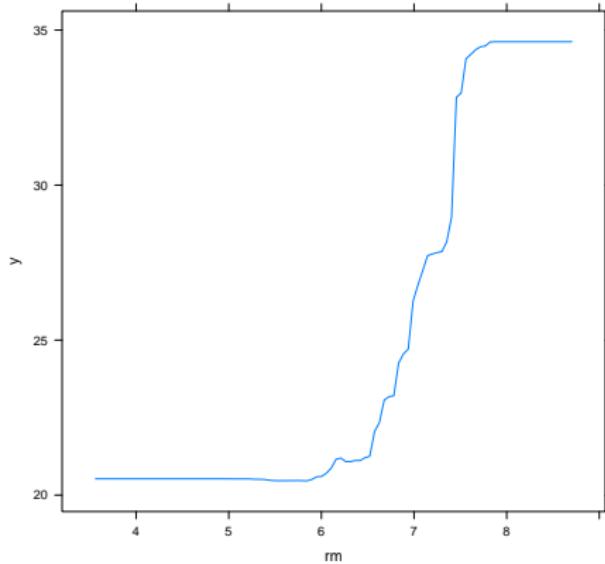
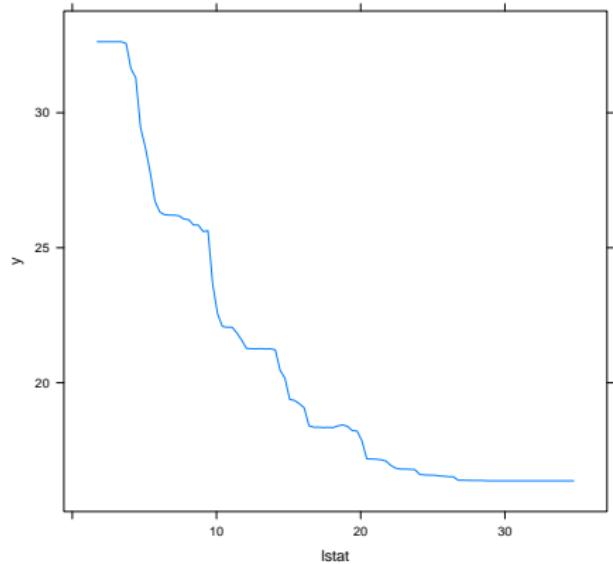
## [1] 9.57
```

# Boston Housing Price



# Boston Housing Price

*Partial dependence plots* illustrate the marginal effect of the selected variables after integrating out the other variables:



# Mode of Transportation

```
transport <- read.csv("Transport.txt")
names(transport) = c("loginc", "distance", "mode")
n = nrow(transport)
train = sample(n,n*0.7) # creat training and test sets
transport_train = transport[train,]
transport_test = transport[-train,]

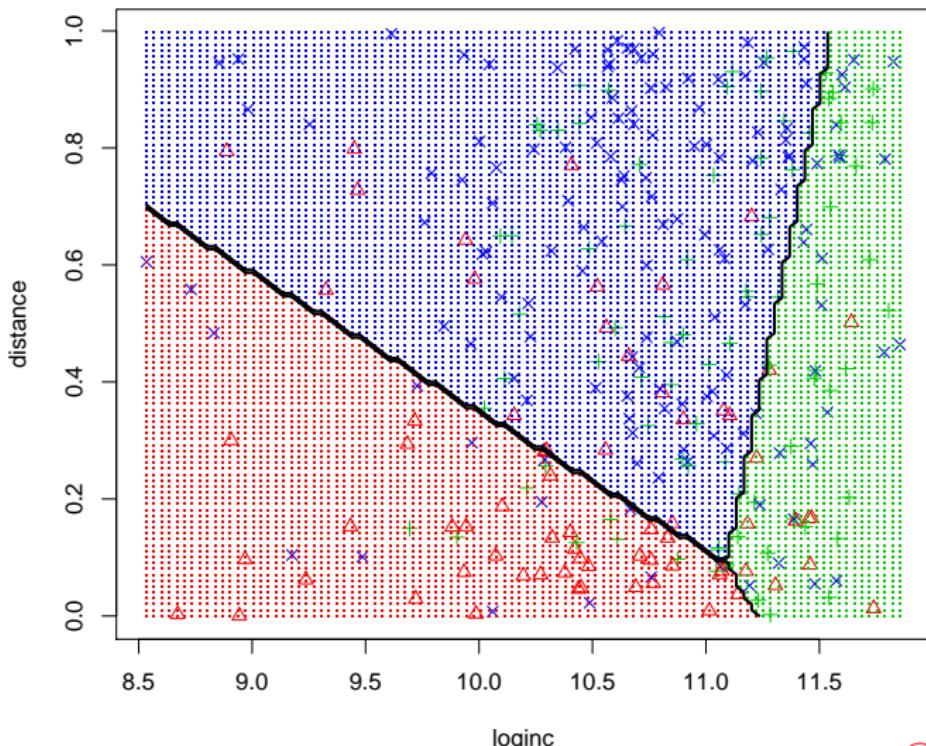
#####
# Logistic Regression #
#####
require(nnet)
fit <- multinom(mode~., data=transport_train)

# test err
yhat = predict(fit, transport_test)
err <- 1-mean(yhat==transport_test$mode)
err

## [1] 0.437
93 / 107
```

# Mode of Transportation

Logistic Regression

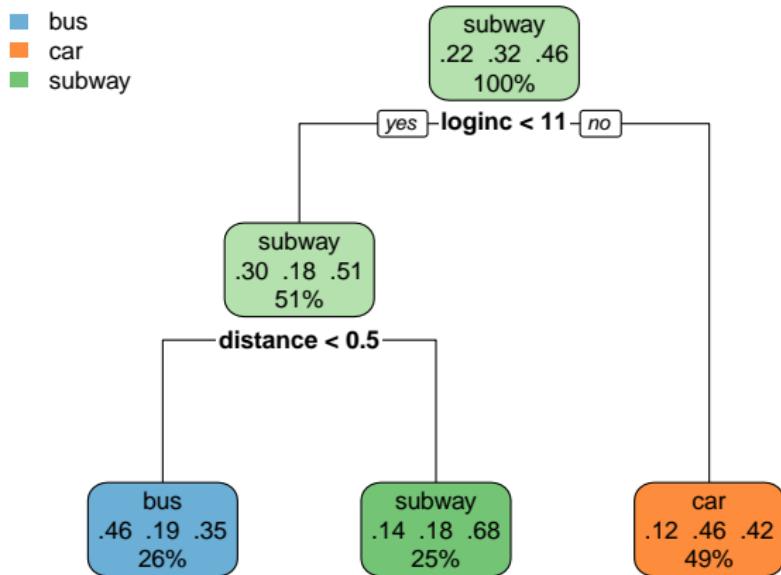


# Mode of Transportation

```
#####
# Classification Tree #
#####
require(rpart)
fit = rpart(mode~.,data=transport_train)
#
# test err
yhat = predict(fit,transport_test,type="class")
err <- 1-mean(yhat==transport_test$mode)
err

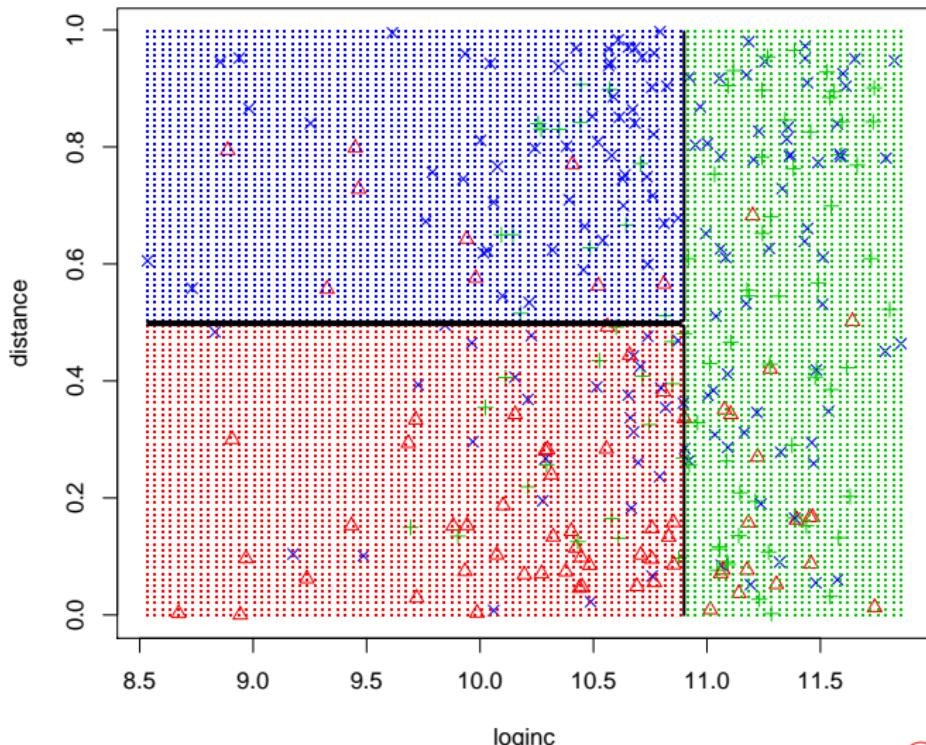
## [1] 0.49
```

# Mode of Transportation



# Mode of Transportation

Decision Tree



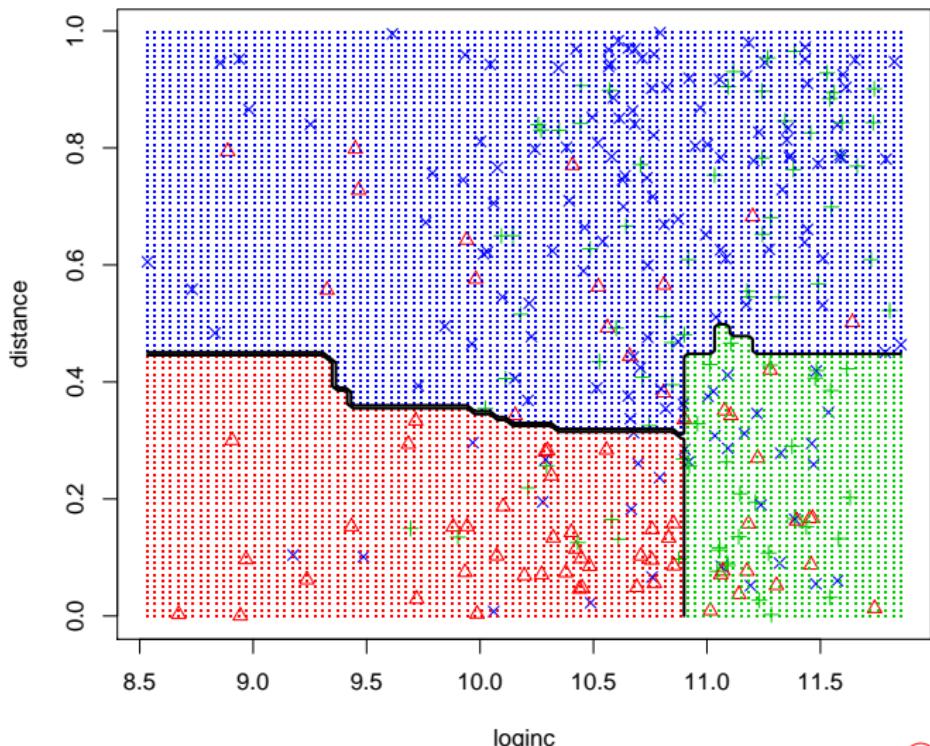
# Mode of Transportation

```
#####
# Random Forest #
#####
require(randomForest)
fit = randomForest(mode~., data=transport_train, mtry=1, maxnodes=4)
#
# test err
yhat = predict(fit, transport_test)
err <- 1-mean(yhat==transport_test$mode)
err

## [1] 0.417
```

# Mode of Transportation

Decision Tree



# Mode of Transportation

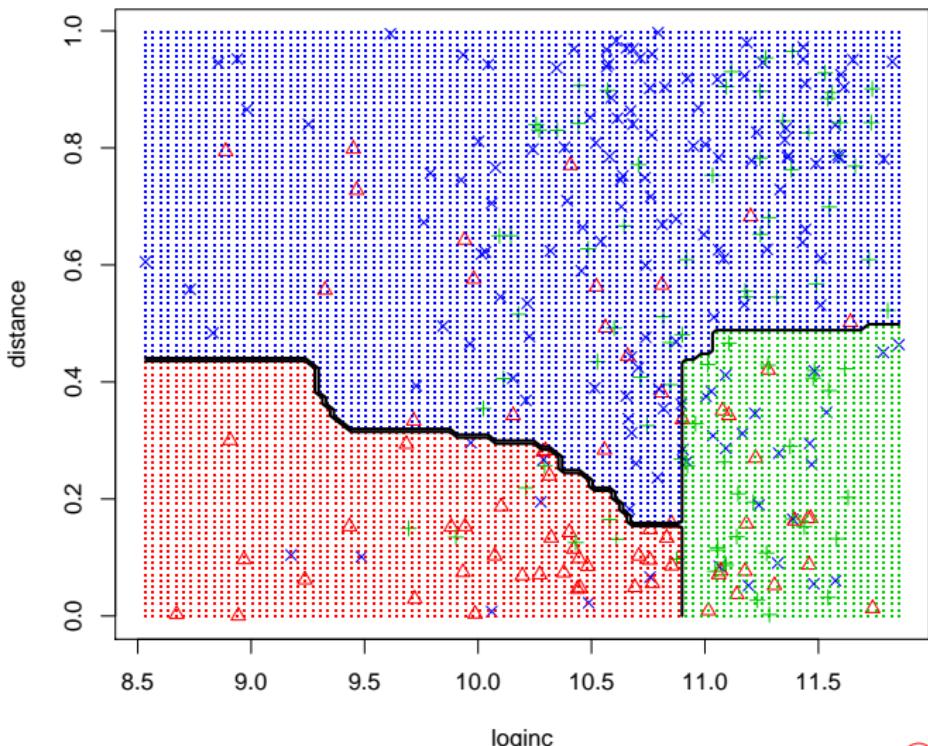
```
#####
# Boosting #
#####

# here we run gradient boosting -- a generalization of AdaBoost
# we use multinomial cross entropy for multiclass classification
require(gbm)
ntree = 500
fit = gbm(mode~, data=transport_train, distribution="multinomial",
           n.trees=ntree, interaction.depth=1, shrinkage=0.001)
#
# test err
require(ramify)
fhat = predict(fit, transport_test, n.trees=ntree)
fhat = drop(fhat)
yhat = argmax(fhat)
err <- 1-mean(yhat==as.numeric(transport_test$mode))
err

## [1] 0.417
```

# Mode of Transportation

Boosting



## Appendix: AdaBoost as Forward Stepwise Modeling

Let  $y \in \{-1, 1\}$ . Start with the forward stepwise procedure outlined on [page 77](#), using the exponential loss function  $\ell(y, f(x)) = \exp(-yf(x))$ .

In step  $m$ , given  $\hat{f}^{(m-1)}(x)$  – the result of the previous  $m - 1$  steps, the goal is find

$$(\hat{\alpha}_m, \hat{\phi}_m) = \arg \min_{\alpha, \phi} \sum_{i=1}^N \exp \left[ -y_i (\hat{f}^{(m-1)}(x) + \alpha \phi(x_i)) \right] \quad (12)$$

$$= \arg \min_{\alpha, \phi} \sum_{i=1}^N w_i^{(m)} \exp [-\alpha y_i \phi(x_i)] \quad (13)$$

, where  $w_i^{(m)} = \exp (-y_i \hat{f}^{(m-1)}(x))$ .

## Appendix: AdaBoost as Forward Stepwise Modeling

Minimizing (13)  $\Rightarrow$

$$\hat{\alpha}_m = \frac{1}{2} \log \frac{1 - \epsilon_m}{\epsilon_m}$$

, where

$$\epsilon_m = \frac{\sum_{i=1}^N w_i^{(m)} \mathcal{I}(y_i \neq \hat{\phi}_m(x_i))}{\sum_{i=1}^N w_i^{(m)}}$$

, and

$$\hat{\phi}_m = \arg \min_{\alpha, \phi} \sum_{i=1}^N w_i^{(m)} \mathcal{I}(y_i \neq \phi(x_i))$$

, i.e.  $\hat{\phi}_m$  is the function that minimizes the weighted misclassification error  $\epsilon_m$ .

## Appendix: AdaBoost as Forward Stepwise Modeling

We can then update  $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{\alpha}_m \hat{\phi}_m(x)$ , and

$$\begin{aligned} w_i^{(m+1)} &= \exp\left(-y_i \hat{f}^{(m)}(x_i)\right) \\ &= \exp\left(-y_i \left[\hat{f}^{(m-1)}(x_i) + \hat{\alpha}_m \hat{\phi}_m(x_i)\right]\right) \\ &= w_i^{(m)} \exp\left(-y_i \hat{\alpha}_m \hat{\phi}_m(x_i)\right) \\ &= w_i^{(m)} \exp(2\hat{\alpha}_m \mathcal{I}(y_i \neq \phi(x_i))) \cdot \exp(-\hat{\alpha}_m) \end{aligned}$$

, where we use the fact that  $-y_i \hat{\phi}_m(x_i) = 2\mathcal{I}(y_i \neq \phi(x_i)) - 1$ .

Therefore, we can multiply  $\hat{\alpha}_m$  by 2, i.e., let  $\hat{\alpha}_m = \log \frac{1-\epsilon_m}{\epsilon_m}$  and update  $w_i$  as

$$w_i^{(m+1)} = w_i^{(m)} \exp(\hat{\alpha}_m \mathcal{I}(y_i \neq \phi(x_i)))$$

This gives us the AdaBoost algorithm.

# Acknowledgement I

Part of this lecture is adapted from the following sources:

- Géron, A. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media.
- Hastie, T., R. Tibshirani, and J. Friedman. 2008. *The Elements of Statistical Learning* (2<sup>nd</sup> ed.). Springer.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Sontag, D. *Introduction To Machine Learning*. Lecture at NYU, retrieved on 2018.01.01. [[link](#)]
- Taddy, M. *Big Data*. Lecture at the University of Chicago Booth School of Business, retrieved on 2017.01.01. [[link](#)]
- Tibshirani, R. *Data Mining*, Lecture at Carnegie Mellon University, retrieved on 2017.01.01. [[link](#)]

## Acknowledgement II

- Van der Schaar, M. and S. Flaxman. *Statistical Machine Learning*. Lecture at Oxford University, retrieved on 2018.01.01. [[link](#)]

# Reference



Dietterich, T. G. 2000. "Ensemble Methods in Machine Learning," In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer.