Ensemble

Mahdi Roozbahani
Georgia Tech

Adapted\inspired from Polo’s slides.

Which Classifier/Model to Choose?

- Possible strategies:
- Go from simplest model to more complex model until you obtain desired accuracy
- Discover a new model if the existing ones do not work for you
- Combine all (simple) models

Common Strategy: Bagging

(Bootstrap Aggregating)

Originally designed for combining multiple models,
to improve classification “stability” [Leo Breiman, 94]

Uses random training datasets
(sampled from one dataset)

http://statistics.about.com/od/Applications/a/What-Is-Bootstrapping.htm

Common Strategy: Bagging
(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\dots,n}$

- Pick a sample S^* with **replacement** of size n
  (S^* called a "bootstrap sample")

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow \quad x^* = \begin{bmatrix} \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \end{bmatrix} \quad y^* = \begin{bmatrix} \_ \\ \_ \\ \_ \\ \_ \end{bmatrix}$$

Common Strategy: Bagging
(<u>B</u>ootstrap <u>Agg</u>regat<u>ing</u>)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\ldots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S called a "bootstrap sample")*

$$S \rightarrow \; x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow \; x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ - & - & - & - \\ - & - & - & - \\ - & - & - & - \end{bmatrix} \quad y^* = \begin{bmatrix} 1 \\ - \\ - \\ - \end{bmatrix}$$

Common Strategy: Bagging
(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\ldots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S called a "bootstrap sample")*

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ \boxed{20 & 21 & 22 & 23} \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ \boxed{-1} \\ -1 \end{bmatrix}$$

$$S^* \rightarrow \quad x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \end{bmatrix} \quad y^* = \begin{bmatrix} 1 \\ \_ \\ \_ \\ \_ \end{bmatrix}$$

Common Strategy: Bagging

(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\dots,n}$

- Pick a sample S^* with **replacement** of size *n*

  (S^* called a "bootstrap sample")

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \end{bmatrix} y^* = \begin{bmatrix} 1 \\ -1 \\ \_ \\ \_ \end{bmatrix}$$

Common Strategy: Bagging
(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\ldots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S^* called a "bootstrap sample")

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \end{bmatrix} y^* = \begin{bmatrix} 1 \\ -1 \\ \_ \\ \_ \end{bmatrix}$$

Common Strategy: Bagging
(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\dots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S called a "bootstrap sample")*

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 9 & 10 & 11 & 12 \\ - & - & - & - \end{bmatrix} y^* = \begin{bmatrix} 1 \\ -1 \\ 1 \\ - \end{bmatrix}$$

Common Strategy: Bagging
(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\ldots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S called a "bootstrap sample")*

$$S \rightarrow \; x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 9 & 10 & 11 & 12 \\ - & - & - & - \end{bmatrix} y^* = \begin{bmatrix} 1 \\ -1 \\ 1 \\ - \end{bmatrix}$$

Common Strategy: Bagging

(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\dots,n}$

- Pick a sample S^* with **replacement** of size *n*
  (S called a "bootstrap sample")*

$$S \rightarrow \quad x = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 5 & 6 & 7 & 8 \end{bmatrix} y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$S^* \rightarrow x^* = \begin{bmatrix} 9 & 10 & 11 & 12 \\ 20 & 21 & 22 & 23 \\ 9 & 10 & 11 & 12 \\ 1 & 2 & 3 & 4 \end{bmatrix} y^* = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

Common Strategy: Bagging

(Bootstrap Aggregating)

Consider the data set $S = \{(X_i, Y_i)\}_{i=1,\ldots,n}$

- Pick a sample S^* with **replacement** of size n
  (S called a "bootstrap sample")*
- Train on S^* to get a classifier f^*
- Repeat above steps B times to get $f_1, f_2, \ldots, f_B$
- Final classifier $f(x) = \text{majority}\{f_b(x)\}_{j=1,\ldots,B}$

Common Strategy: Bagging

Why would bagging work?

- Combining multiple classifiers reduces the variance of the final classifier

When would this be useful?

- We have a classifier with high variance

Bagging decision trees

Consider the data set S

- Pick a sample S^* with replacement of size n
- Grow a decision tree T_b
- Repeat B times to get $T_1,\ldots,T_B$
- The final classifier will be

$$f(x) = \text{majority}\{f_{T_b}(x)\}_{b=1,\ldots,B}$$

Random Forests

Almost identical to <u>bagging decision trees</u>, except we introduce some <u>randomness</u>:

- Randomly pick *m* of the *d* available attributes, at every split when growing the tree
  (i.e., d - m attributes ignored)

Bagged **random** decision trees =
Random forests

What are our Hyper-Parameters in Random Forest

$m = Number\ of\ randomly\ chosen\ attributes$

Usual values for $m = \sqrt{d}, 1, 10$

d is number of dimensions or features or attributes

How to optimize m? Cross-Validation

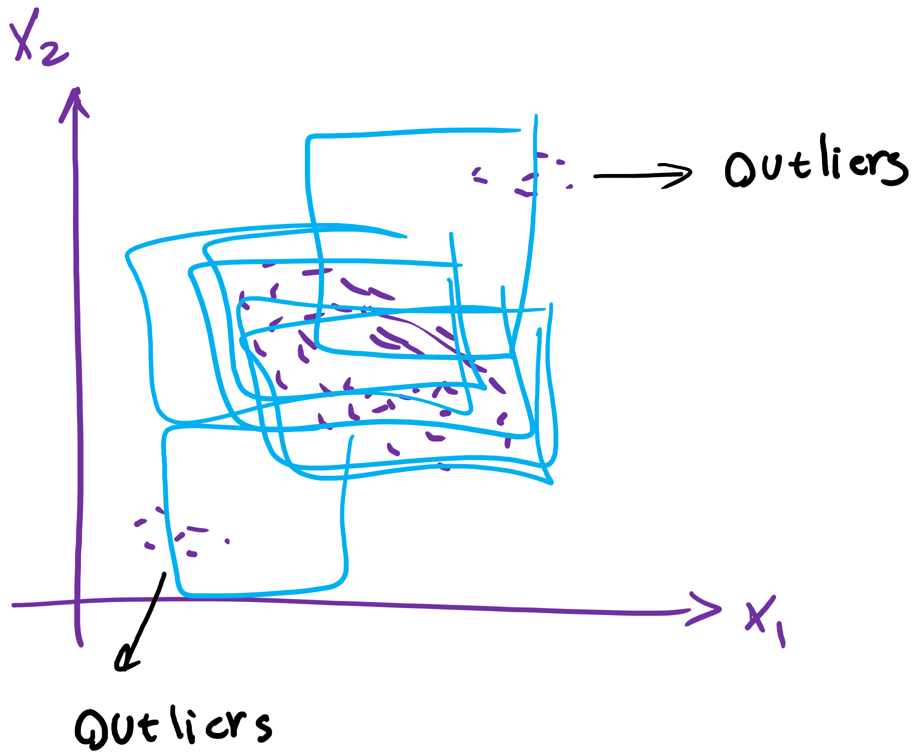
$B = Number\ of\ models\ or\ decision\ trees\ in\ Random\ Forest$

Keep adding trees until training error stabilizes (reaches to a plateau)

Important points about random forests

Algorithm (hyper) parameters

- Size/#nodes of each tree
- as in when building a decision tree
- May randomly pick an attribute, and may even randomly pick the split point!
  - Significantly simplifies implementation and increases training speed
    - PERT - Perfect Random Tree Ensembles
      http://www.interfacesymposia.org/I01/I2001Proceedings/ACutler/ACutler.pdf
    - Extremely randomized trees
      http://orbi.ulg.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf



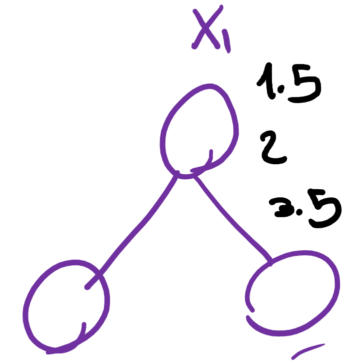
"Random Forest Feature Selection"

Boosting

$$X = \begin{bmatrix} 1 & \\ & \rightarrow 1.5 \\ 2 & \\ & \rightarrow 2.5 \\ 3 & \\ & \rightarrow 3.5 \\ 4 & \end{bmatrix} \quad y_a = \begin{bmatrix} 20 \\ 30 \\ 70 \\ 80 \end{bmatrix}$$

House size   rental

Gradient boosting



$[2, 3, 4]$

$\text{avg(Residual)} = -30 \; [1]$

left

$[-30]$ $[-20, 20, 30]$

↓ avg → Right

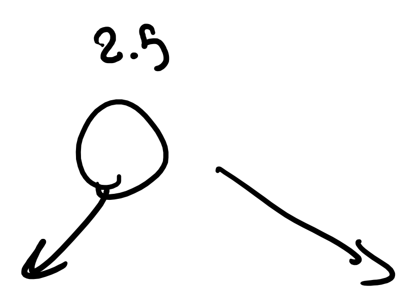
$$(-30 - (-30))^2 = 0 \qquad \frac{-20+20+30}{3} = 10$$

$$(-20-10)^2 + (20-10)^2 + (30-10)^2$$

RSS

$$F_0 = \hat{y}_p = \frac{20+30+70+80}{4} = 50$$

$$\text{residual} = \begin{array}{c} y_a - \hat{y}_p \\ \begin{bmatrix} 20 - 50 = -30 \\ -20 \\ 20 \\ 30 \end{bmatrix} \end{array}$$



$[-30, -20]$ $[20, 30]$

-25 25

$$\text{residuals} = \begin{bmatrix} 20 + 25 = 45 \\ 30 + 25 = 55 \\ 70 - 25 = 45 \\ 80 - 25 = 55 \end{bmatrix}$$

learning rate

$$F_1 = F_0 + \eta \, \underline{h(x)}$$

-25

$$F_2 = F_1 + \eta \, h(x)$$

$$F_3 = F_2 + \eta \, h(x)$$

$x = 2.5$

[1, 2]	[3, 4]
[−30, −20]	[20, 30]

$$h_L^{(1)} = \frac{-30-20}{2} = -25$$

$$h_R^{1} = 25$$

$$RSS = \underbrace{(-30-(-25))^2 + (-20+25)^2}_{L} + \underbrace{(20-25)^2 + (30-25)^2}_{R}$$

$$\hat{y}_p = F$$

learning rate = 1

$$\underline{F_1} = \underline{F_0} + (\eta)\, \underbrace{h^{(1)}}_{\text{avg residuals}}$$

$$F_1 = \begin{array}{c} \\ ① \\ ② \\ ③ \\ ④ \end{array} \begin{array}{c} \hat{y}_p \\ \begin{bmatrix} 50 + (-25) = 25 \\ 50 + (-25) = 25 \\ 50 + (25) = 75 \\ 50 + (25) = 75 \end{bmatrix} \end{array}$$

$\hat{y}_p \rightarrow F_1$

$$R^{(2)} = \begin{array}{c} y_a - \hat{y}_p \\ \begin{bmatrix} 20 - 25 = -5 \\ 30 - 25 = 5 \\ 70 - 75 = -5 \\ 80 - 75 = 5 \end{bmatrix} \end{array}$$

$$y_a = \begin{bmatrix} 20 \\ 30 \\ 70 \\ 80 \end{bmatrix}$$

$$R^{(2)} = \begin{bmatrix} -5 \\ 5 \\ -5 \\ 5 \end{bmatrix} \quad \begin{matrix} \to 1.5 \\ \to 2.5 \\ \to 3.5 \end{matrix}$$

$X = 2.5$

$[-5, 5]$ $\quad$ $[-5, 5]$

$h_L^{(2)} = 0$ $\quad$ $h_R^{(2)} = 0$

$$RSS = (-5)^2 + (5)^2 + (-5)^2 + (5)^2 = 100$$

$$F_1 = \begin{bmatrix} 25 \\ 25 \\ 75 \\ 75 \end{bmatrix} \qquad Y_a = \begin{bmatrix} 20 \\ 30 \\ 70 \\ 80 \end{bmatrix}$$

$X = 1.5$

$h_R^{(2)} = -5$

$[-5]$ $\quad$ $[5, -5, 5]$

$$h_{(R)}^{(2)} = \frac{5-5+5}{3} \approx 1.7$$

$$\underbrace{(-5-\overset{h_R^{(2)}}{(-5)})^2}_{0} + (5-\overset{h_R^{(1)}}{1.7})^2 + (-5-1.7)^2 + (5-1.7)^2 \approx 67$$

$Y_a + \hat{Y}_p \to F_2$

$$F_2 = F_1 + \overset{\text{learning rate}=1}{\eta}\, h^{(2)}$$

$$F_2 = \text{①} \begin{bmatrix} 25 + 1 * (-5) \\ 25 + 1 * (1.7) \\ 75 + (1.7) \\ 75 + (1.7) \end{bmatrix}$$

$$R^{(3)} = \begin{bmatrix} 20 - 20 \\ \cdot\ -- \\ -- \\ \cdot\ - \end{bmatrix}$$