

10.1 Figure 10.5 shows three trees each built on a different bootstrap sample. For each (rectangular) tree leaf, the labels of the training observations that fall into the leaf are shown. The labels are y (yes), n (no), and m (maybe). The order of the observations within a leaf does not matter. Based on these three trees, use bagging to predict a new observation with the x-values: $(x_1, x_2, x_3)^T = (2, -3, 12)^T$.

- (a) What is the bagged prediction based on majority vote?
 (b) What is the bagged prediction based on average probability?

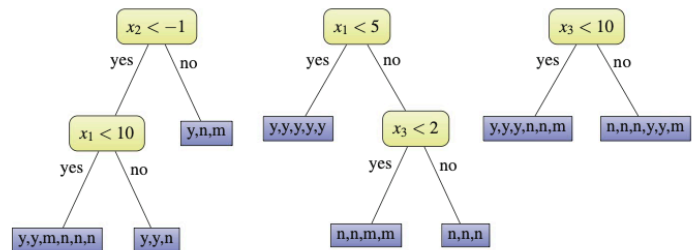


Fig. 10.5 Three trees built from bootstrap samples. Each observation label (y,n,m) for the training data is listed

0.1

$$x_1, x_2, x_3 = 2, -3, 12$$

ⓐ Bagged prediction on majority vote: n

$$1^{st} \text{ tree} \rightarrow n \left(\frac{1}{2} \right)$$

$$2^{nd} \text{ tree} \rightarrow y \left(\frac{1}{3} \right)$$

$$3^{rd} \text{ tree} \rightarrow n \left(\frac{1}{3} \right)$$

ⓑ Based on average probability:

$$1^{st} \Rightarrow \hat{f}_k(x) = \frac{1}{B} \sum_{b=1}^B P_b(x)$$

$$\hat{G}_{bag}(x) = \arg\max_k \hat{f}_k(x)$$

$$\hat{f}_{(y)}(x) = \frac{1}{3} \left[\frac{1}{2} + 1 + \frac{1}{3} \right] = \frac{5}{6} \times \frac{1}{3} = \frac{5}{9} \leftarrow \text{highest. } y \text{ is the predicted class}$$

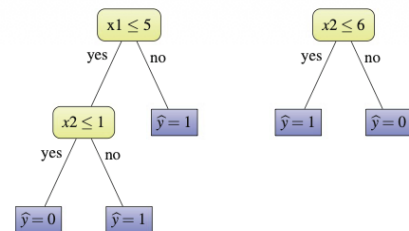
$$\hat{f}_{(n)}(x) = \frac{1}{3} \left[\frac{1}{2} + 0 + \frac{1}{2} \right] = \frac{1}{3}$$

$$\hat{f}_{(m)}(x) = \frac{1}{3} \left[\frac{1}{6} + 0 + \frac{1}{6} \right] = \frac{1}{9}$$

10.2 A random forest algorithm trained on a binary outcome, $y = 0$ or $y = 1$, and two x-variables. The random forest only has two trees which are shown in Fig. 10.6. Alongside each tree, the corresponding out-of-bag sample is shown in Tables 10.4 and 10.5.

- (a) Fill in values for “shuffled x1” for OOB sample 2. The solution is not unique; you are only asked to provide one solution. (For the purpose of this question, your “shuffled x1” should not be identical to that for OOB sample 1).
 (b) Based on the shuffled values of x1 in the tables, compute the importance of x1.
 (c) The out-of-bag samples in Tree 1 and Tree 2 have the same number of observations. Is that a coincidence or do all out-of-bag samples have the same number of observations? Explain briefly.

Fig. 10.6 Tree 1 built from bootstrap sample 1 and Tree 2 built from bootstrap sample 2



Tree 1

Tree 2

Table 10.4 OOB sample 1

ID	y	x1	x2	Shuffled x1
1	0	1	-1	6
8	1	8	15	10
3	1	3	8	8
14	1	6	2	3
5	0	10	3	1

Table 10.5 OOB sample 2

ID	y	x1	x2	Shuffled x1
23	1	12	-3	8
8	1	8	15	9
17	1	4	1	12
21	0	3	7	10
5	0	10	3	3

Tree 2

$$x_1 \text{ accuracy} = \frac{3}{5}$$

$$\text{shuffled } x_1 \text{ accuracy} = \frac{2}{5} \quad \text{Importance} = 0$$

$$\text{Tree 1 } x_1 \text{ accuracy} = \frac{4}{5}$$

$$\text{error} = \frac{1}{5}$$

$$\text{Shuffled error} = \frac{2}{5}, \text{ accuracy} = \frac{3}{5}$$

$$\text{Importance} = \frac{1}{5}$$

$$\text{Overall importance of } x_1 = \frac{0.2+0}{2} = 0.1$$

$$\textcircled{c} P \text{ not selecting } \frac{1}{n} \text{ observation } n \text{ times in a row} = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = 0.376.$$

Each observation has 0.376 chance of being excluded from a bootstrap sample (Unique observations)