CS 360: Machine Learning

Sara Mathieson, Sorelle Friedler

Spring 2024

HAVERFORD COLLEGE

Admin

- **Lab 3** due TONIGHT
- **Sorelle office hours** today 4-5pm in H110
- **Lab 4** posted tonight
- **Reading**: Geron Chap 7 (Ensembles) + optional reading up

Outline for Feb 15

- Finish AdaBoost

- Entropy and weighted entropy

- Gradient Boosting

Ensemble methods quiz (discuss with partner)

1. Briefly describe one advantage of an ensemble method over a single classifier.

   Usually decreases our testing error

2. When using ensemble methods, we want the base classifiers that have:
   - low variance, high bias
   - high variance, low bias
   - high variance, high bias

3. Ensembles often require us to run a base classifier on different training datasets. Name and briefly describe one method for generating multiple training datasets.

   Bootstrap (sampling with replacement)

4. (Bonus) for decision trees, is it possible to use the same feature twice on different paths from the root?

   Yes! (but not on the *same* path)

Outline for Feb 15

- Finish AdaBoost

- Entropy and weighted entropy

- Gradient Boosting

AdaBoost (adaptive boosting)

- **Train**
  - Start with all examples weighted equally (1/n)
  - for T iterations
    - Learn base classifier using weights
    - Change examples weights (up-weight incorrectly classified examples, down-weight correctly classified examples)
- **Test**
  - Get predictions from all base classifiers
  - Vote based on how well each classifier did during training

AdaBoost

set $w_i^{(1)} = \frac{1}{n}$ for $i = 1, \ldots, n$

for $t = 1 \ldots T$:

ⓐ fit classifier to weighted examples

true $y \in \{-1, 1\}$

binary classification

$h^{(t)}(\vec{x}) \in \{-1, 1\}$

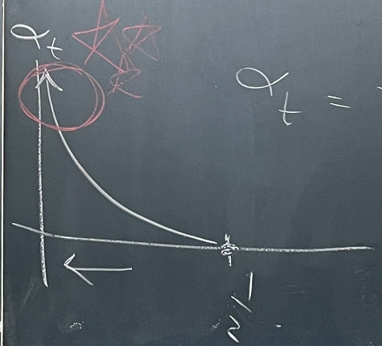
ⓑ compute weighted classification error

$$\varepsilon_t = \sum_{i=1}^{n} w_i^{(t)} \mathbb{1}\left(y_i \neq h^{(t)}(\vec{x}_i)\right)$$

$0 \le \varepsilon_t \le 1$

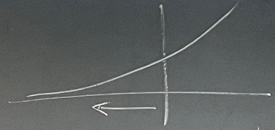
ⓒ compute classifier score

$$\alpha_t = \frac{1}{2} \log\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$$



as $\varepsilon_t \to 0$, $\alpha_t \to \infty$

(d) update weights



$$w_i^{(t+1)} = c_t \, w_i^{(t)} \exp\left(-y_i \alpha_t h^{(t)}(\vec{x}_i)\right)$$

- c_t: normalize so weights sum to 1
- y_i: $\{-1, 1\}$
- $h^{(t)}(\vec{x}_i)$: $\{-1, 1\}$

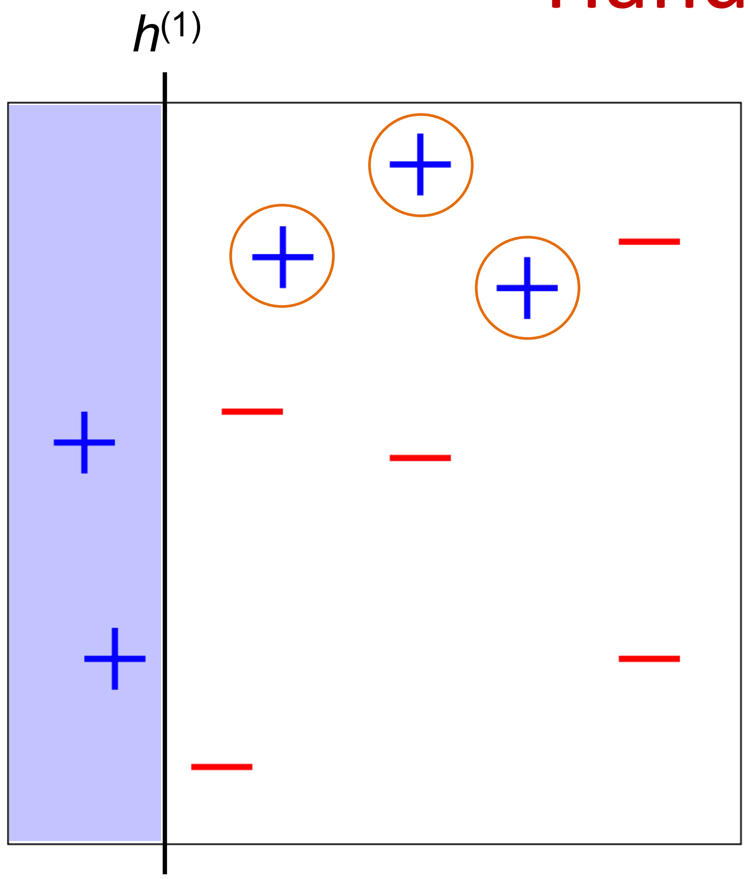
$$\exp\left(-y_i \alpha_t h^{(t)}(\vec{x}_i)\right) = \begin{cases} e^{-\alpha_t} & \text{if } y_i = h^{(t)}(\vec{x}_i) \quad \text{(downweight)} \\ e^{\alpha_t} & \text{if } y_i \neq h^{(t)}(\vec{x}_i) \quad \text{(upweight)} \end{cases}$$

after loop

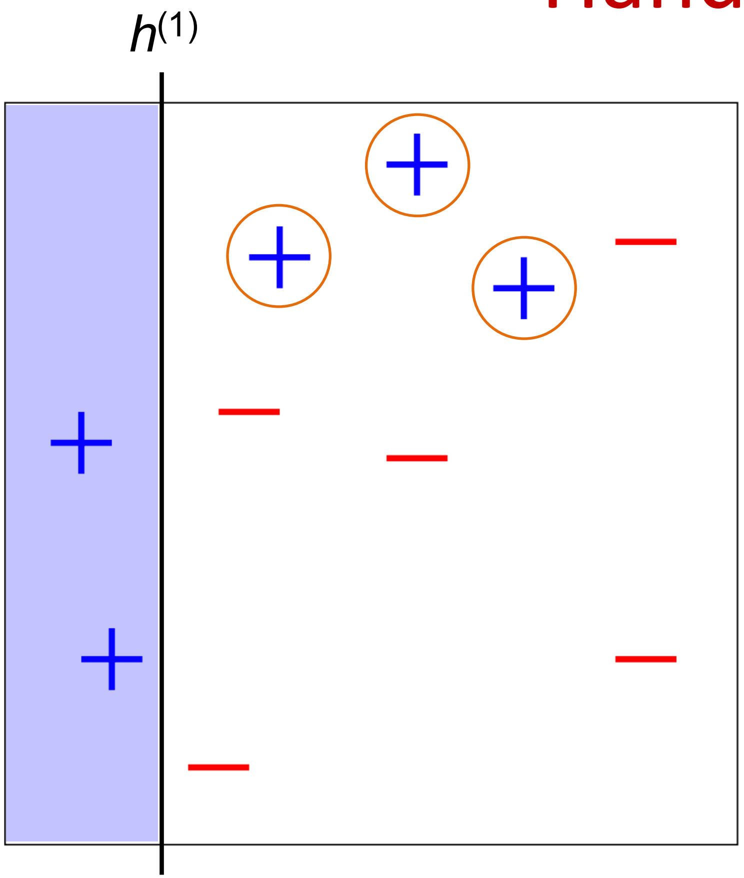
$$h(\vec{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h^{(t)}(\vec{x})\right)$$

Handout 8

Handout 8: Round 1



Handout 8: Round 1



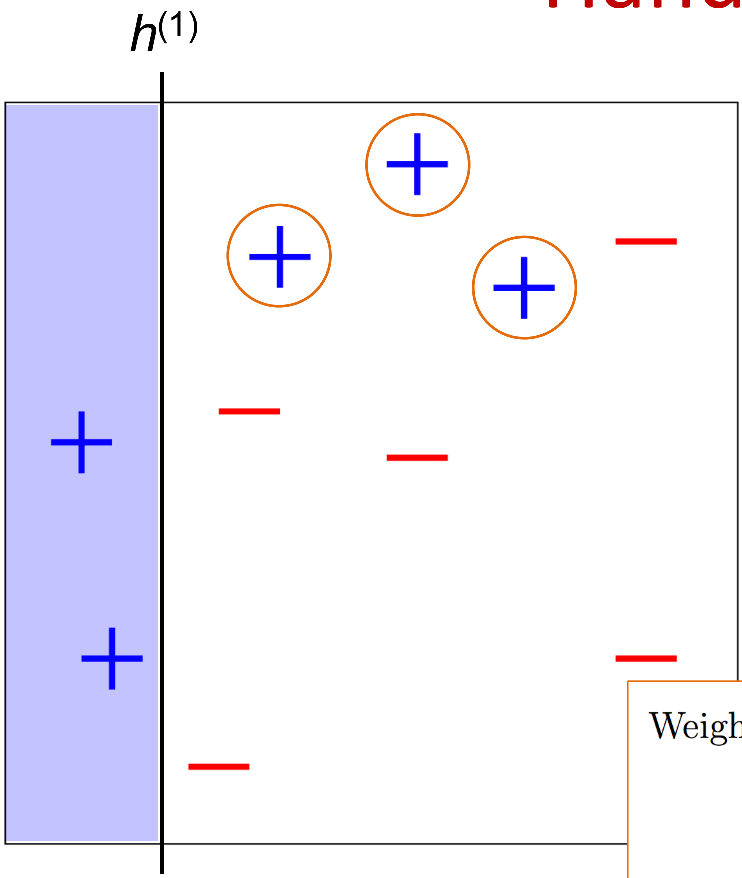
$w_i^{(1)} = \frac{1}{10}$ for all $i = 1, 2, \cdots, 10$.

$\epsilon_1 = \frac{3}{10}$ (three points incorrectly classified, all with weight $\frac{1}{10}$)

$\alpha_1 = \frac{1}{2} \ln \left(\frac{1 - \frac{3}{10}}{\frac{3}{10}} \right) = \ln \sqrt{\frac{7}{3}} \approx 0.42$

- correctly classified: $w_i^{(2)} = c_1 \cdot \frac{1}{10} \exp \left(- \ln \sqrt{\frac{7}{3}} \right)$
- incorrectly classified: $w_i^{(2)} = c_1 \cdot \frac{1}{10} \exp \left(\ln \sqrt{\frac{7}{3}} \right)$

Handout 8: Round 1



$w_i^{(1)} = \frac{1}{10}$ for all $i = 1, 2, \cdots, 10$.

$\epsilon_1 = \frac{3}{10}$ (three points incorrectly classified, all with weight $\frac{1}{10}$)

$\alpha_1 = \frac{1}{2} \ln \left(\frac{1 - \frac{3}{10}}{\frac{3}{10}} \right) = \ln \sqrt{\frac{7}{3}} \approx 0.42$

- correctly classified: $w_i^{(2)} = c_1 \cdot \frac{1}{10} \exp \left(- \ln \sqrt{\frac{7}{3}} \right)$
- incorrectly classified: $w_i^{(2)} = c_1 \cdot \frac{1}{10} \exp \left(\ln \sqrt{\frac{7}{3}} \right)$

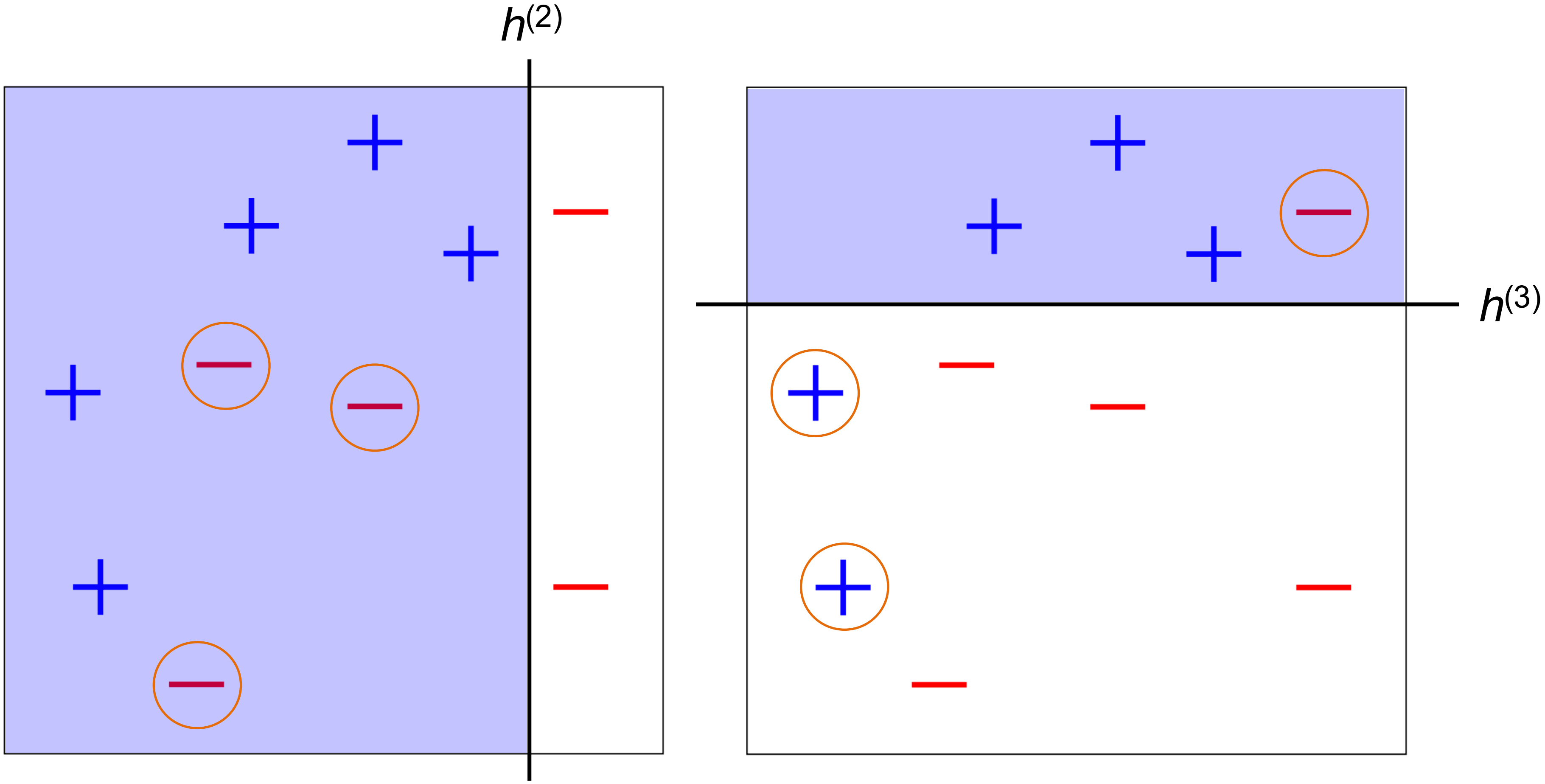
Weights must sum to 1, $\Rightarrow$

$$7 \cdot \frac{c_1}{10} \exp \left(- \ln \sqrt{\frac{7}{3}} \right) + 3 \cdot c_1 \cdot \frac{1}{10} \exp \left(\ln \sqrt{\frac{7}{3}} \right) = 1$$

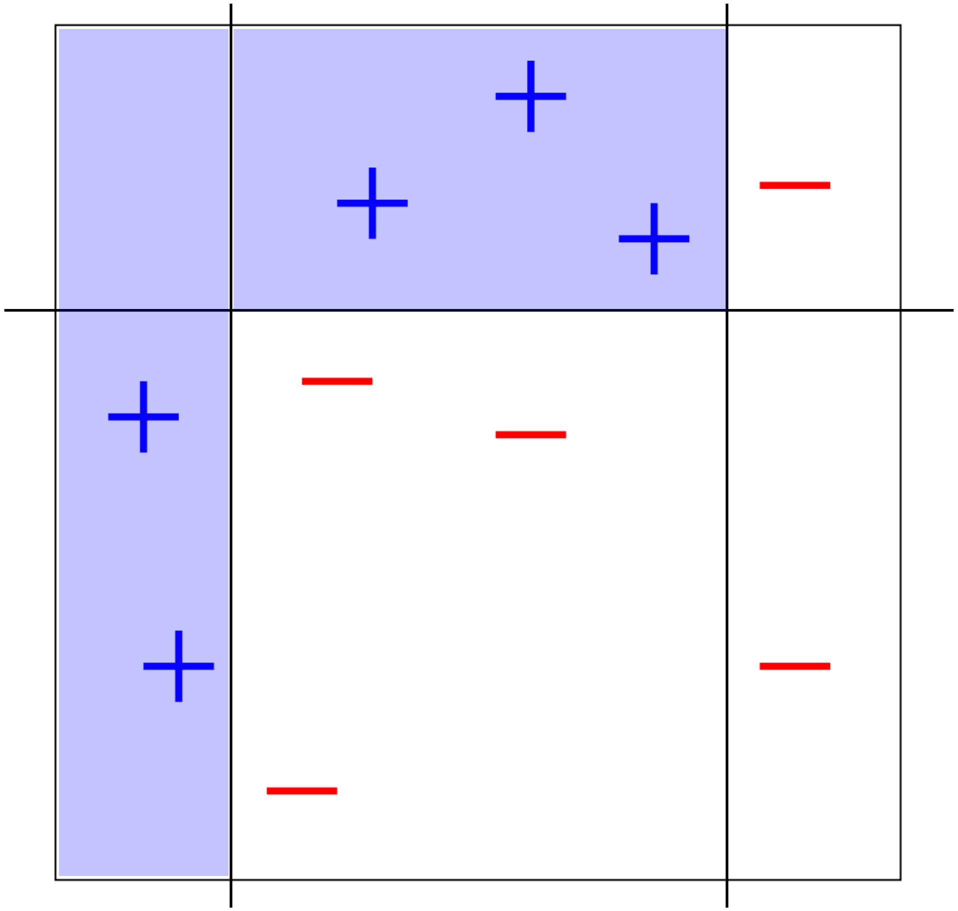
$$\Rightarrow \quad c_1 = \frac{5}{\sqrt{21}}$$

- correctly classified: $w_i^{(2)} = \frac{5}{\sqrt{21}} \cdot \frac{1}{10} \sqrt{\frac{3}{7}} = \frac{1}{14}$ decrease!
- incorrectly classified: $w_i^{(2)} = \frac{5}{\sqrt{21}} \cdot \frac{1}{10} \sqrt{\frac{7}{3}} = \frac{1}{6}$ increase!

Handout 8: Round 2 & 3 (exercise!)



Handout 8: final classifier



$$h(\boldsymbol{x}) = \text{sign}\left(0.42 \cdot h^{(1)}(\boldsymbol{x}) + 0.65 \cdot h^{(2)}(\boldsymbol{x}) + 0.92 \cdot h^{(3)}(\boldsymbol{x})\right)$$

Outline for Feb 15

- Finish AdaBoost

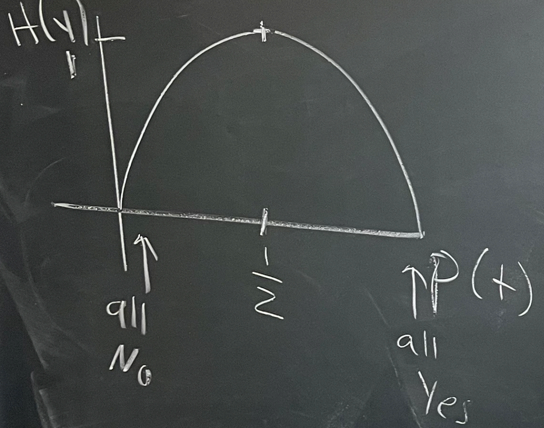
- Entropy and weighted entropy

- Gradient Boosting

Entropy + Info Gain

No Yes
[5, 9]

$$H(y) = -\sum_{c \in vals(y)} p(c) \log_2 p(c) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right)$$



Conditional Entropy

$$H(y|x) = \sum_{v \in vals(x)} P(x=v)\,H(y|x=v) = \frac{5}{14}H(y|x=$$

x ↑ one feature

[3, 2]

⇓

predict No

[5, 9]

Outlook

sun — [illegible], 2] → predict no

rain — [2, 3] → predict yes

overcast — [0, 4] → predict yes

$$H(Y \mid X=v) = -\sum_{c \in vals(Y)} P(Y=c \mid X=v) \log_2 P(Y=c \mid X=v)$$

$$P(Y=c \mid X=v) = \frac{count(Y=c, X=v)}{count(X=v)} = \frac{3}{5}$$

$$P(B \mid A) = \frac{P(A, B)}{P(A)}$$

$c = no$, $v = sun$

$$\ldots \mid X = sun) + \frac{5}{14} H(Y \mid X = rain) + \frac{4}{14} H(Y \mid X = overcast)$$

Weighted Examples

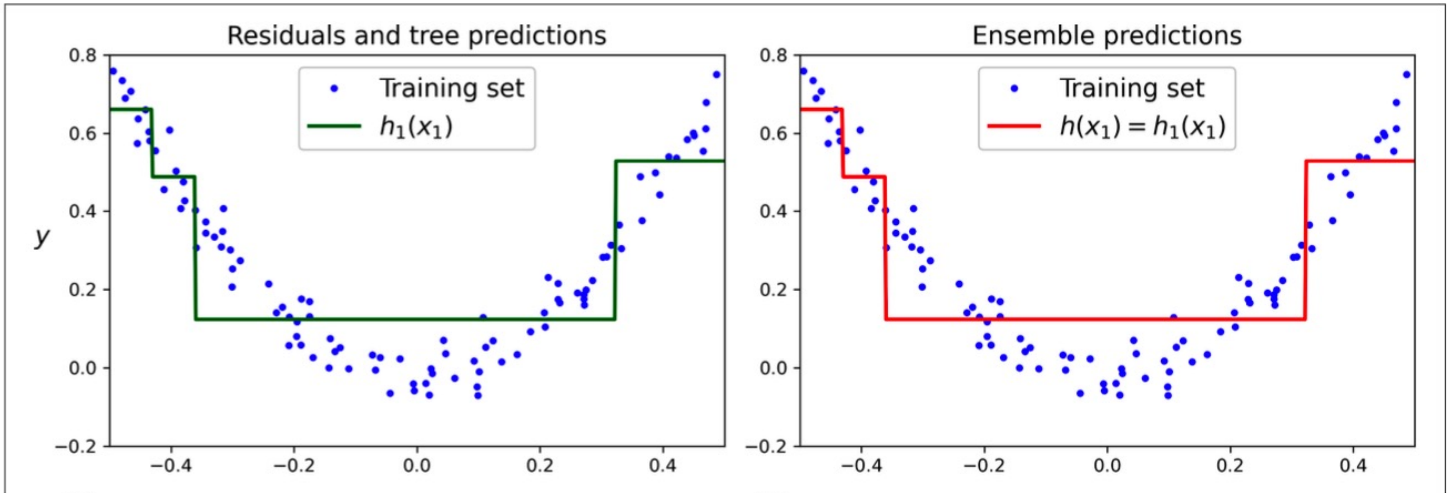
$$P(Y=c \mid X_j = v) = \frac{\sum_{i=1}^{n} w_i^{(t)} \mathbb{1}(y_i = c, x_{ij} = v)}{\sum_{i=1}^{n} w_i^{(t)} \mathbb{1}(x_{ij} = v)}$$

leaves

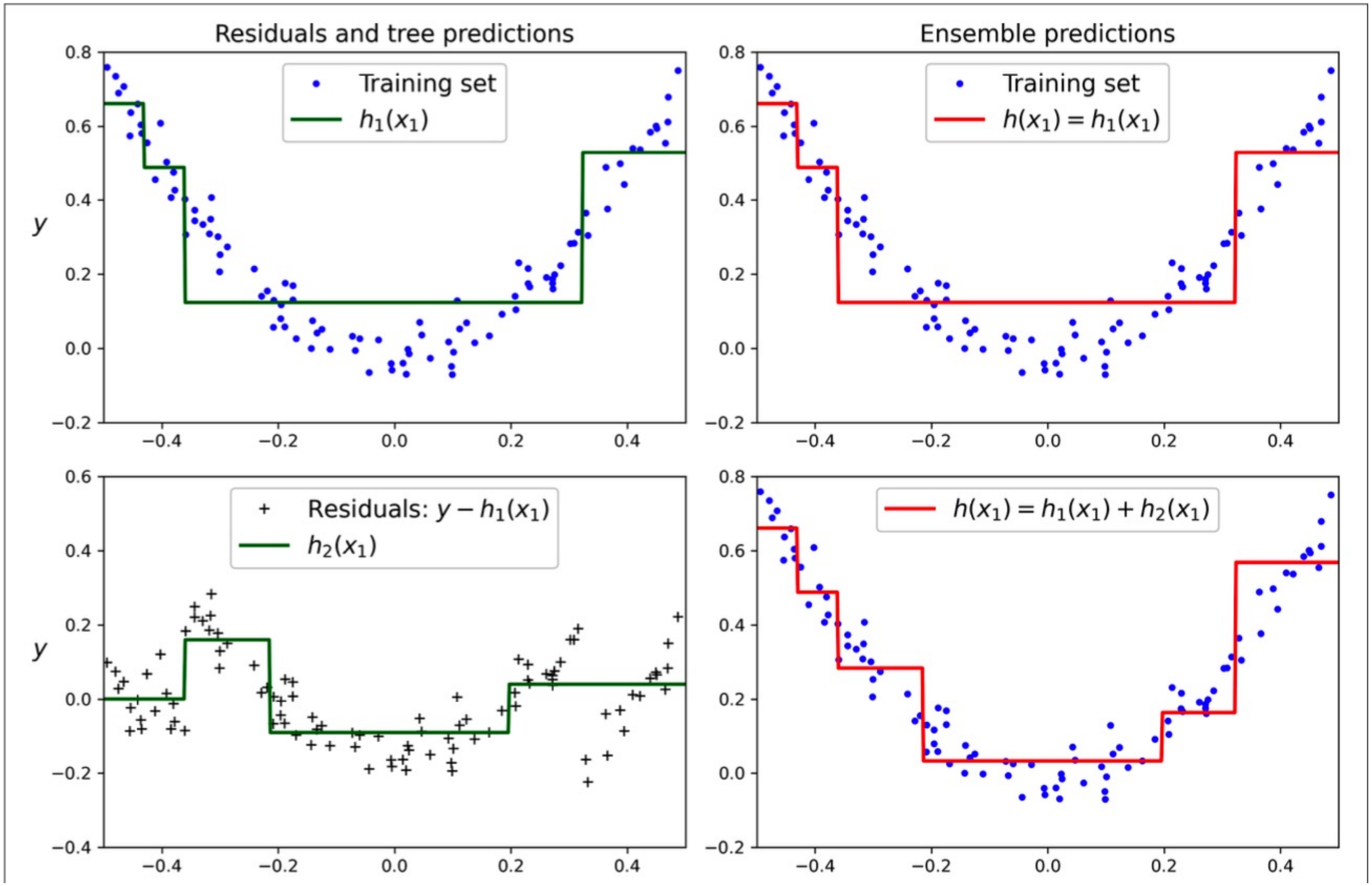
$$P(\text{leaf label is } 1) = \frac{\sum_{i=1}^{n} w_i^{(t)} \mathbb{1}(y_i = 1)}{\sum_{i \text{ in leaf}} w_i^{(t)}}$$

Outline for Feb 15

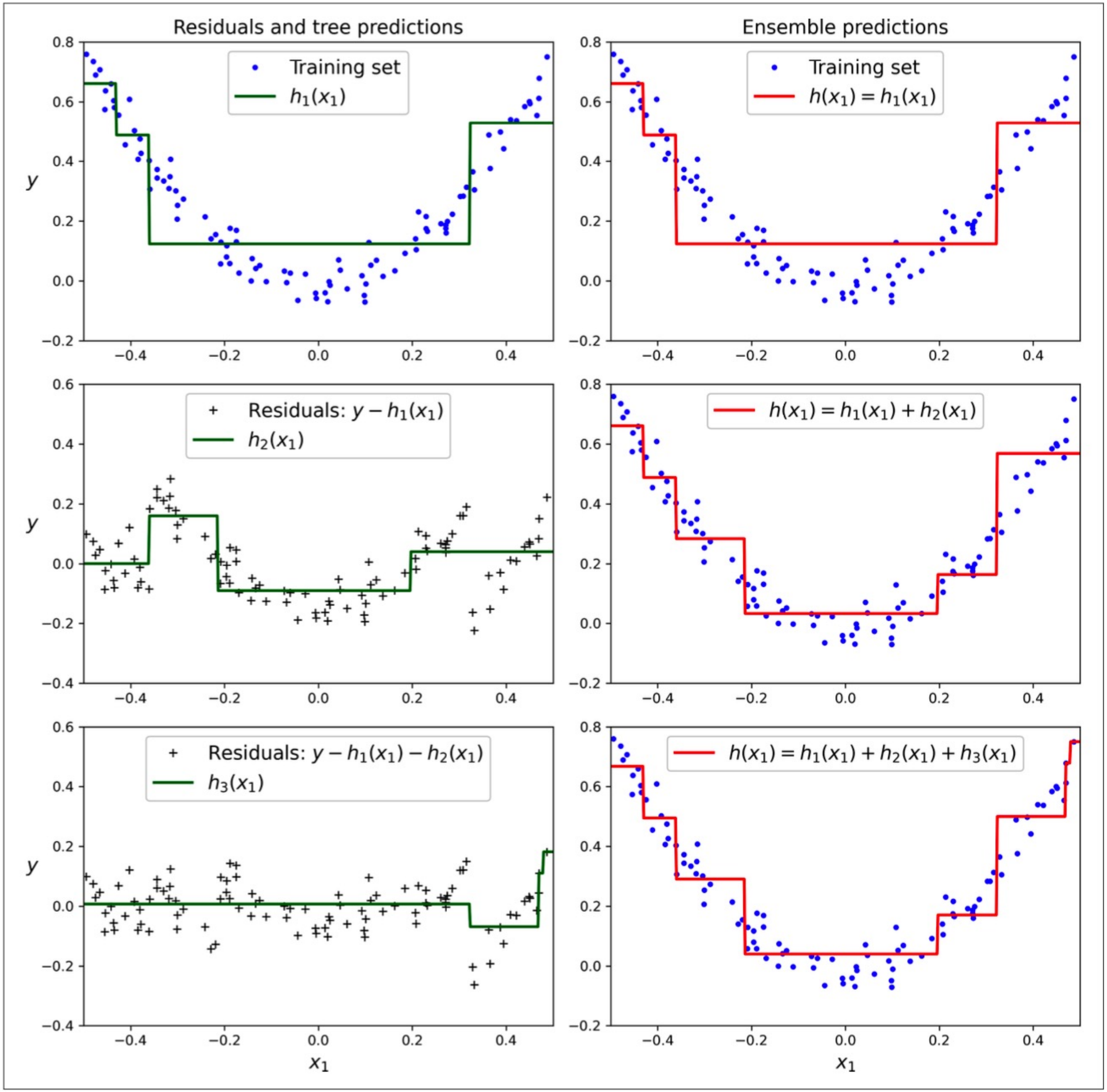
- Finish AdaBoost
- Entropy and weighted entropy
- Gradient Boosting



Gradient Boosting example (regression)



Gradient Boosting example (regression)



Gradient Boosting example (regression)

Geron: Chap 7

Gradient Boosting

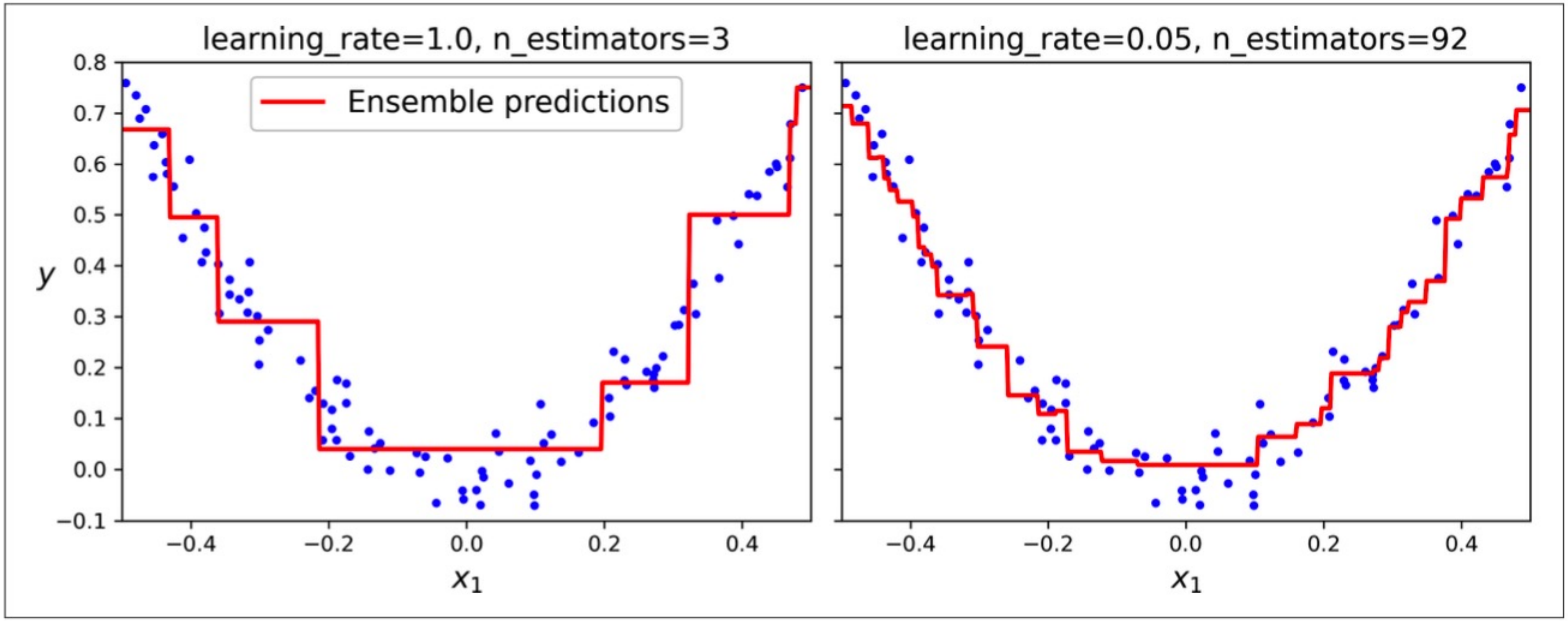


Figure 7-10. GBRT ensembles with not enough predictors (left) and just enough (right)