Decision trees

Readings for today

- Chapter 8: Tree-based methods. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer

Topics

1. Decision trees
2. Classification trees
3. Trees vs. linear models

Decision trees

Interpretability

Linear Models: $\hat{y}_i = \hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

unit change in y that happens with each unit change in x

Assumes:
1. $Y \sim N(\mu, \sigma)$
2. X is iid

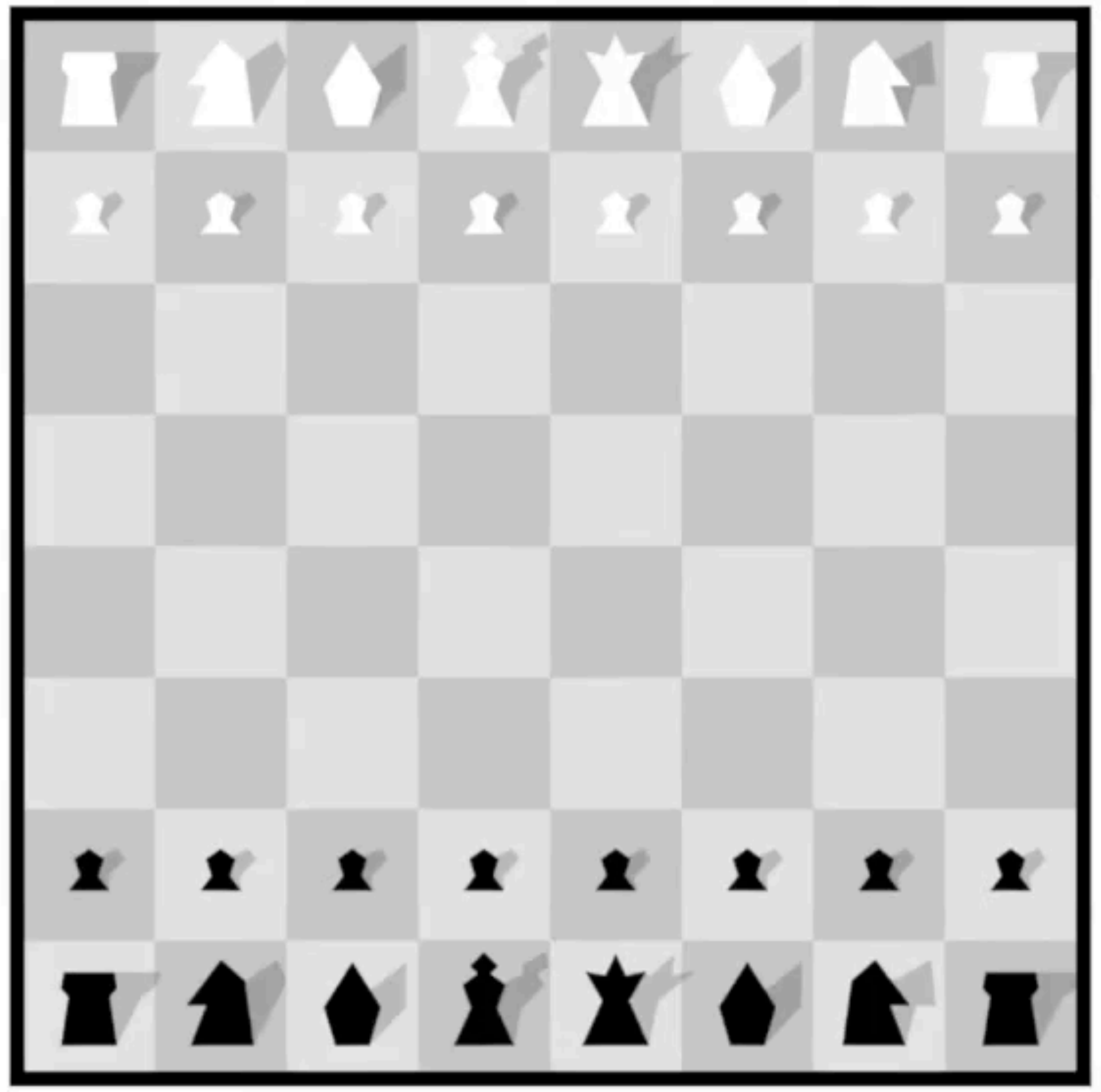
As soon as your data violates these assumptions, the interpretability of your effects is impaired.

(James et al. 2013)

Tree methods

Goal: Treat $f(X)$ as a series of "if/then" rules.

Approach: Identify the best decision rules for parsing X to explain Y.



Google DeepMind

Non-parametric way of identifying subregions in X relevant to Y.

(James et al. 2013)

Recursive Binary Splitting

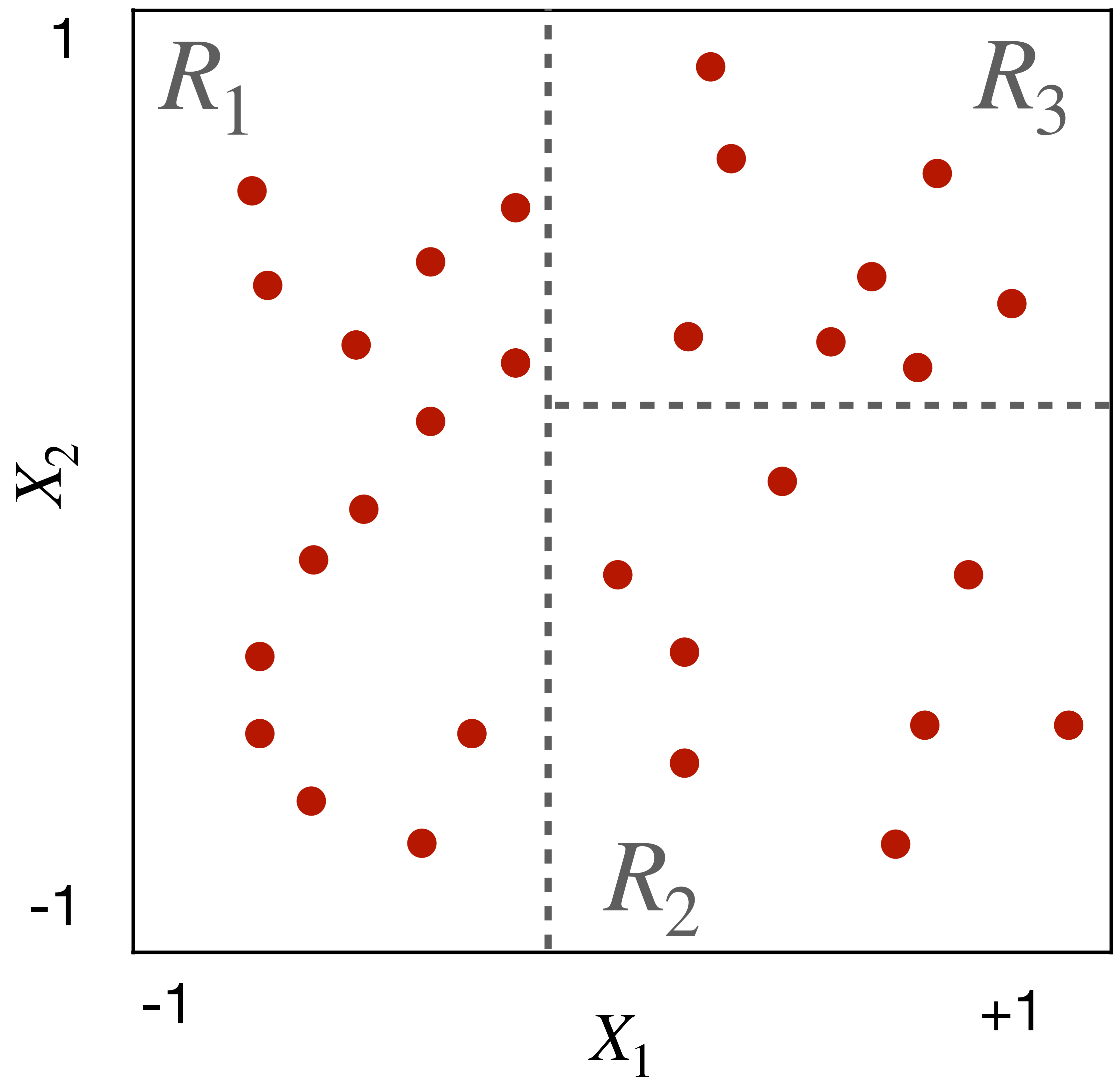
<u>Steps</u>: 1. Divide X recursively into j subregions $R_1, \ldots, R_j$.

2. Assign $\hat{y}$ as the mode of every y_i in region R_m

$X_1 > -0.13$

$X_2 < 0.21$

$$\hat{y}_{R_1} = \frac{1}{m}\sum_{i=1}^{m} y_i^{R_1} \qquad \hat{y}_{R_2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{R_2} \qquad \hat{y}_{R_3} = \frac{1}{k}\sum_{i=1}^{k} y_i^{R_3}$$



(James et al. 2013)

Objective function

<u>Goal:</u> $$\min\left(\sum_{j=1}^{J}\sum_{i\in R_j}^{n}(y_i - \hat{y}_{R_j})^2\right) \rightarrow RSS$$

$\hat{y}_{R_j}$ is the mean/modal value of all observations in region R_j.

<u>Top-down:</u> Split most relevant feature first and then work your way down successively.

<u>Greedy:</u> *Only* the "best" split (i.e., split that explains most variance) is chosen at each step.

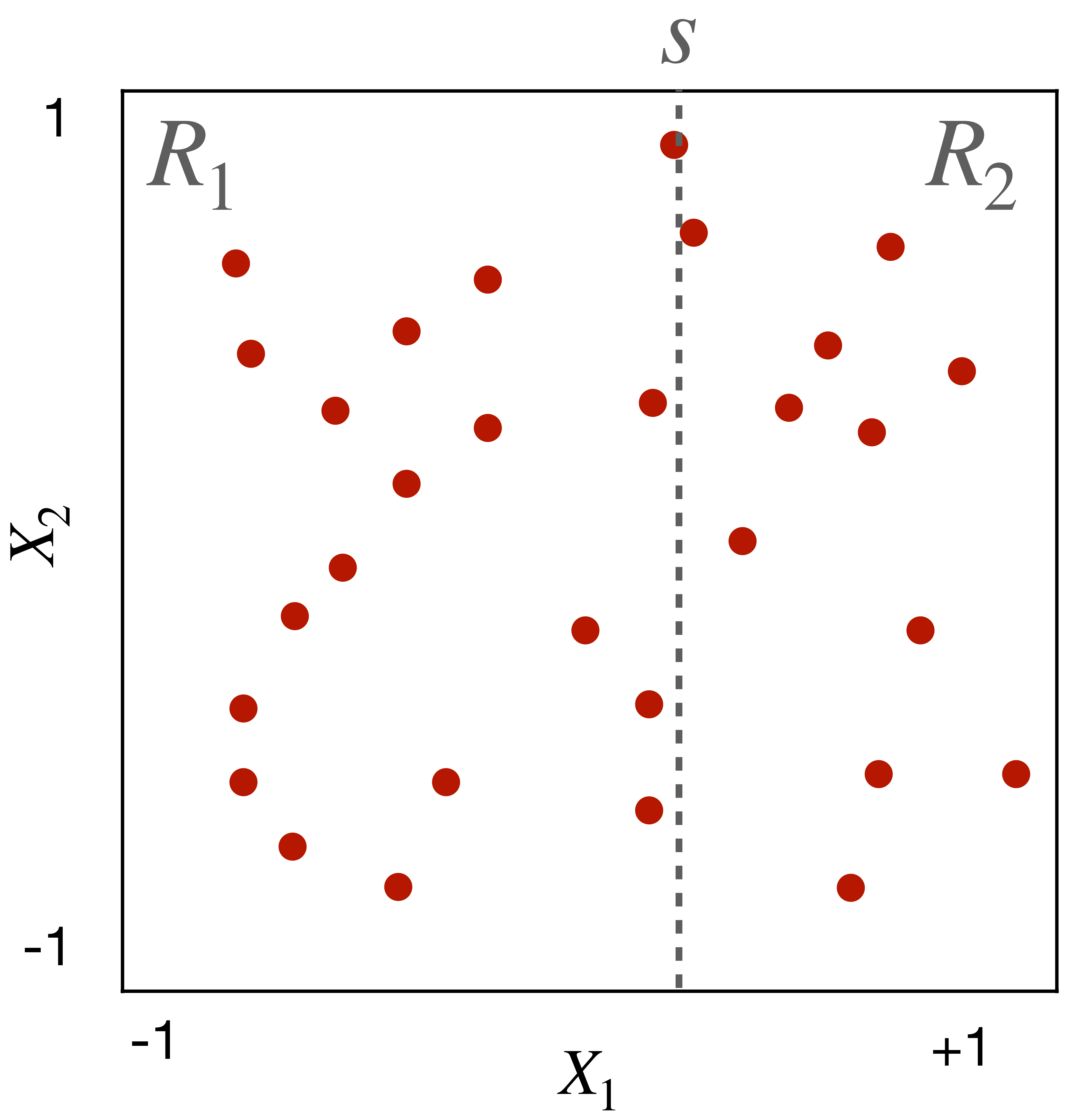
(James et al. 2013)

Splitting

Choosing the right split means finding the best border, s, that minimizes two residual estimates.

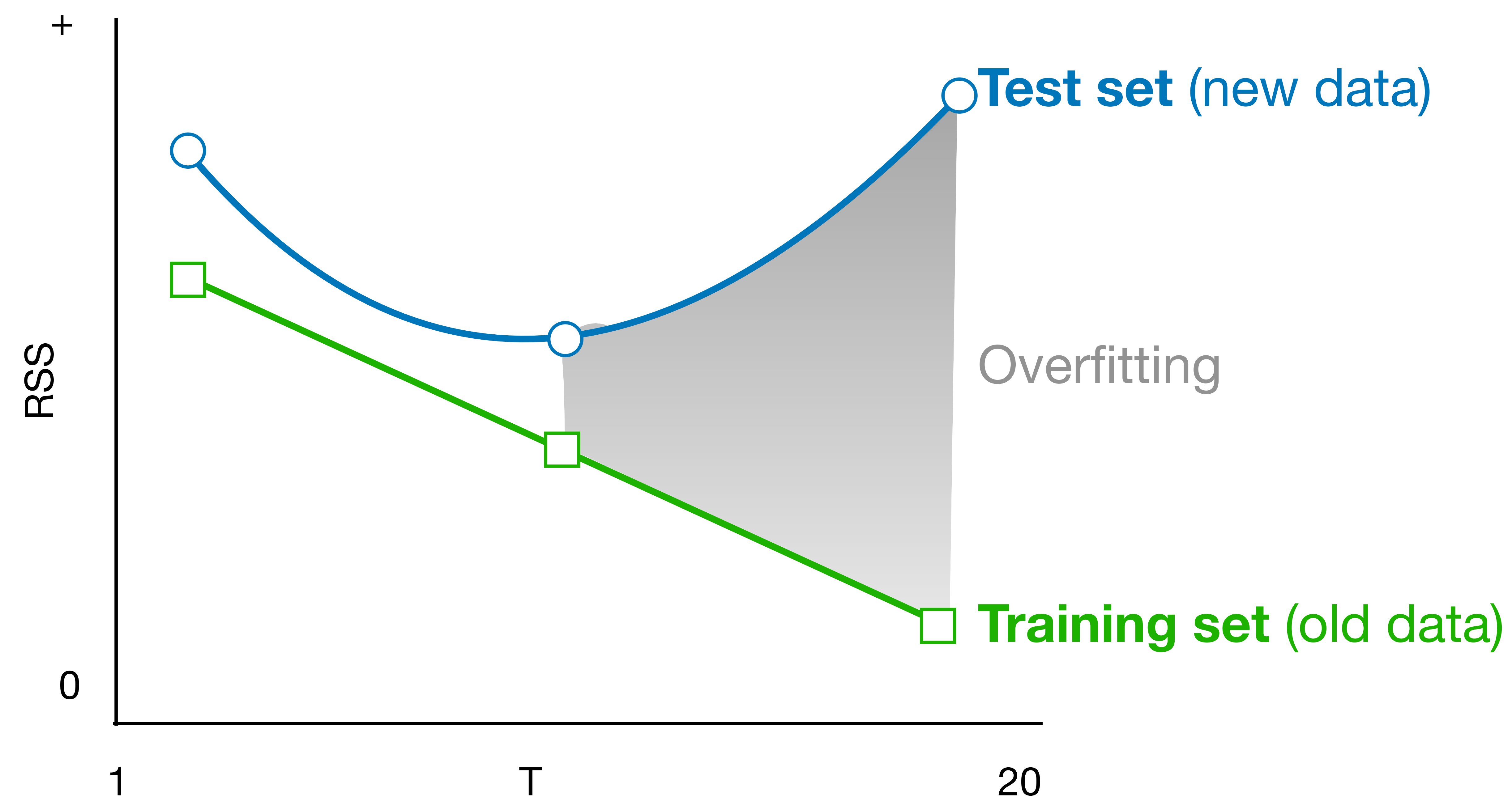
$$\overbrace{R_1(j,s) = \{X|X_j < s\}}^{\text{Region 1}} \qquad \overbrace{R_2(j,s) = \{X|X_j \geq s\}}^{\text{Region 2}}$$

<u>minimize:</u> $$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$



(James et al. 2013)

Bias-variance tradeoff



Number of regions you choose to split along, T, determines the flexibility of your model

$$\uparrow T = \uparrow \text{variance}$$

<u>Cost-Complexity Tuning:</u>

sparsity parameter to be tuned

$$\min\left(\sum_{m=1}^{T} \sum_{i:x_i \in R_m}^{n} (y_i - \hat{y}_{R_m})^2 + \hat{\alpha}\,|T|\right)$$

$$\uparrow \alpha = \downarrow T$$

(James et al. 2013)

Classification trees

Classification trees

Goal: Create a tree for qualitative (categorical) Y.

classification error rate

$$E_m = 1 - \max_k(\hat{p}_{mk})$$

Proportion of training observations in region m that are members of class k.

Objective: $\min(\sum_{j=1}^{J} \sum_{i \in m}^{K} E_m)$

Minimize the classification error across classes.

(James et al. 2013)

Node purity

Node purity: Measure of uniqueness of categories in each R_m.

1. Gini Index (G): $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$

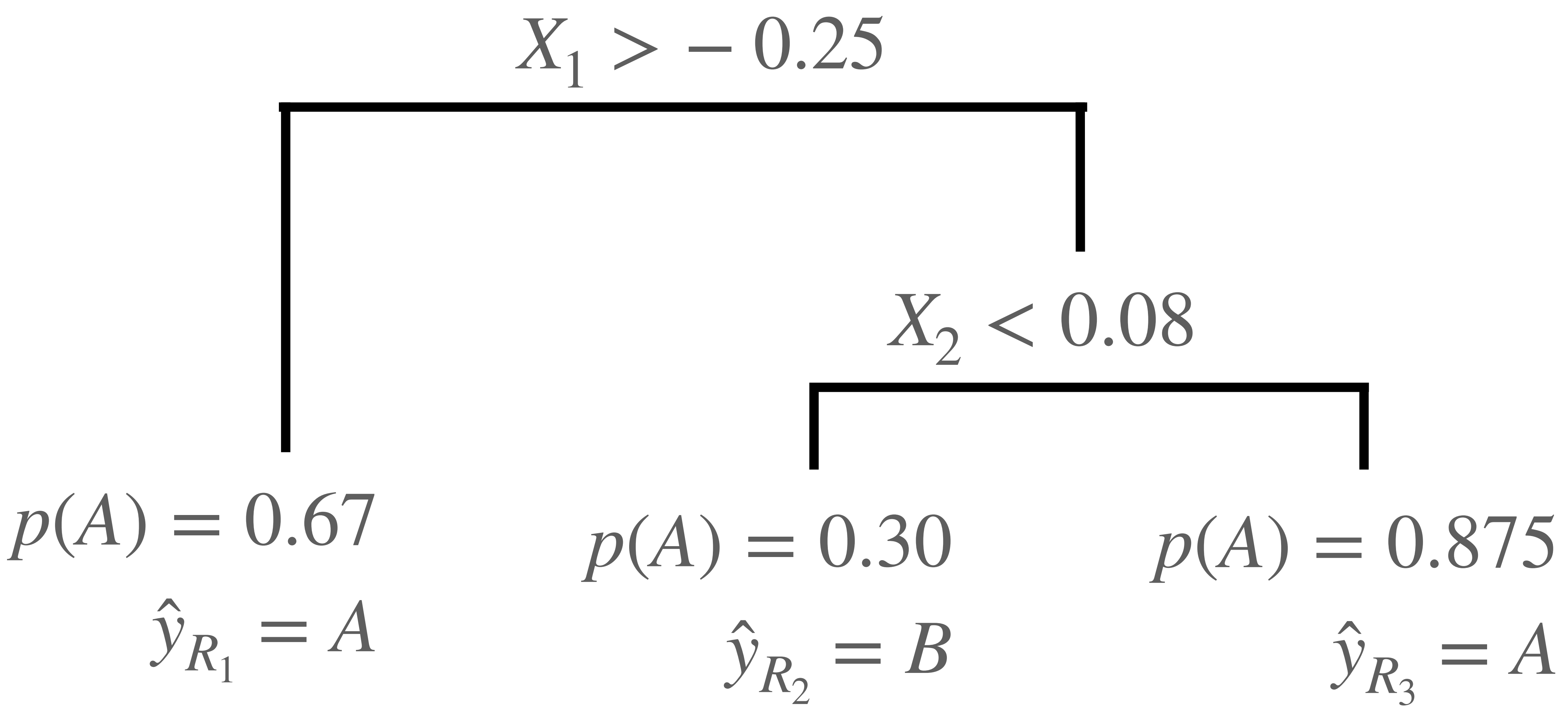
2. Cross-entropy (D): $D = \sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$

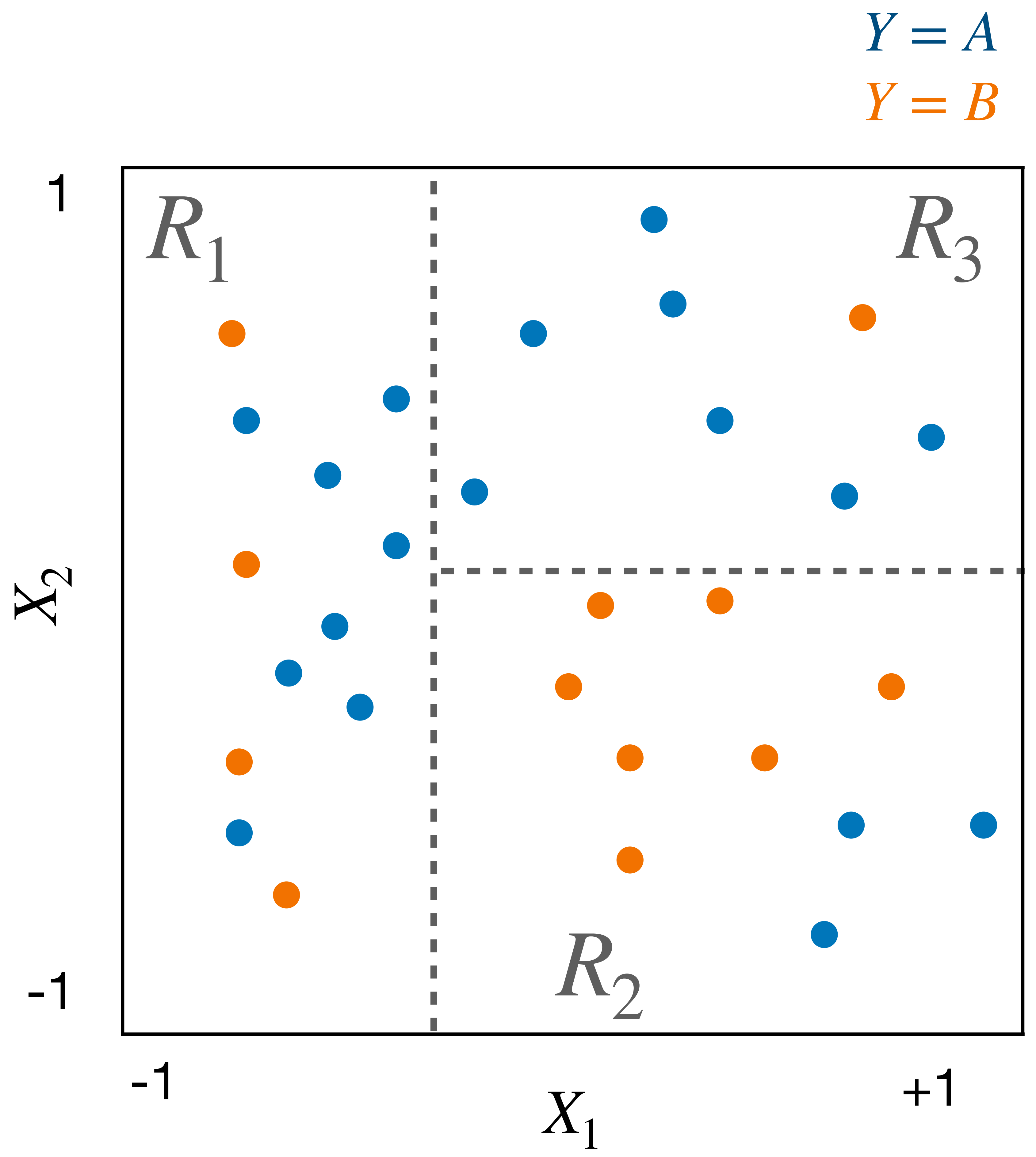
- $\downarrow G = \uparrow$ purity
- $\downarrow D = \uparrow$ purity

Both measures determine whether a node/region contains predominantly members of one class.

(James et al. 2013)

Example: 2 group classification



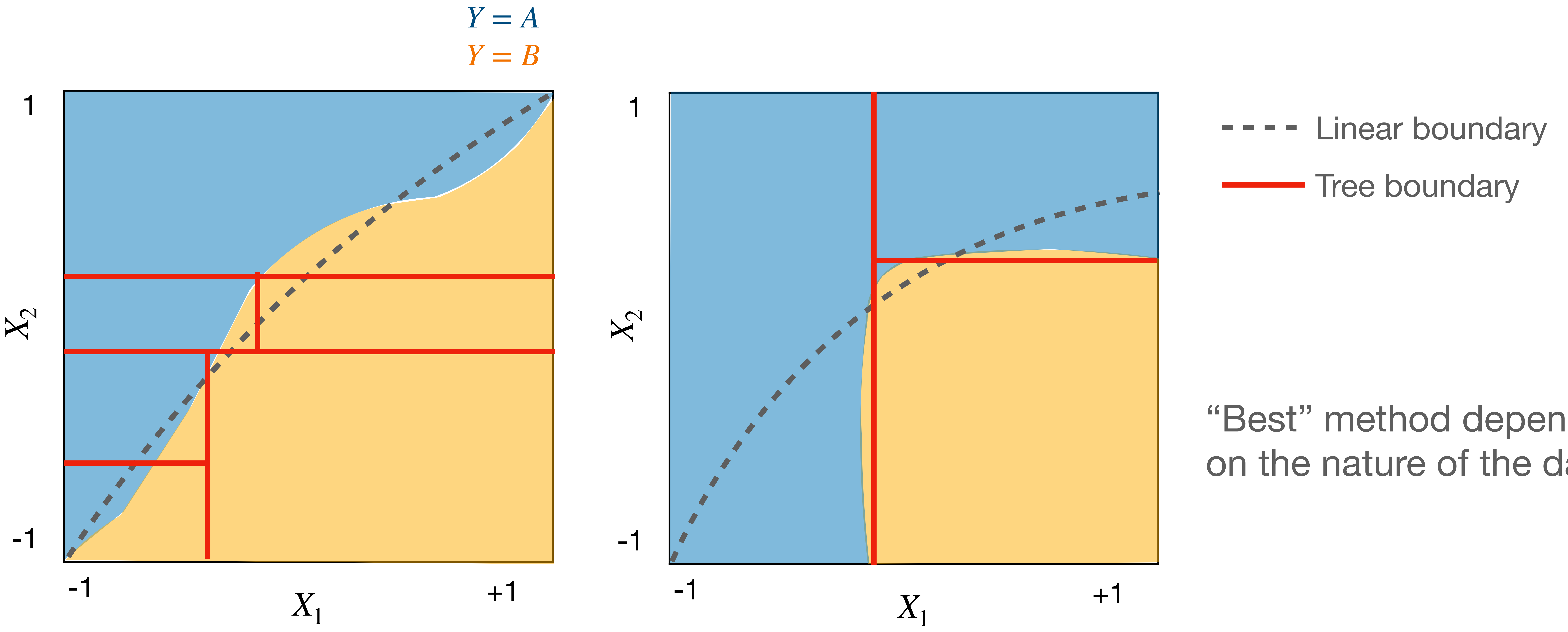


$G = 0.54$ ⟵ Moderate node purity.

(James et al. 2013)

Trees vs. linear models

Trees vs. linear models



"Best" method depends on the nature of the data.

(James et al. 2013)

The pros and cons

Advantages:

- Easy to explain
- Easily visualized
- No need for dummy coding
- “Mirrors” human heuristics
- Applicable to data that does not meet the assumptions of linear or parametric models.

Disadvantages:

- Lower predictive accuracy.
- Non-robust

small changes in data have huge impacts on model fits.

(James et al. 2013)

Take home message

- Decision trees are intuitive, but finicky, methods for determining $X \rightarrow Y$ relationships that are especially powerful in contexts where the data does not meet the assumptions of standard linear models.