Growing decision trees

STAT 4710

October 26, 2023

Rolling into a new unit!

- ✓ **Unit 1:** R for data mining
- ✓ **Unit 2:** Prediction fundamentals
- ✓ **Unit 3:** Regression-based methods
- **Unit 4:** Tree-based methods
- **Unit 5:** Deep learning

Lecture 1: Growing decision trees

Lecture 2: Tree pruning and bagging

Lecture 3: Random forests

Lecture 4: Boosting

Lecture 5: Unit review and quiz in class

Leaving the land of linearity

Leaving the land of linearity

Most methods covered so far based on $\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_{p-1} X_{p-1}$:

- Linear regression
- Logistic regression
- Ridge, lasso, elastic net

Leaving the land of linearity

Most methods covered so far based on $\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_{p-1} X_{p-1}$:

- Linear regression
- Logistic regression
- Ridge, lasso, elastic net

Notable exception: K-nearest neighbors (recall Unit 2)

Leaving the land of linearity

Most methods covered so far based on $\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_{p-1} X_{p-1}$:

- Linear regression
- Logistic regression
- Ridge, lasso, elastic net

Notable exception: K-nearest neighbors (recall Unit 2)

In Unit 4 we will leave the land of linearity.

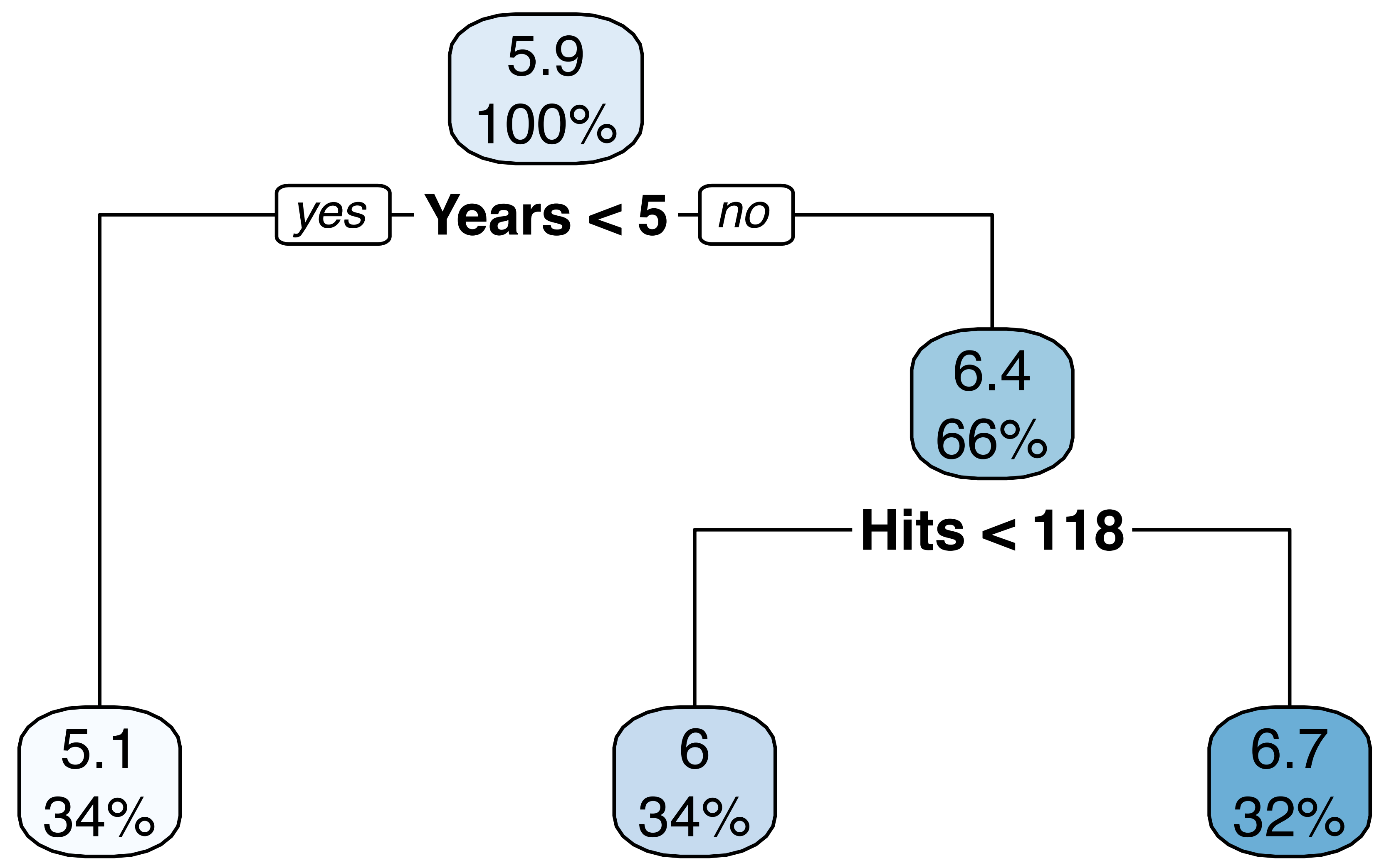
Entering the land of trees and forests

Entering the land of trees and forests

Decision trees (lectures 1 and 2) are predictive models based on recursively partitioning the feature space.

Entering the land of trees and forests

Decision trees (lectures 1 and 2) are predictive models based on recursively partitioning the feature space.

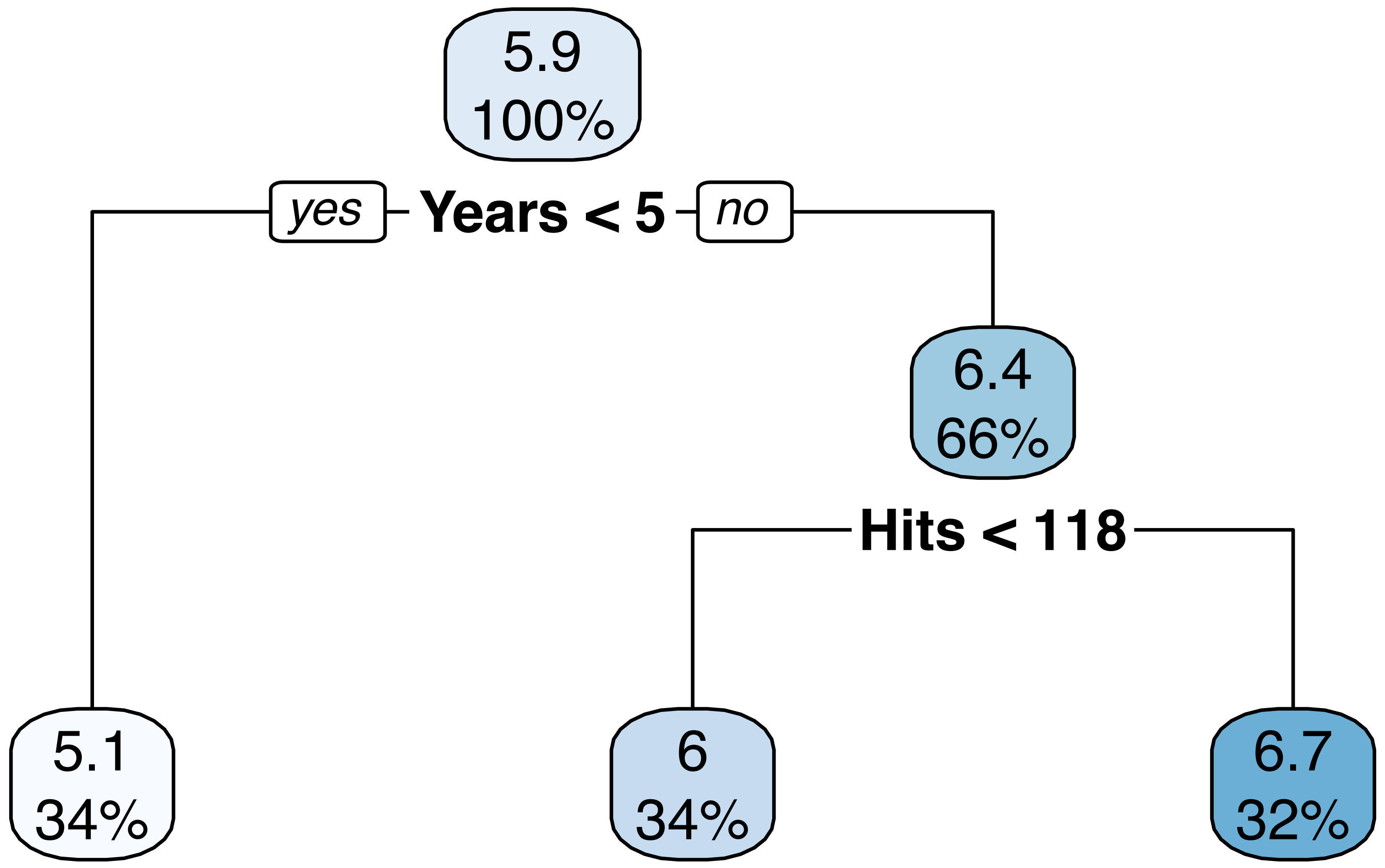


Predicting baseball players' salaries based on years played and number of hits in the previous year.

Entering the land of trees and forests

Decision trees (lectures 1 and 2) are predictive models based on recursively partitioning the feature space.

Their prediction rules can be nicely illustrated and are very interpretable.



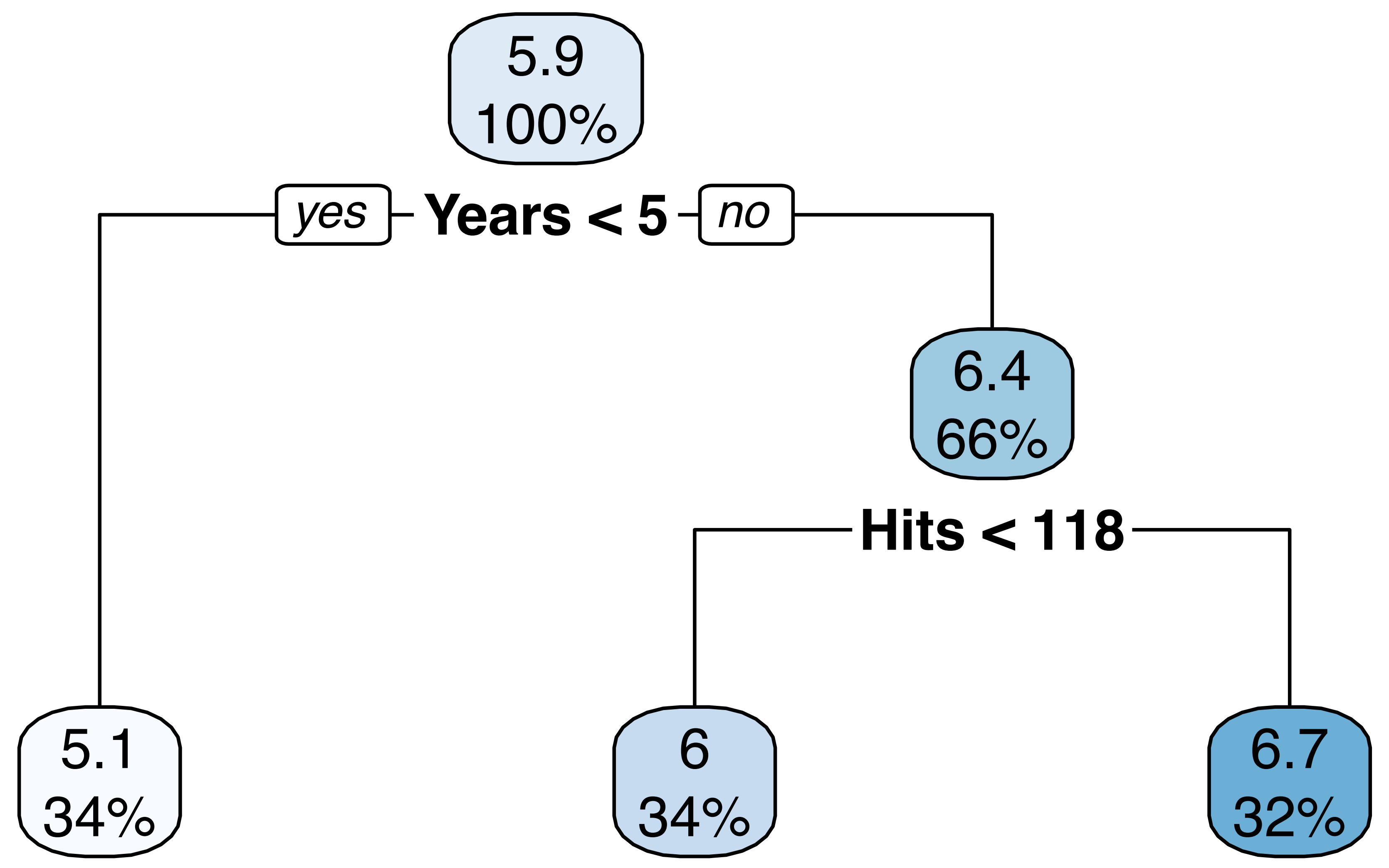
Predicting baseball players' salaries based on years played and number of hits in the previous year.

Entering the land of trees and forests

Decision trees (lectures 1 and 2) are predictive models based on recursively partitioning the feature space.

Their prediction rules can be nicely illustrated and are very interpretable.

However, trees are somewhat unstable and do not give the best prediction performance.



Predicting baseball players' salaries based on years played and number of hits in the previous year.

Entering the land of trees and forests

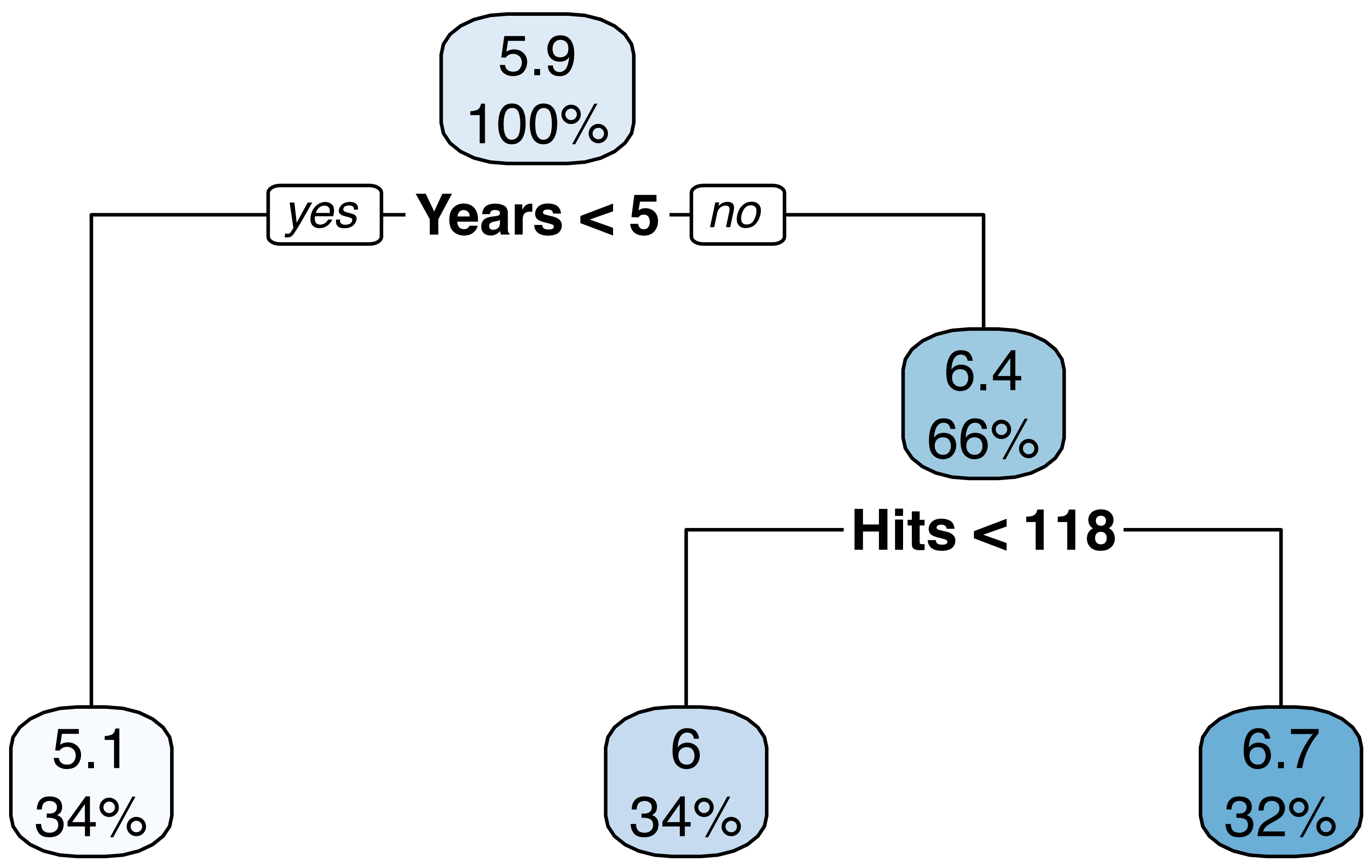
Decision trees (lectures 1 and 2) are predictive models based on recursively partitioning the feature space.

Their prediction rules can be nicely illustrated and are very interpretable.

However, trees are somewhat unstable and do not give the best prediction performance.

Nevertheless, trees can be used as building blocks for state-of-the-art prediction performance:

- Random forests (lecture 3)
- Boosting (lecture 4)



Predicting baseball players' salaries based on years played and number of hits in the previous year.

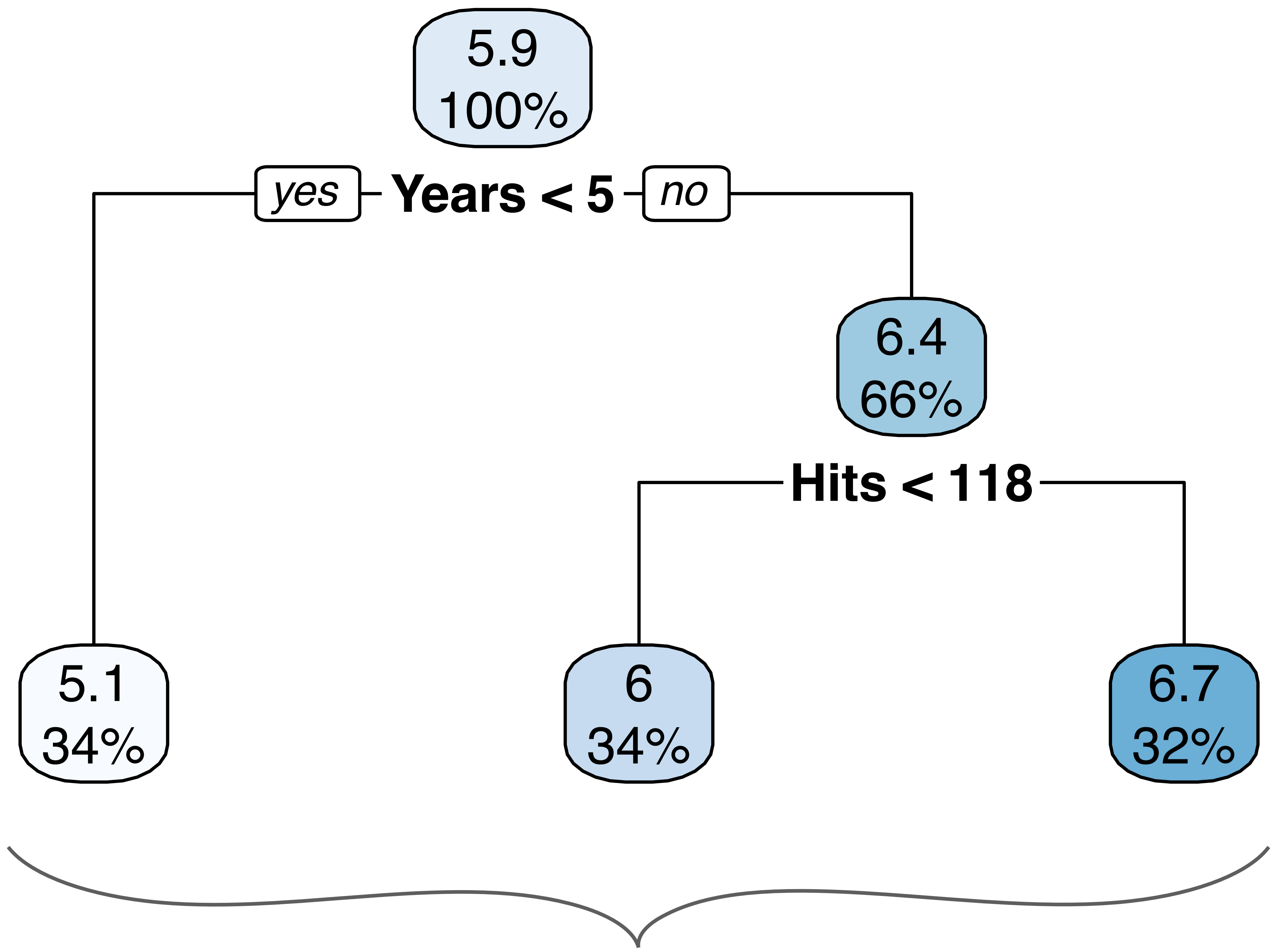
Hitters data

Major League Baseball Data from the 1986 and 1987 seasons.

- Observations: 322 MLB players
- Response: Salary (1987 annual salary on opening day in thousands of dollars)
- Features: Assists, AtBat,…,Hits,…,Years (19 total)

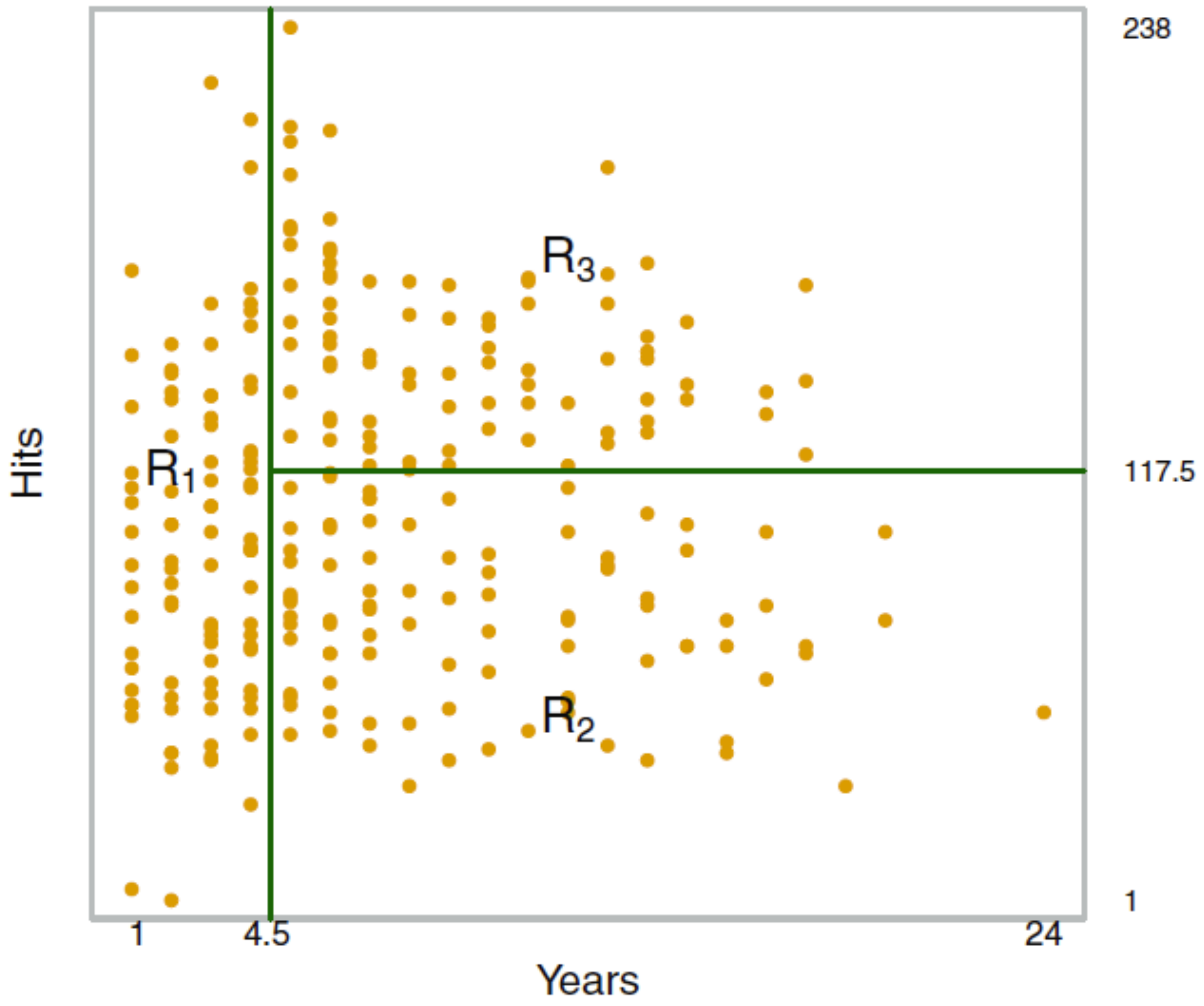
Image credit: DALL-E 3

Tree $\Longleftrightarrow$ Partition into nested, axis-aligned rectangles



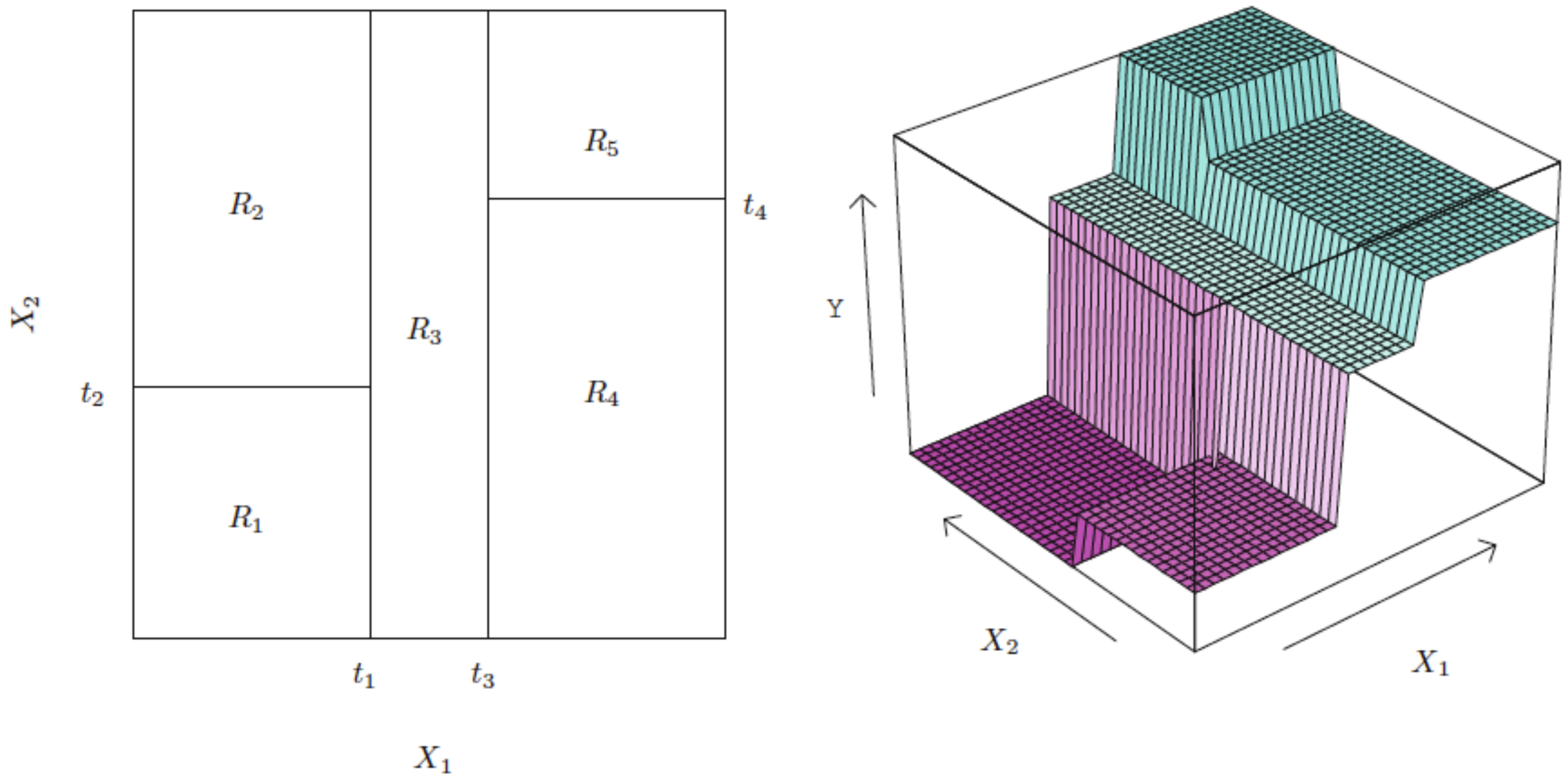
Nodes without any descendants called
leaf nodes or terminal nodes (equivalent)

$\Longleftrightarrow$



Each terminal node corresponds to a rectangular region of feature space.

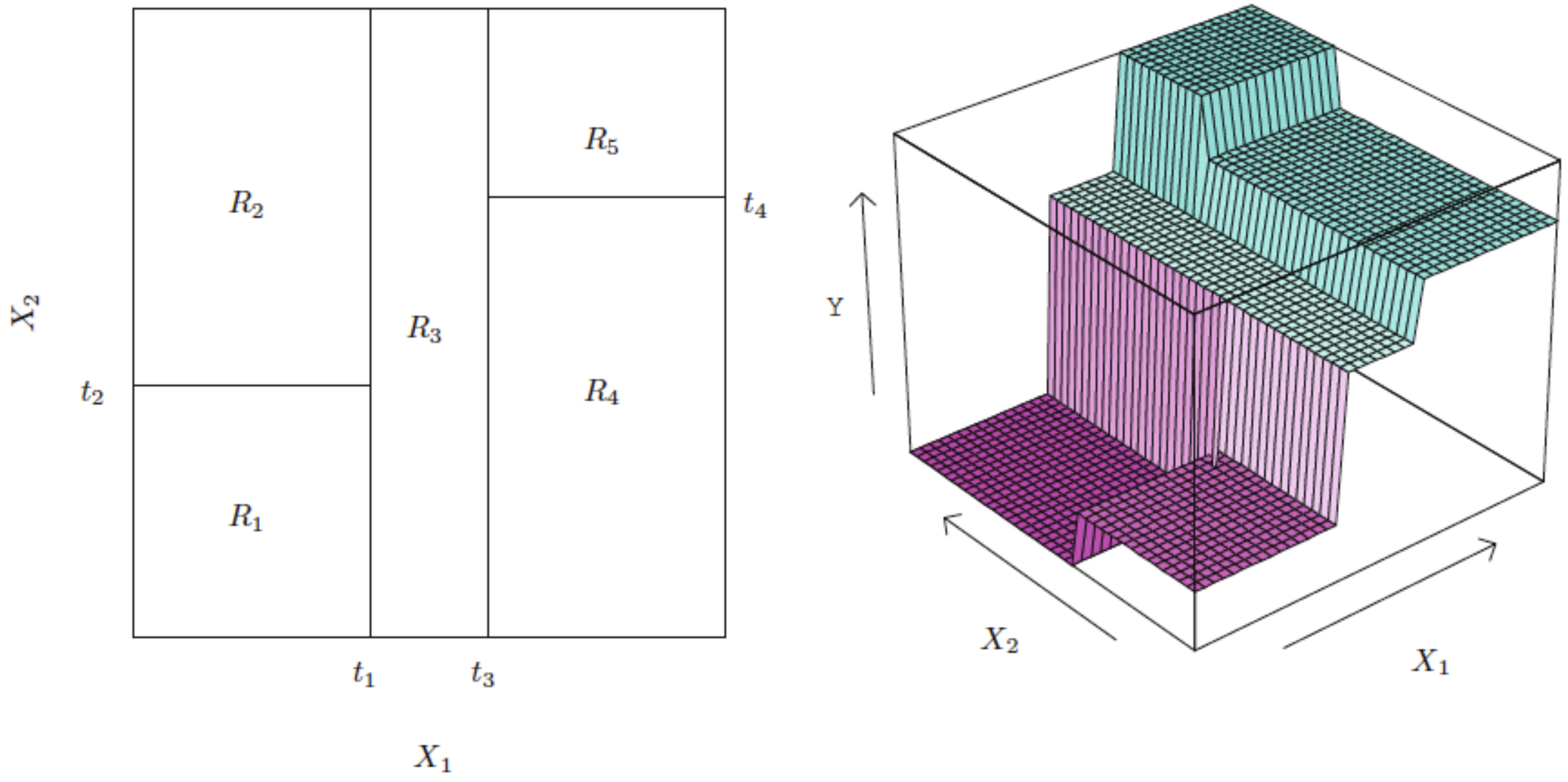
Mathematical expression of the prediction rule



Mathematical expression of the prediction rule

A trained tree consists of:

- M regions $\widehat{R}_1, \ldots, \widehat{R}_M$
- response values $\hat{c}_1, \ldots, \hat{c}_M$



Mathematical expression of the prediction rule

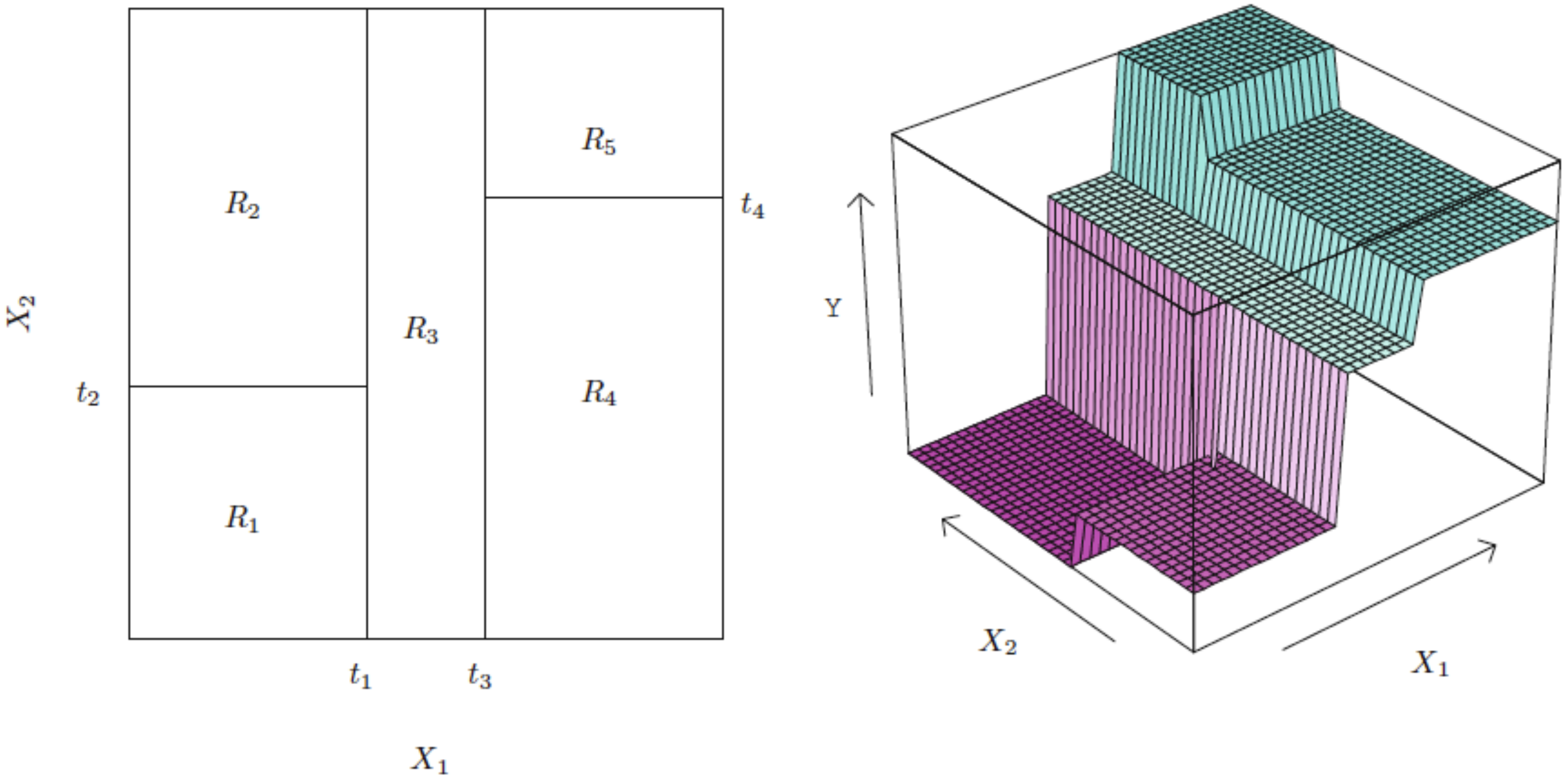
A trained tree consists of:

- M regions $\widehat{R}_1, \ldots, \widehat{R}_M$
- response values $\widehat{c}_1, \ldots, \widehat{c}_M$

For a new feature vector X^{test}, predict the constant value $\widehat{c}_m$ for region $\widehat{R}_m$:

$$\widehat{Y}^{\text{test}} = \widehat{c}_m \text{ if } X^{\text{test}} \in \widehat{R}_m.$$

(continuous or categorical response)



Partitioning for continuous and categorical features

Partitioning for continuous and categorical features

Suppose we partition on X_j.

Partitioning for continuous and categorical features

Suppose we partition on X_j.

- If X_j is continuous, we just find a split point s and split into $\{X : X_j < s\}$ and $\{X : X_j \geq s\}$.

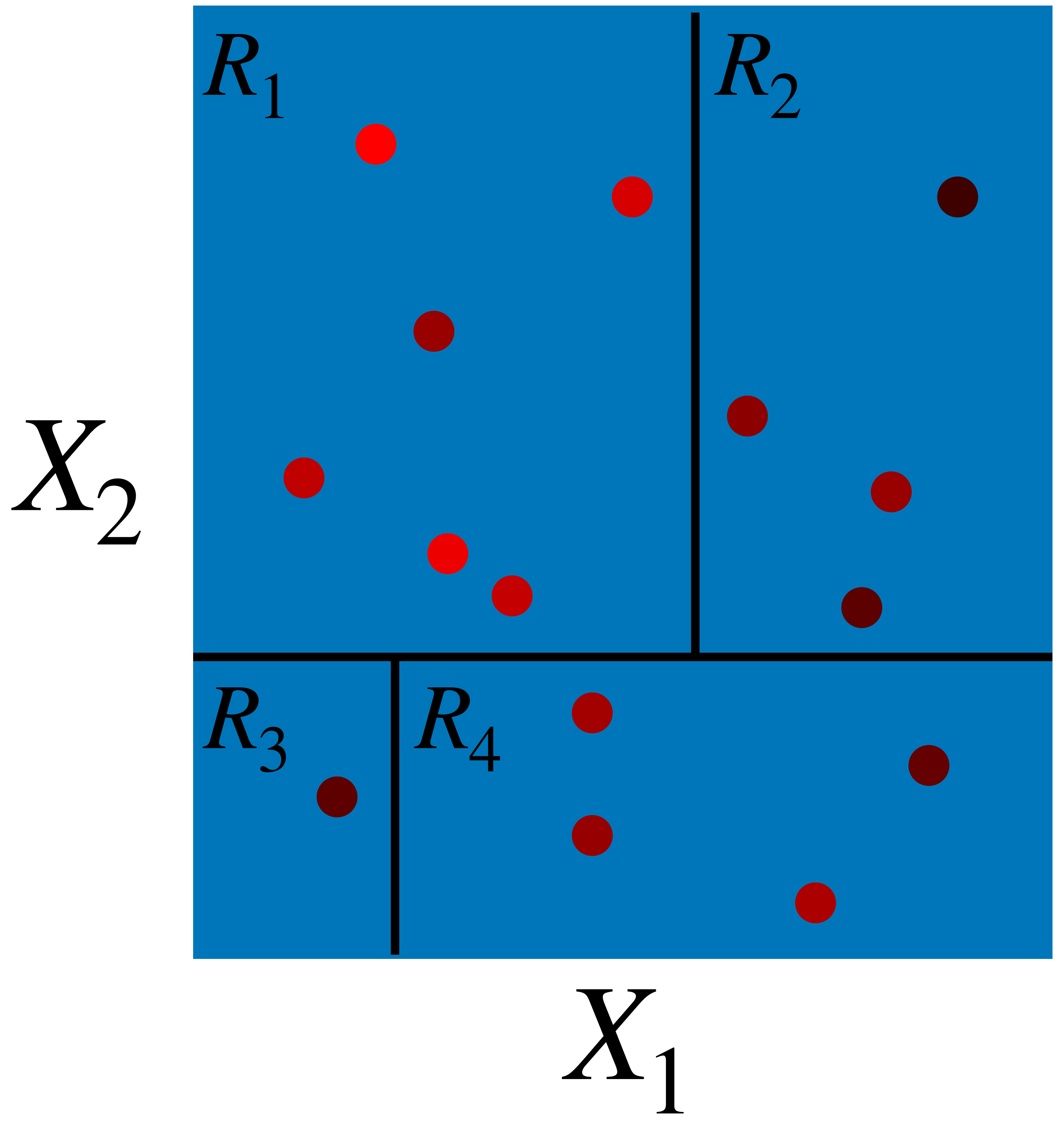
Partitioning for continuous and categorical features

Suppose we partition on X_j.

- If X_j is continuous, we just find a split point s and split into $\{X : X_j < s\}$ and $\{X : X_j \geq s\}$.
- If X_j is categorical, e.g. with levels $\{a, b, c, d, e\}$, then we need to split the levels into two groups, e.g. $\{a, c\}$ and $\{b, d, e\}$, giving the partitions $\{X : X_j \in \{a, c\}\}$ and $\{X : X_j \in \{b, d, e\}\}$.

Training a regression tree

The squared error objective



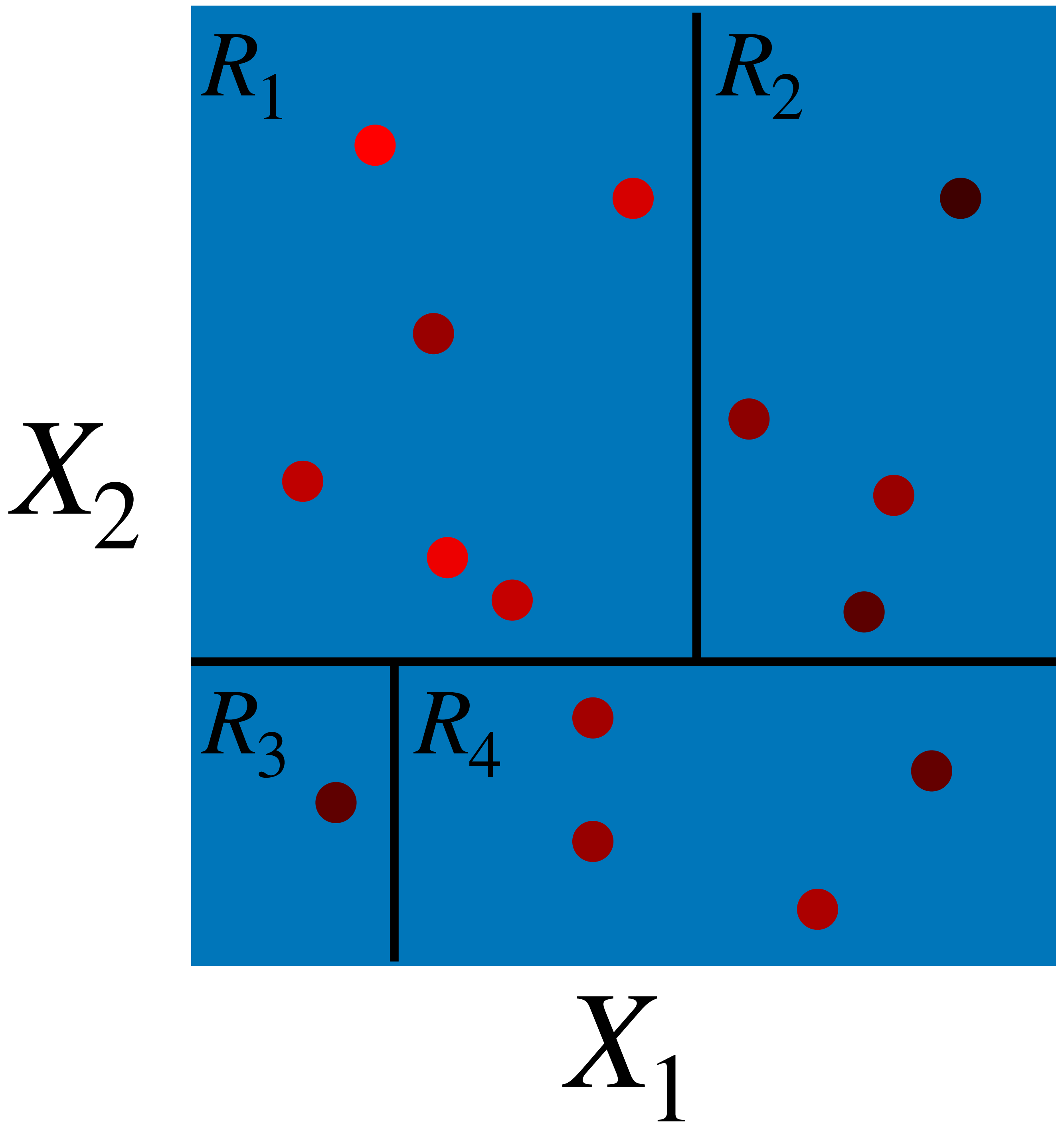
Training a regression tree

The squared error objective

As usual, we are given a training dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$.

For a fixed M, we seek rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ and values $\widehat{c}_1, \ldots, \widehat{c}_M$ to minimize the residual sum of squares (RSS):

$$\widehat{R}_1, \ldots, \widehat{R}_M, \widehat{c}_1, \ldots, \widehat{c}_M = \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



Training a regression tree

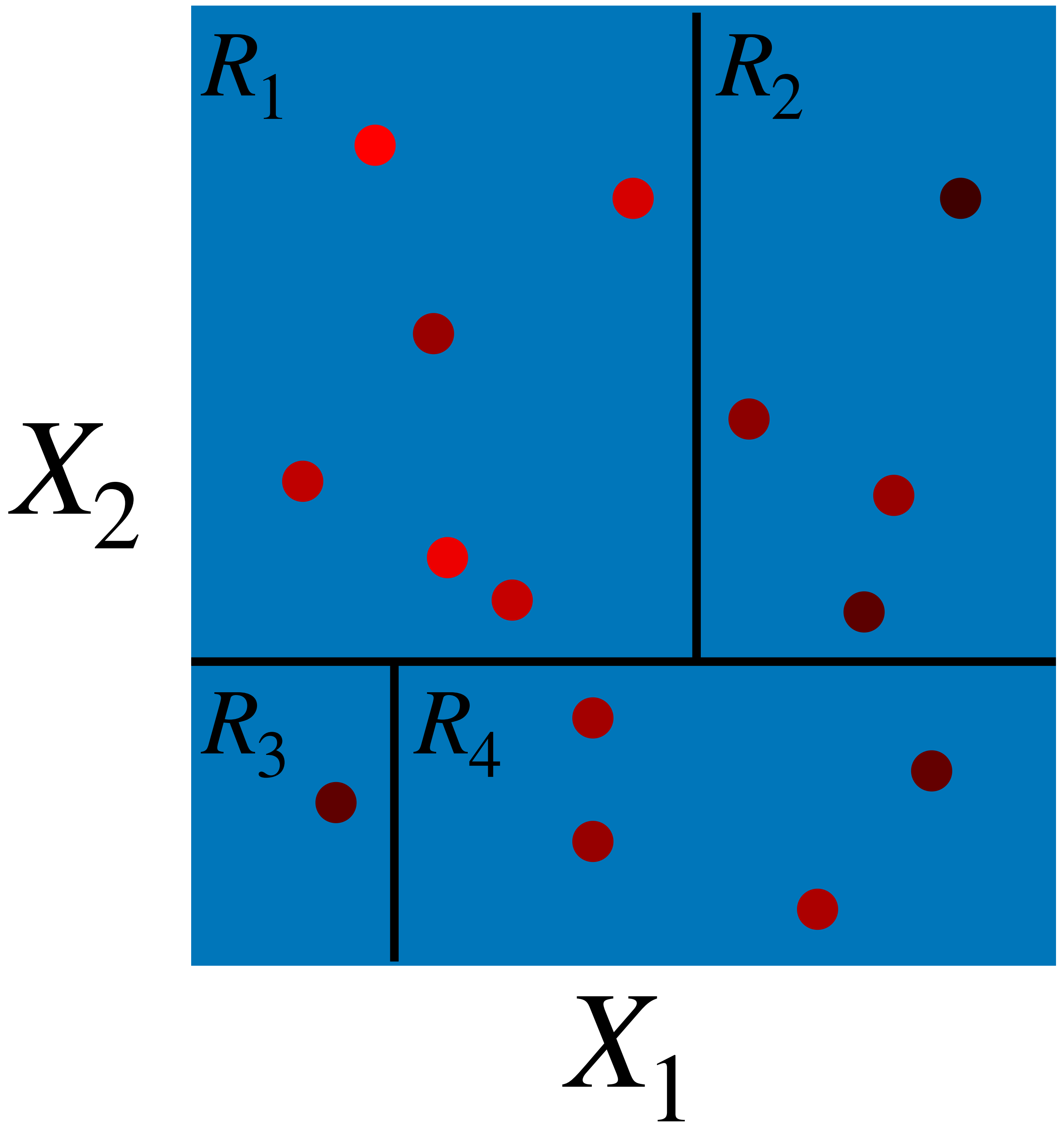
The squared error objective

As usual, we are given a training dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$.

For a fixed M, we seek rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ and values $\hat{c}_1, \ldots, \hat{c}_M$ to minimize the residual sum of squares (RSS):

$$\widehat{R}_1, \ldots, \widehat{R}_M, \hat{c}_1, \ldots, \hat{c}_M = \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$= \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \left\{ \sum_{i:X_i\in R_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in R_M} (Y_i - c_M)^2 \right\}$$



Training a regression tree

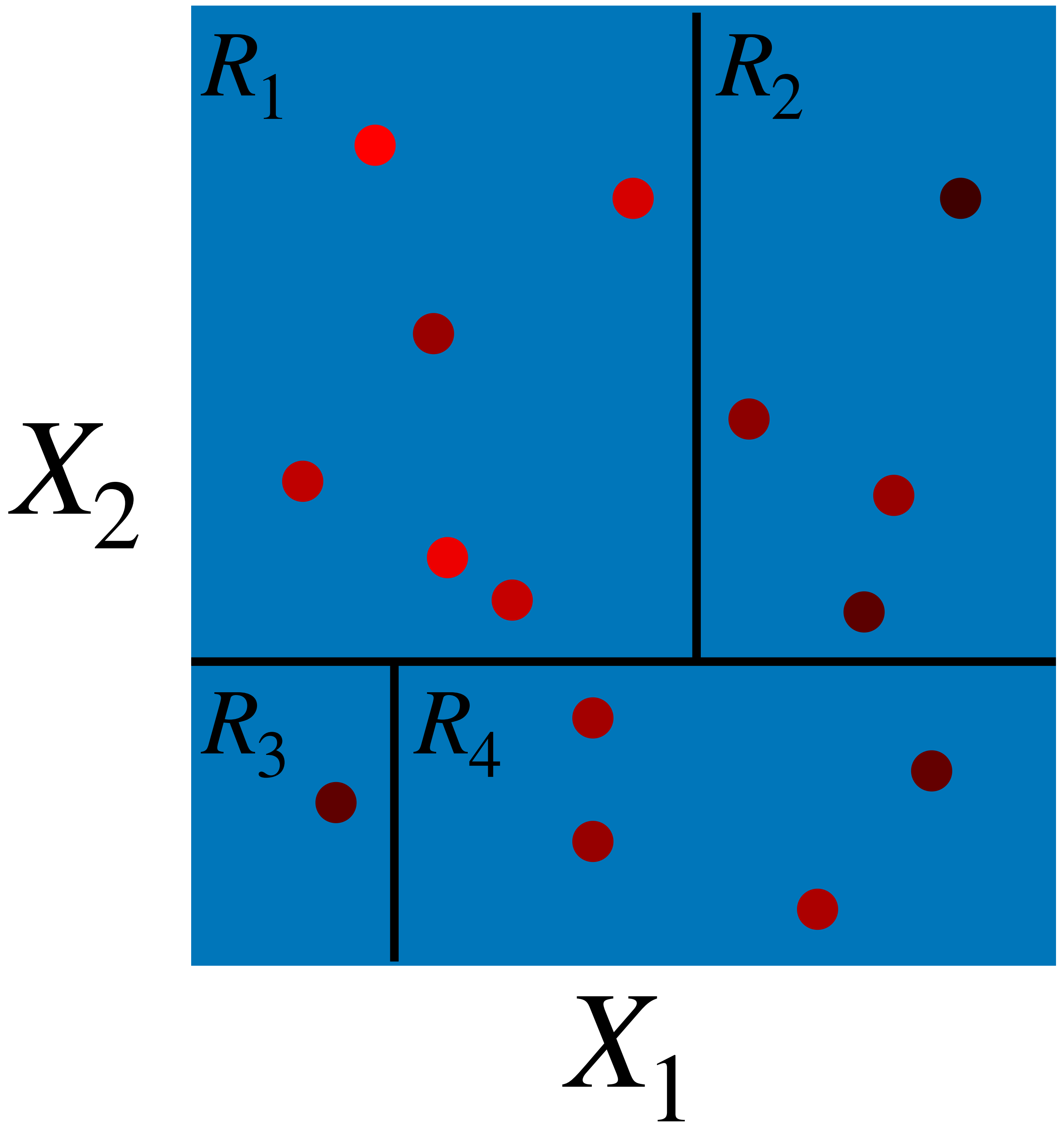
The squared error objective

As usual, we are given a training dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$.

For a fixed M, we seek rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ and values $\widehat{c}_1, \ldots, \widehat{c}_M$ to minimize the residual sum of squares (RSS):

$$\widehat{R}_1, \ldots, \widehat{R}_M, \widehat{c}_1, \ldots, \widehat{c}_M = \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

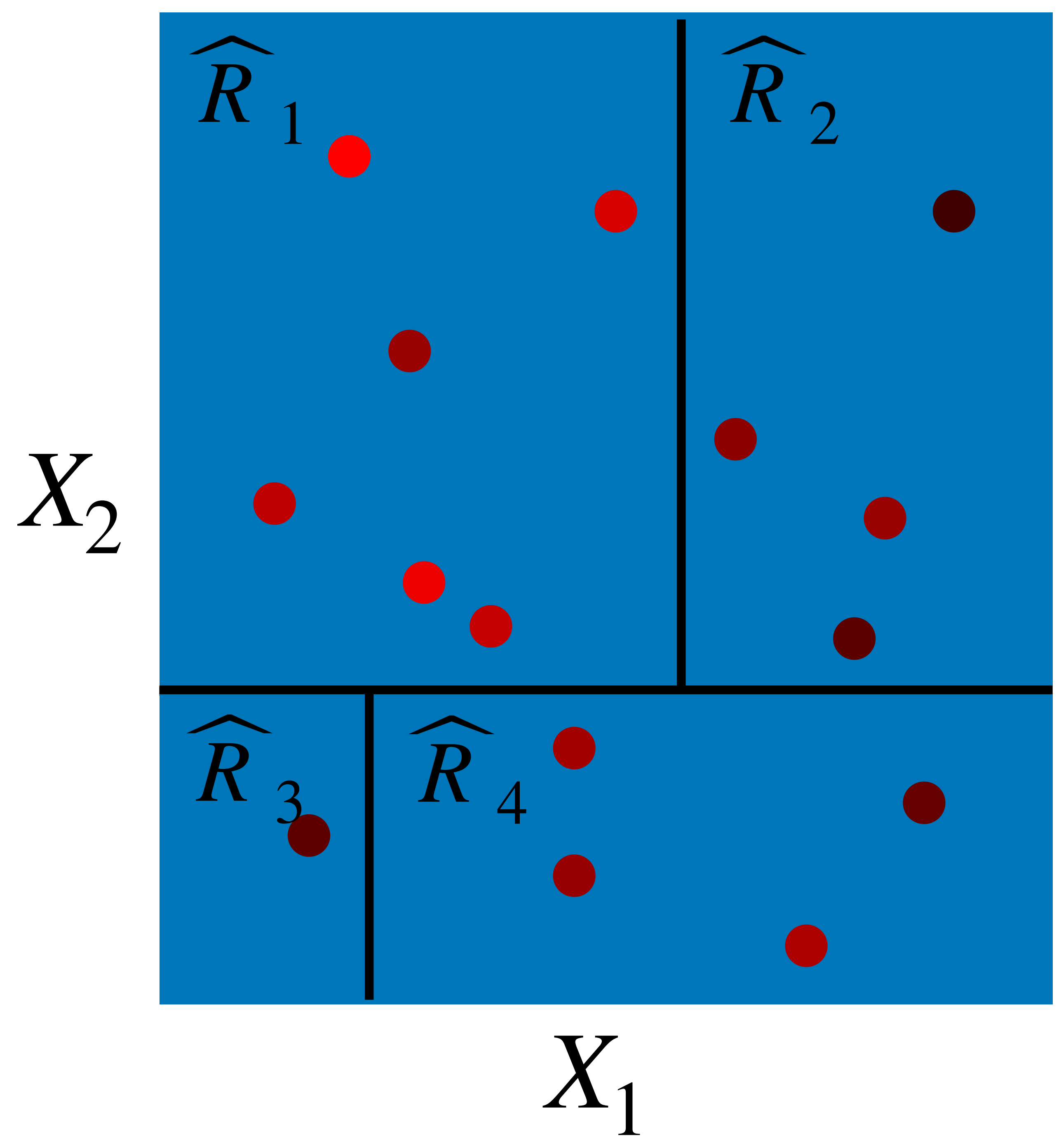
$$= \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \left\{ \underbrace{\sum_{i:X_i \in R_1} (Y_i - c_1)^2}_{\text{RSS for } R_1} + \cdots + \underbrace{\sum_{i:X_i \in R_M} (Y_i - c_M)^2}_{\text{RSS for } R_M} \right\}$$



Training a regression tree

Optimal $\widehat{c}_m$ given $\widehat{R}_m$

First let's consider a simpler problem, where rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ are given:

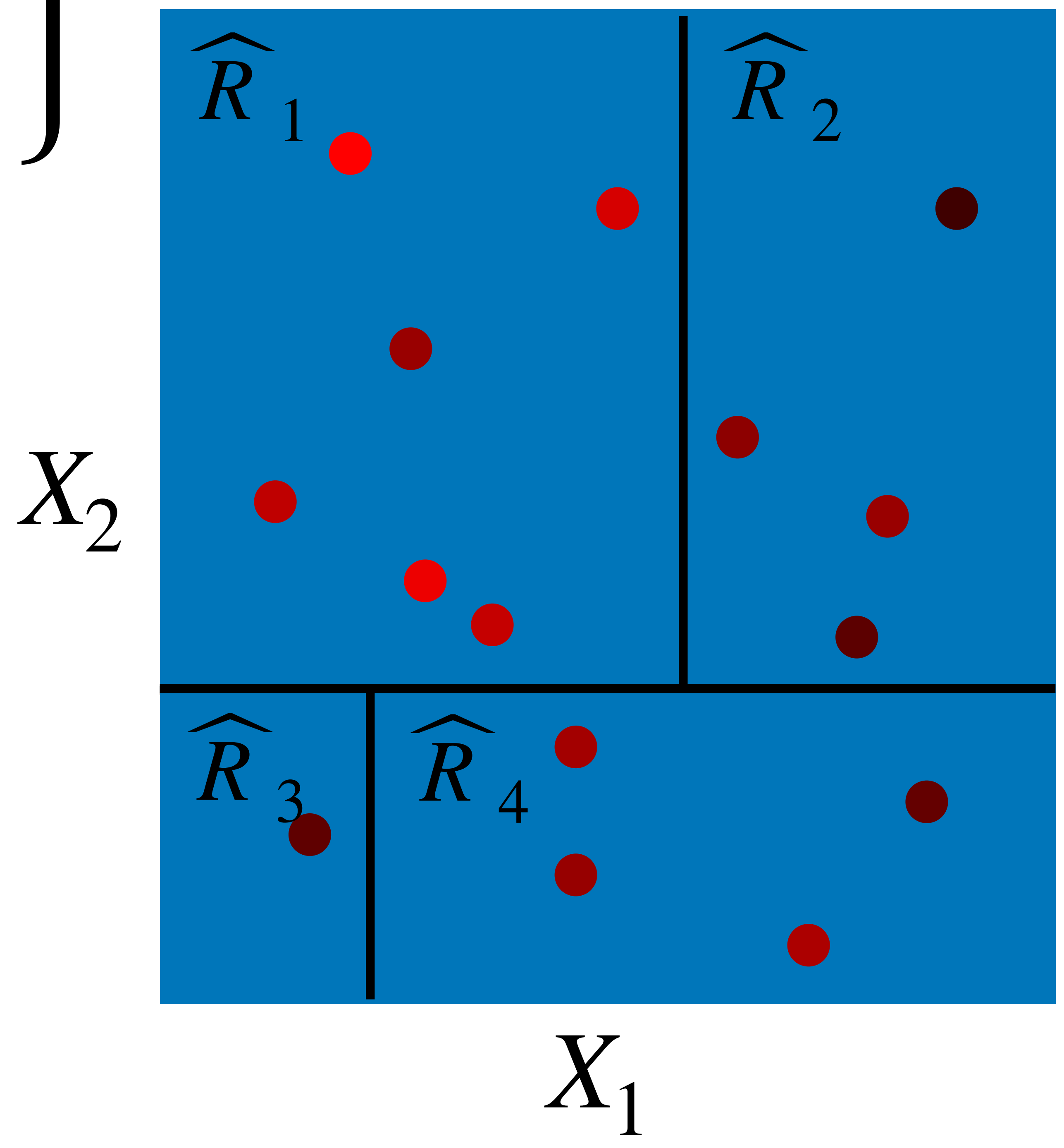


Training a regression tree

Optimal $\widehat{c}_m$ given $\widehat{R}_m$

First let's consider a simpler problem, where rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ are given:

$$\widehat{c}_1, \ldots, \widehat{c}_M = \underset{c_1,\ldots,c_M}{\arg\min} \left\{ \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2 \right\}$$



Training a regression tree

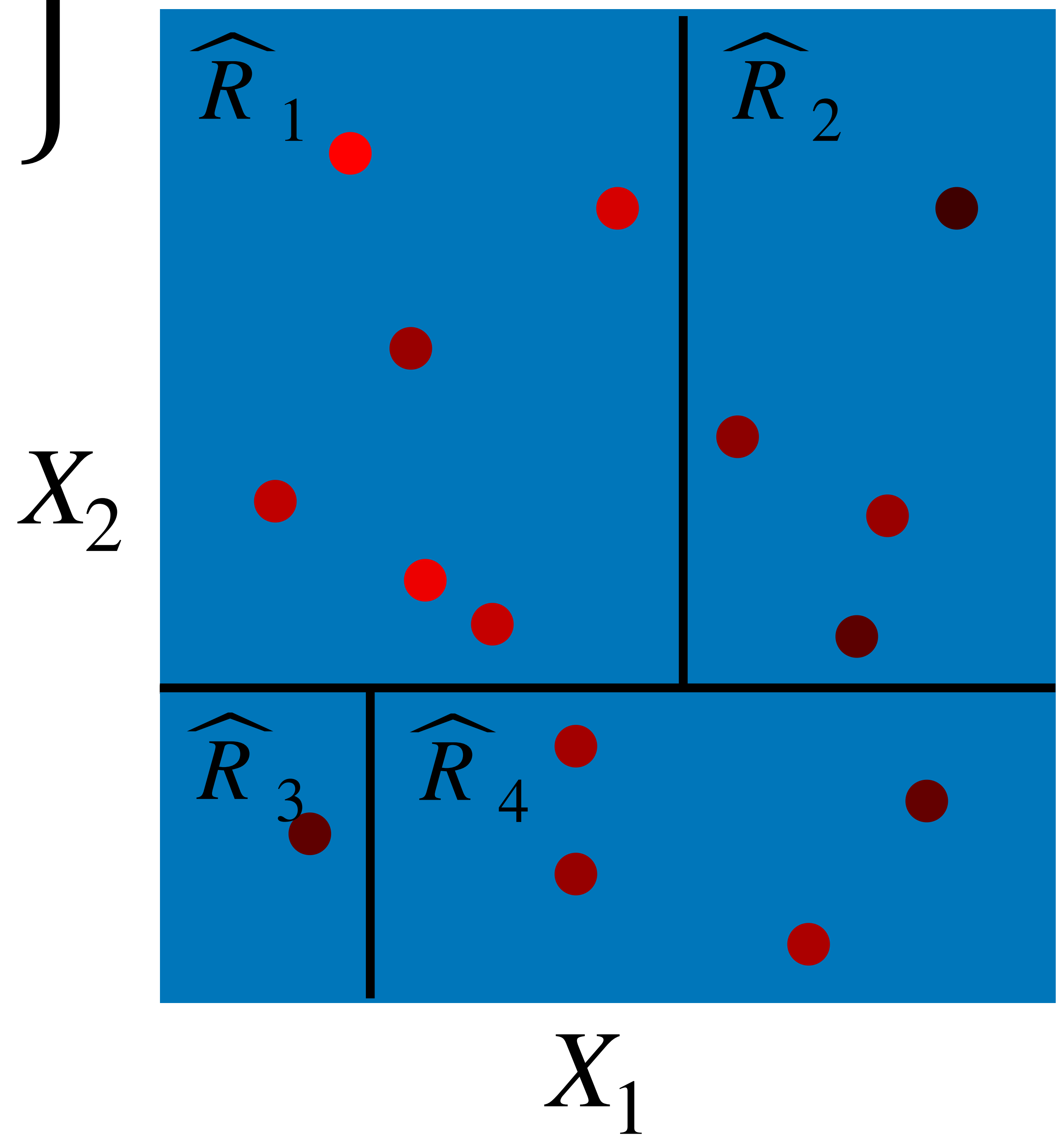
Optimal $\widehat{c}_m$ given $\widehat{R}_m$

First let's consider a simpler problem, where rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ are given:

$$\widehat{c}_1, \ldots, \widehat{c}_M = \underset{c_1,\ldots,c_M}{\arg\min} \left\{ \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2 \right\}$$

We're fitting a constant to each region, so the solution is

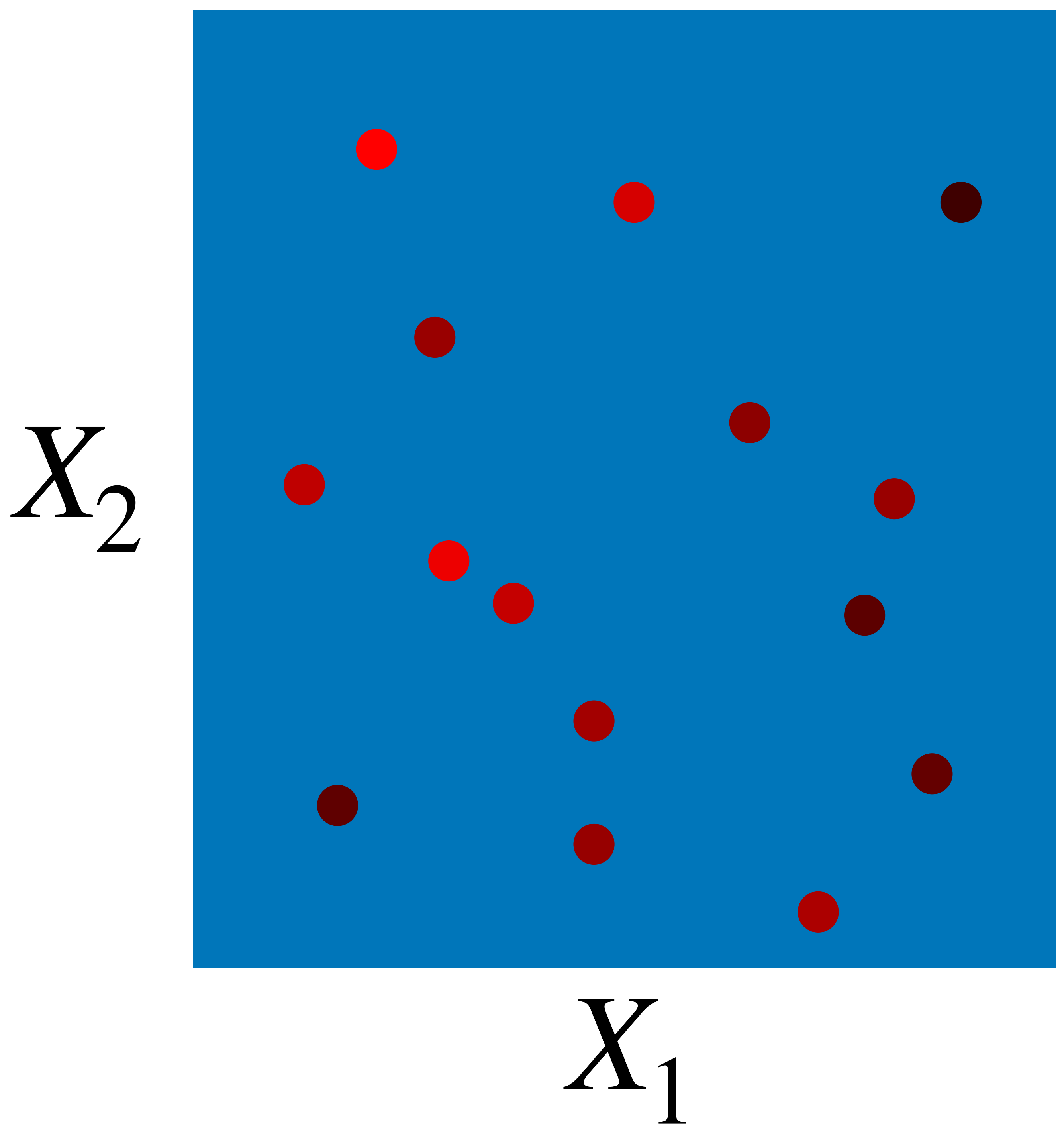
$$\widehat{c}_m = \text{mean}\left(\{ Y_i : X_i \in \widehat{R}_m \} \right).$$



Training a regression tree

Finding the rectangles $\widehat{R}_m$

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

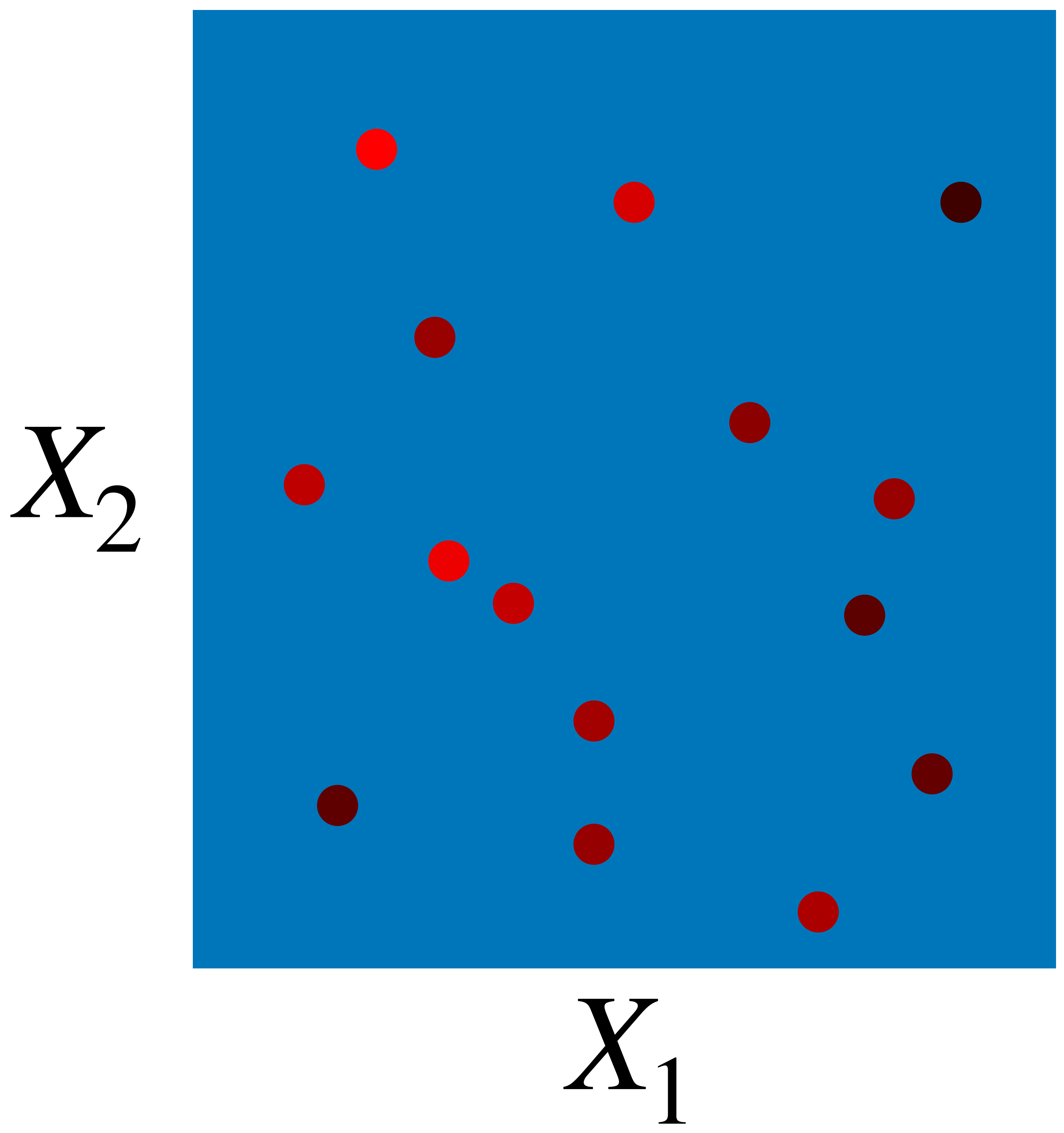


Training a regression tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



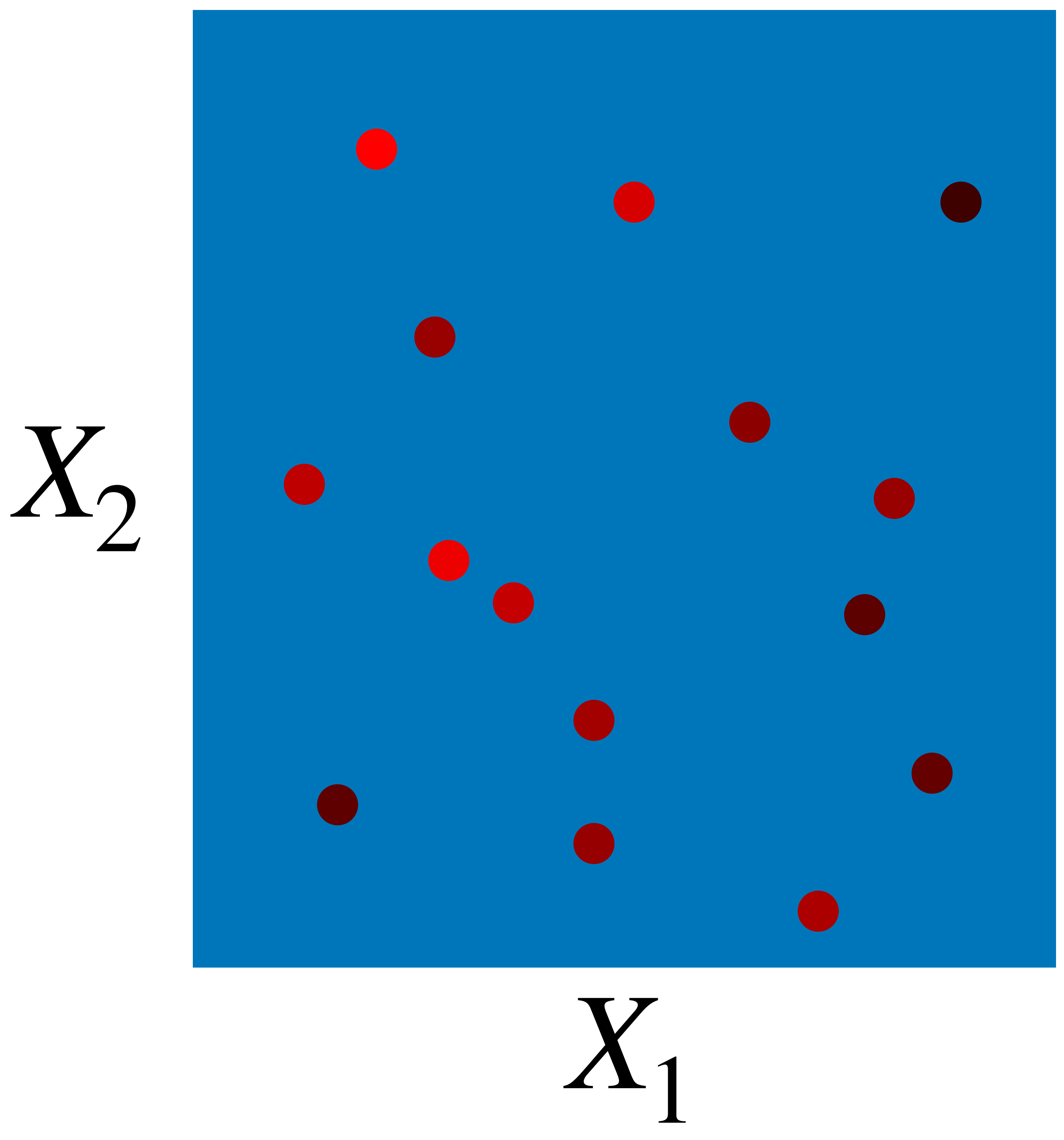
Training a regression tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

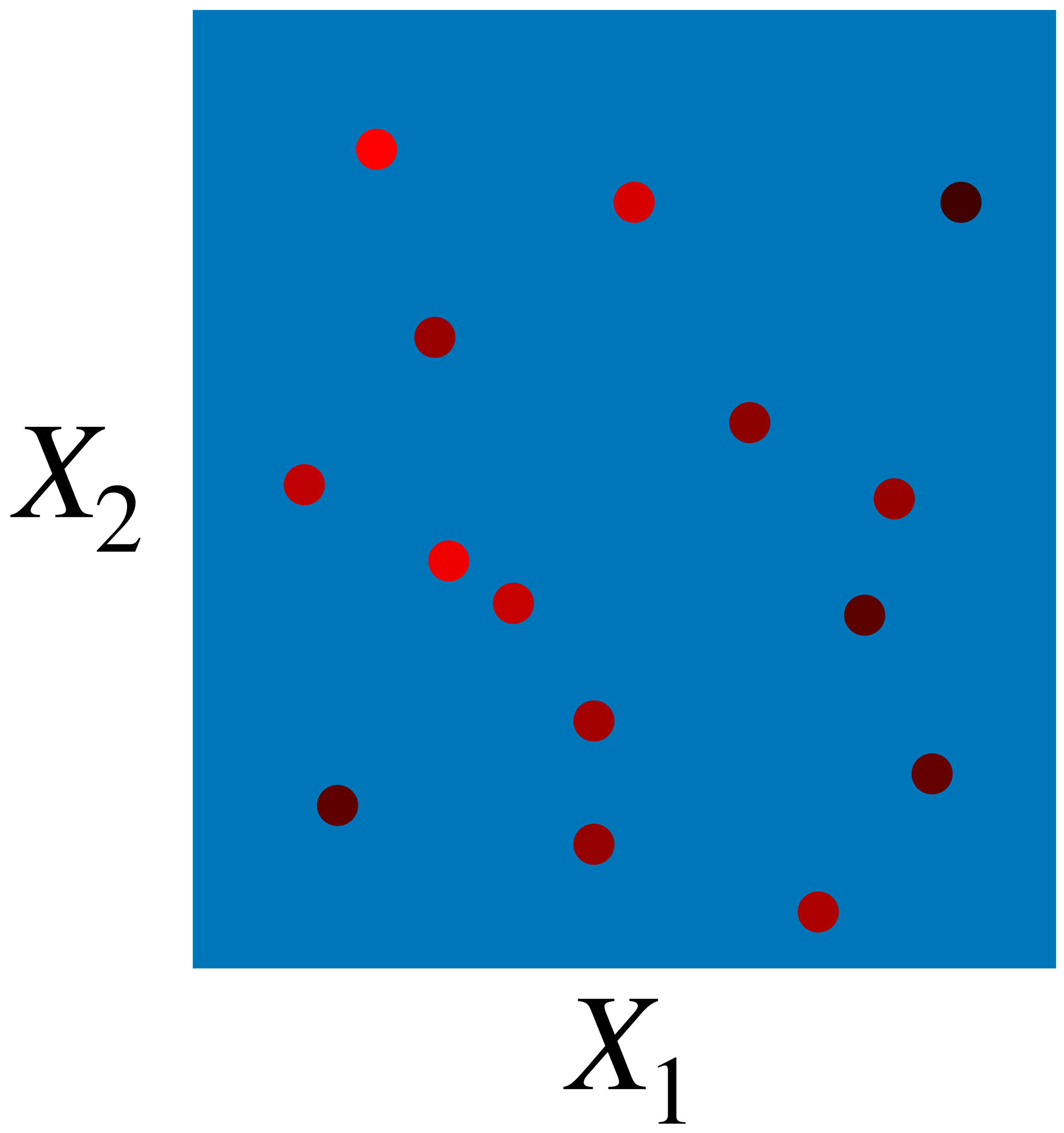
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

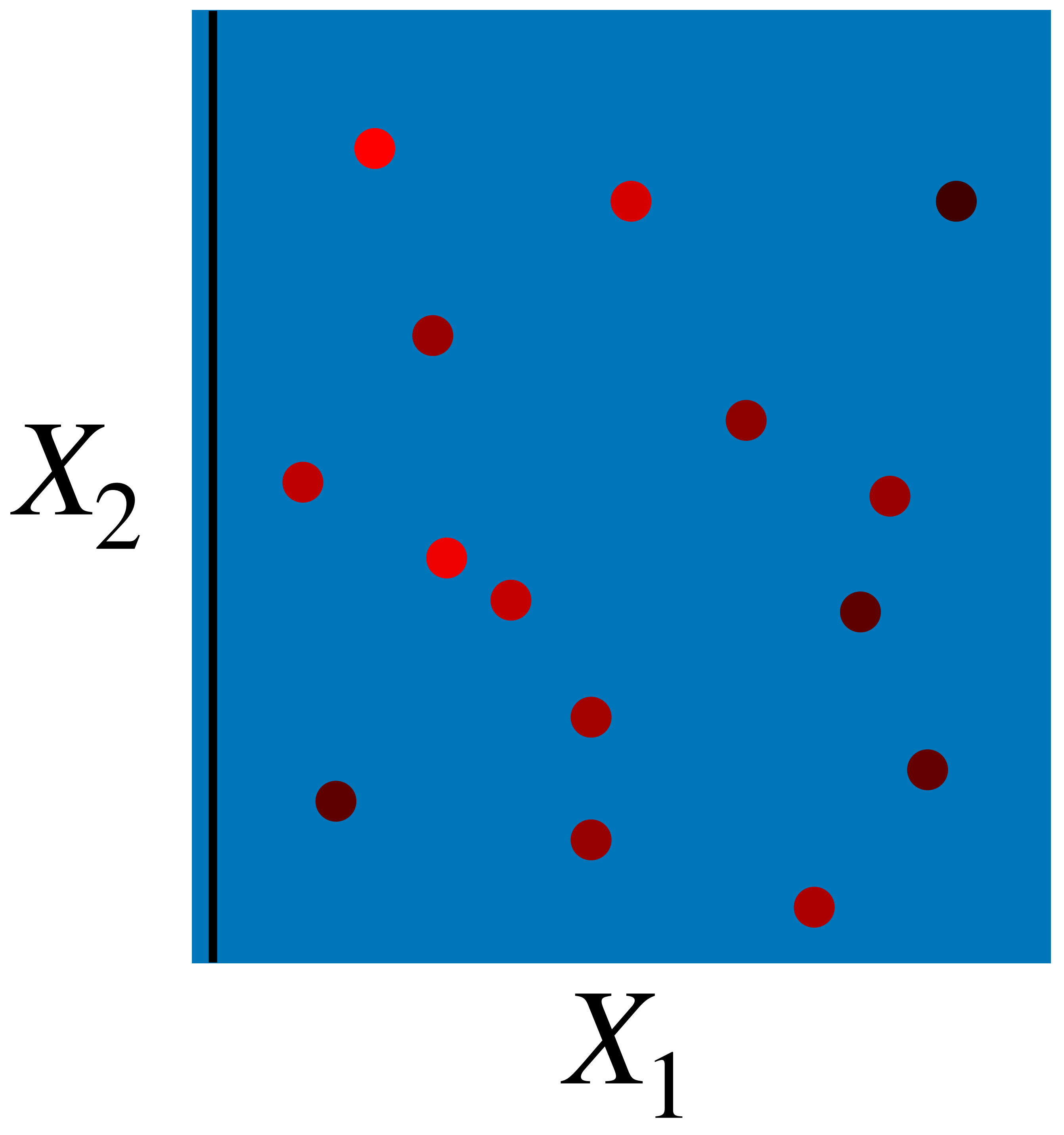
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

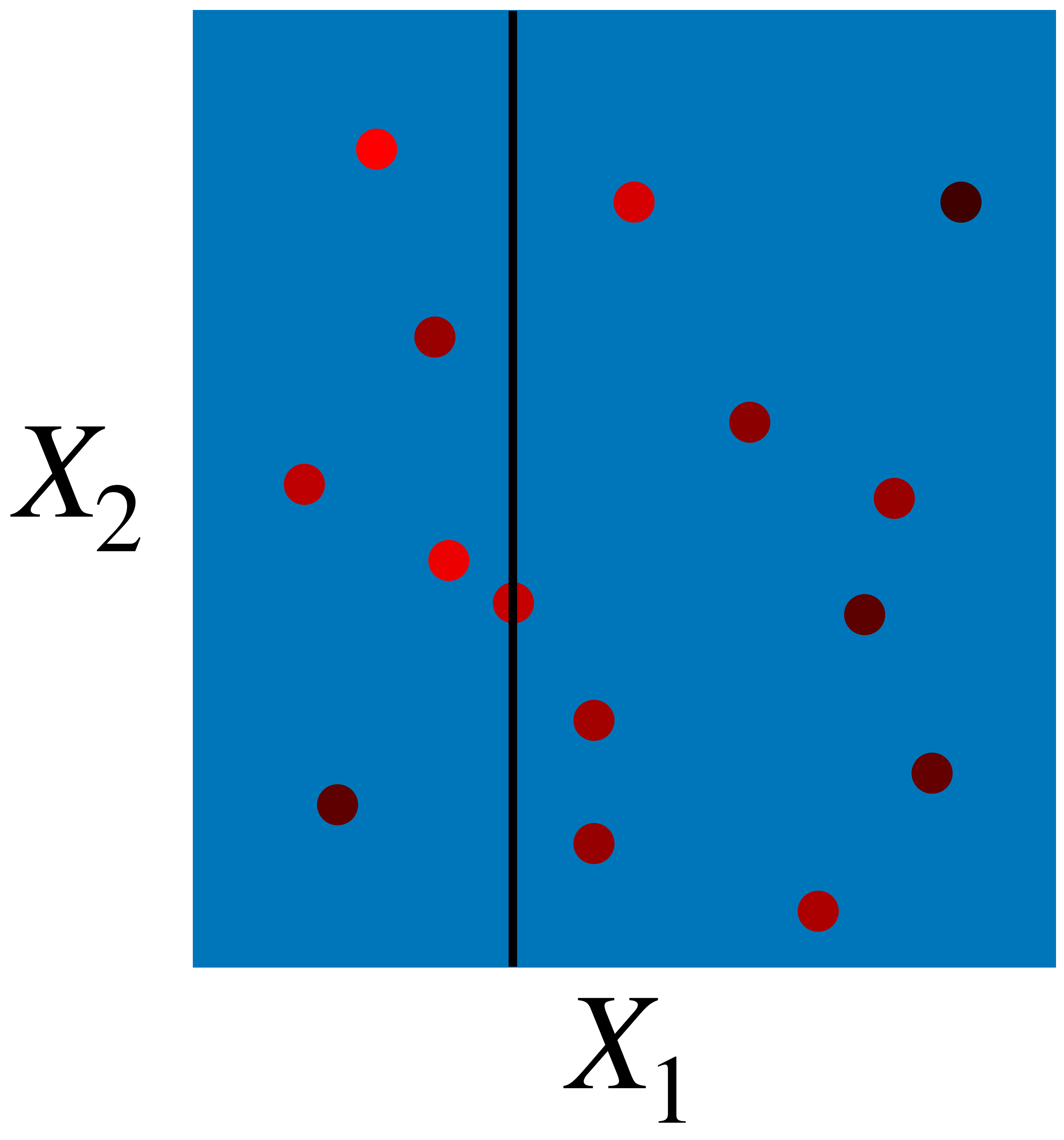
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

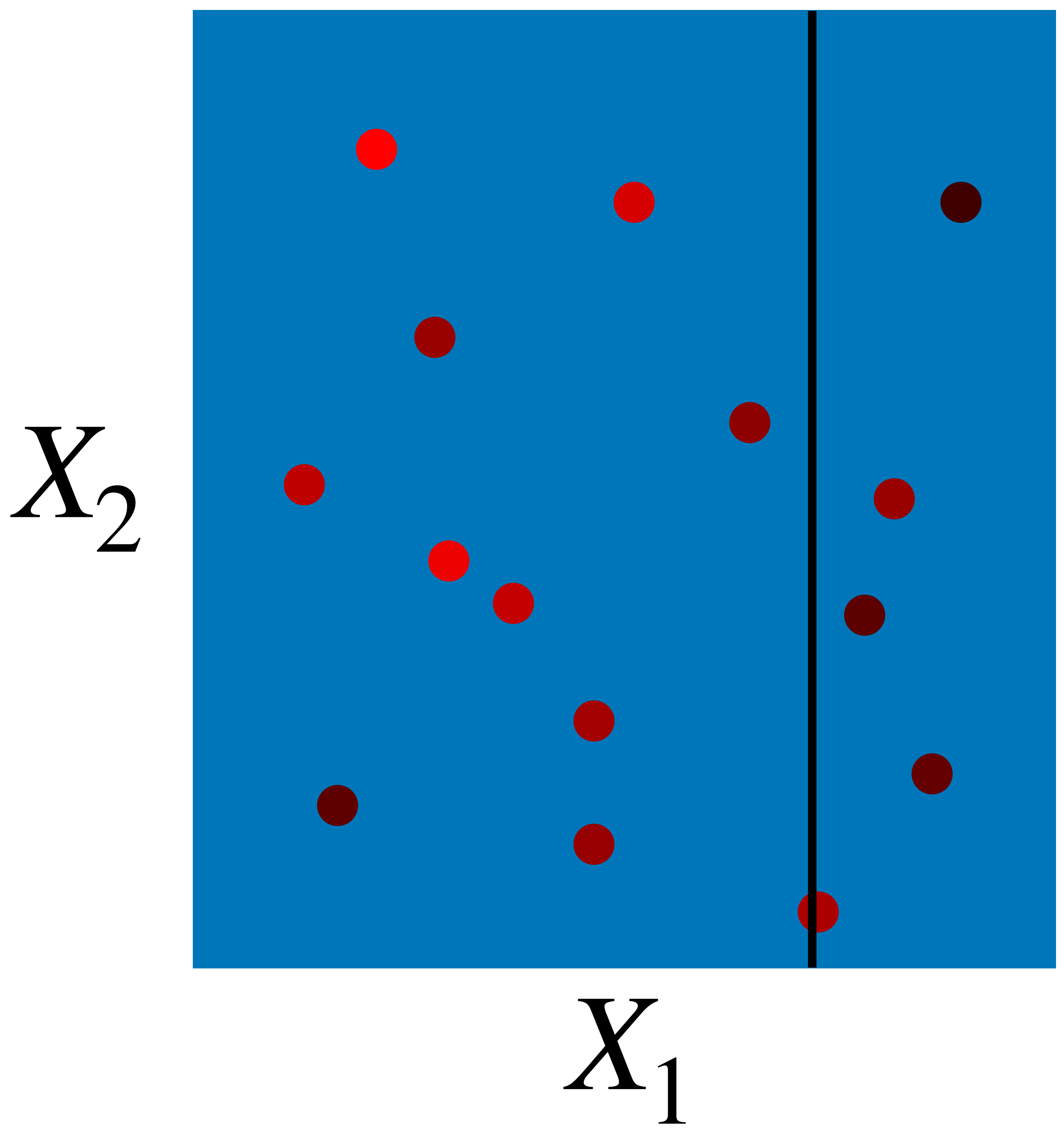
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

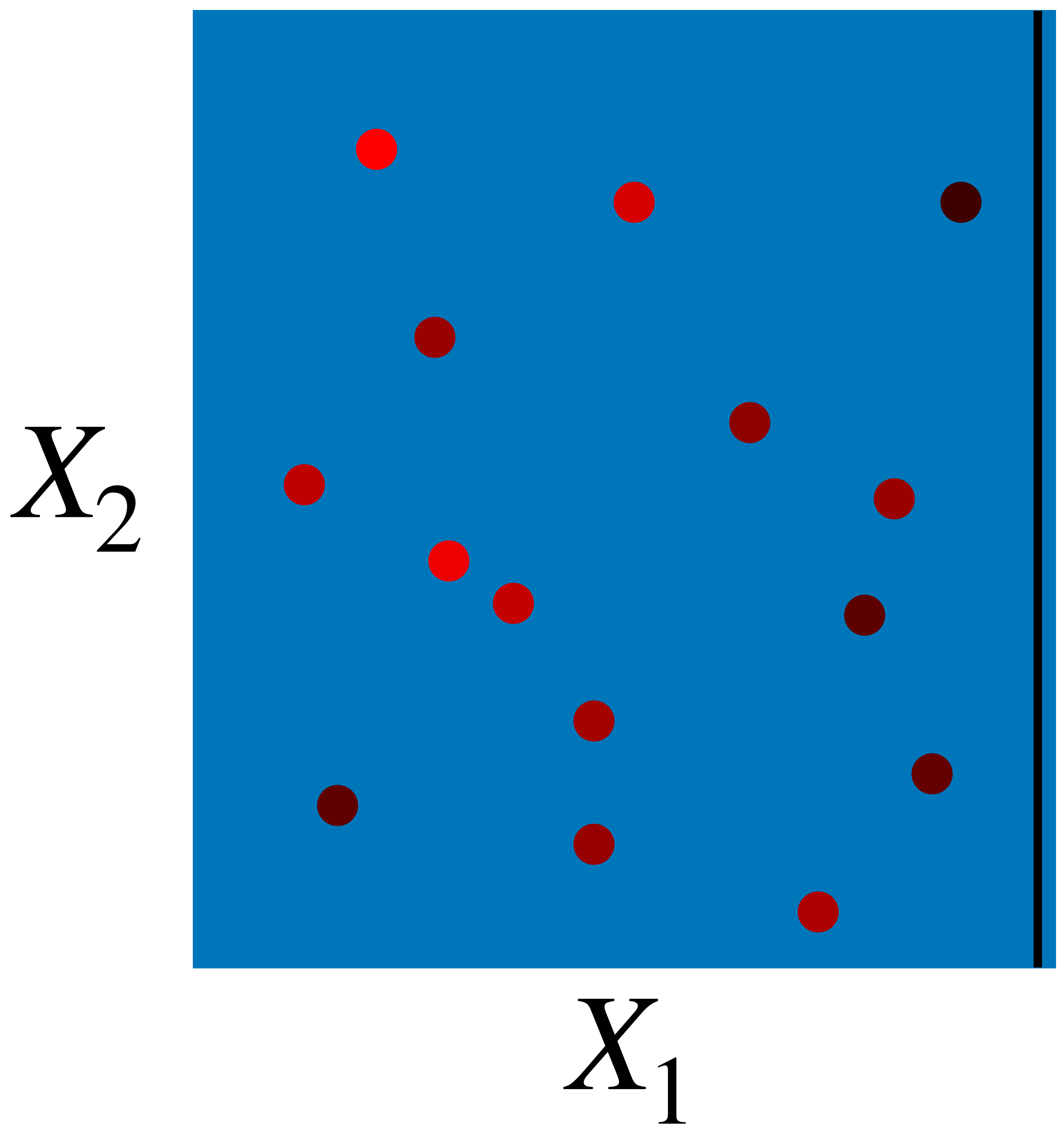
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.



Training a regression tree

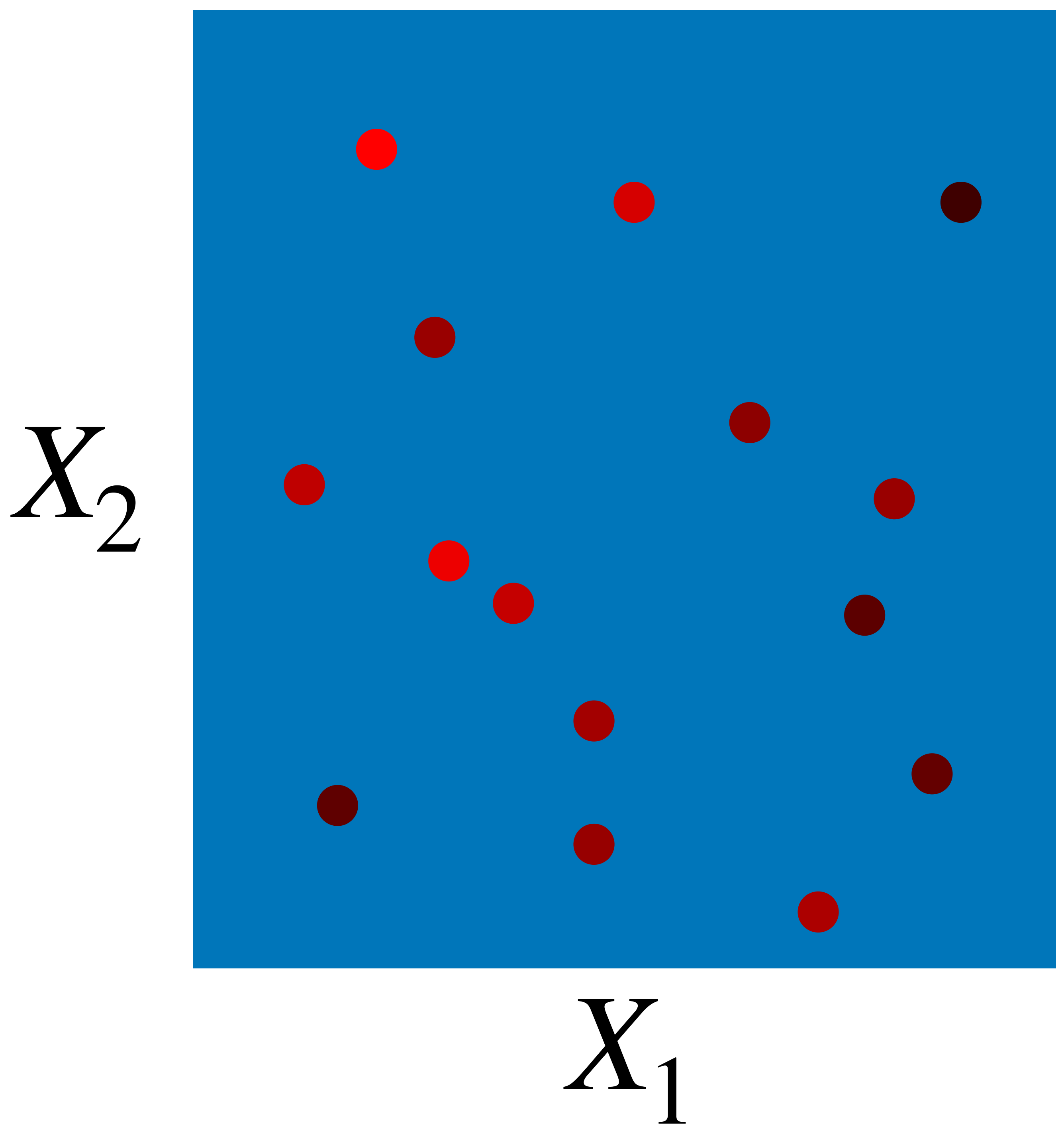
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

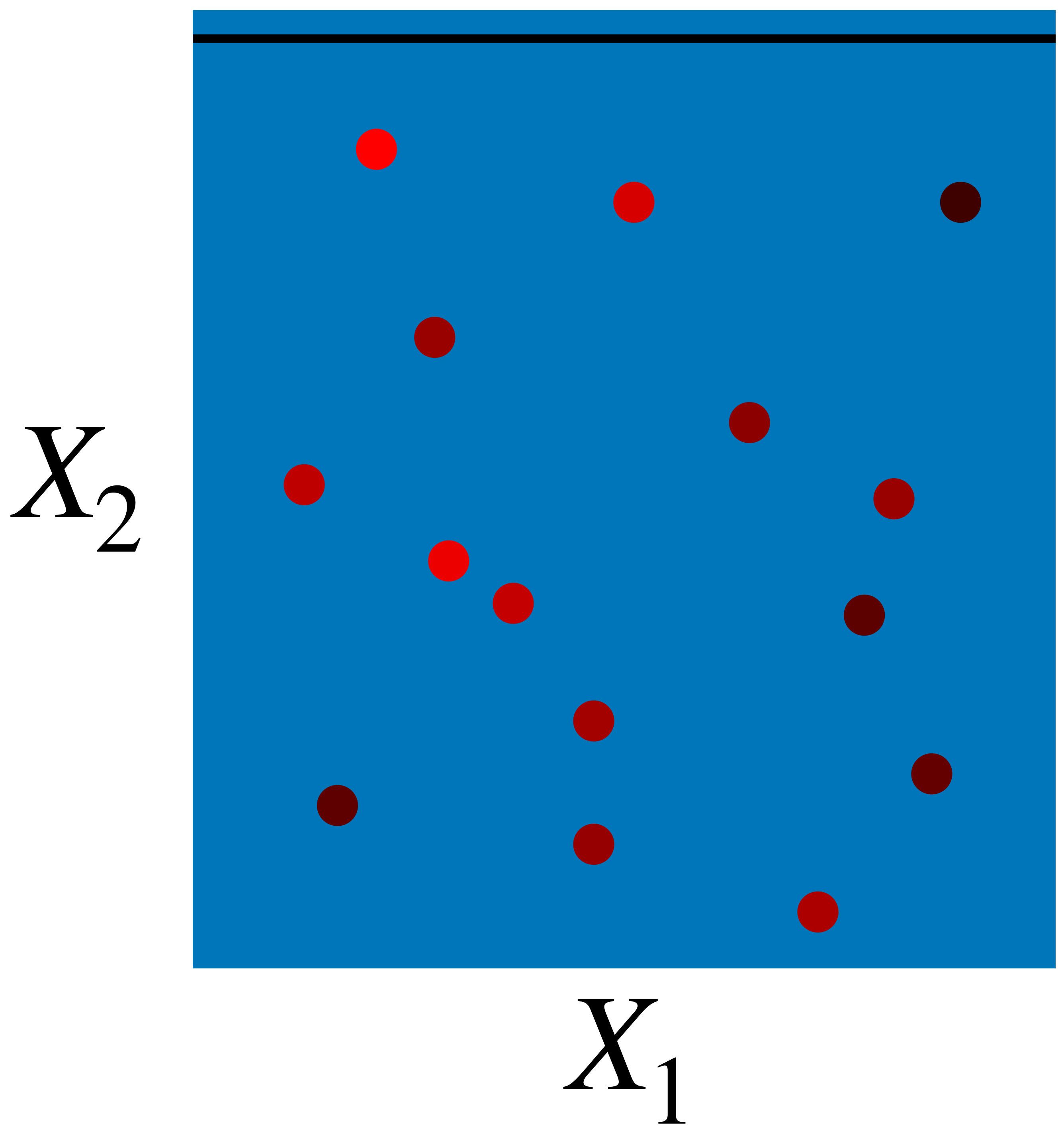
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

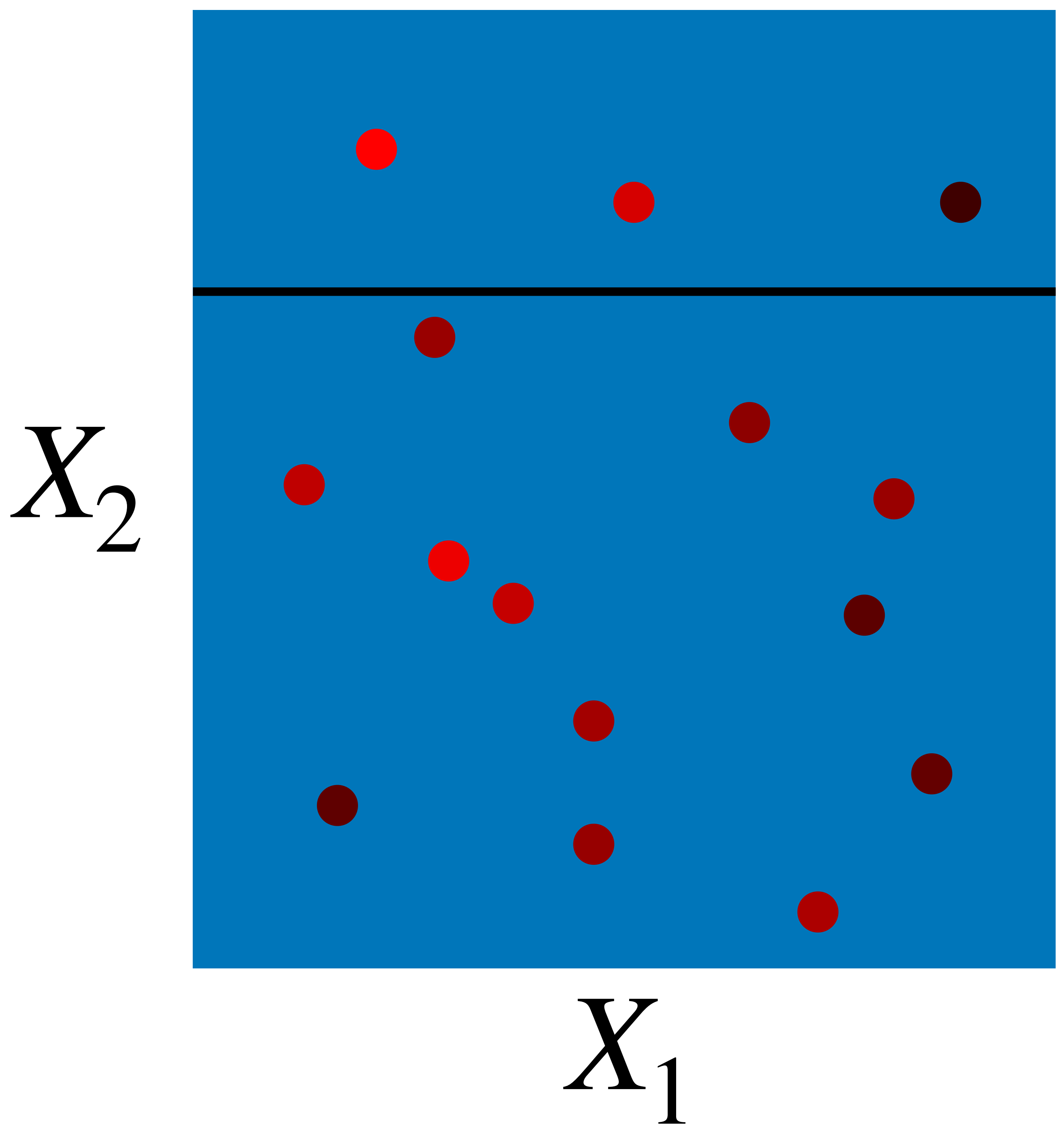
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

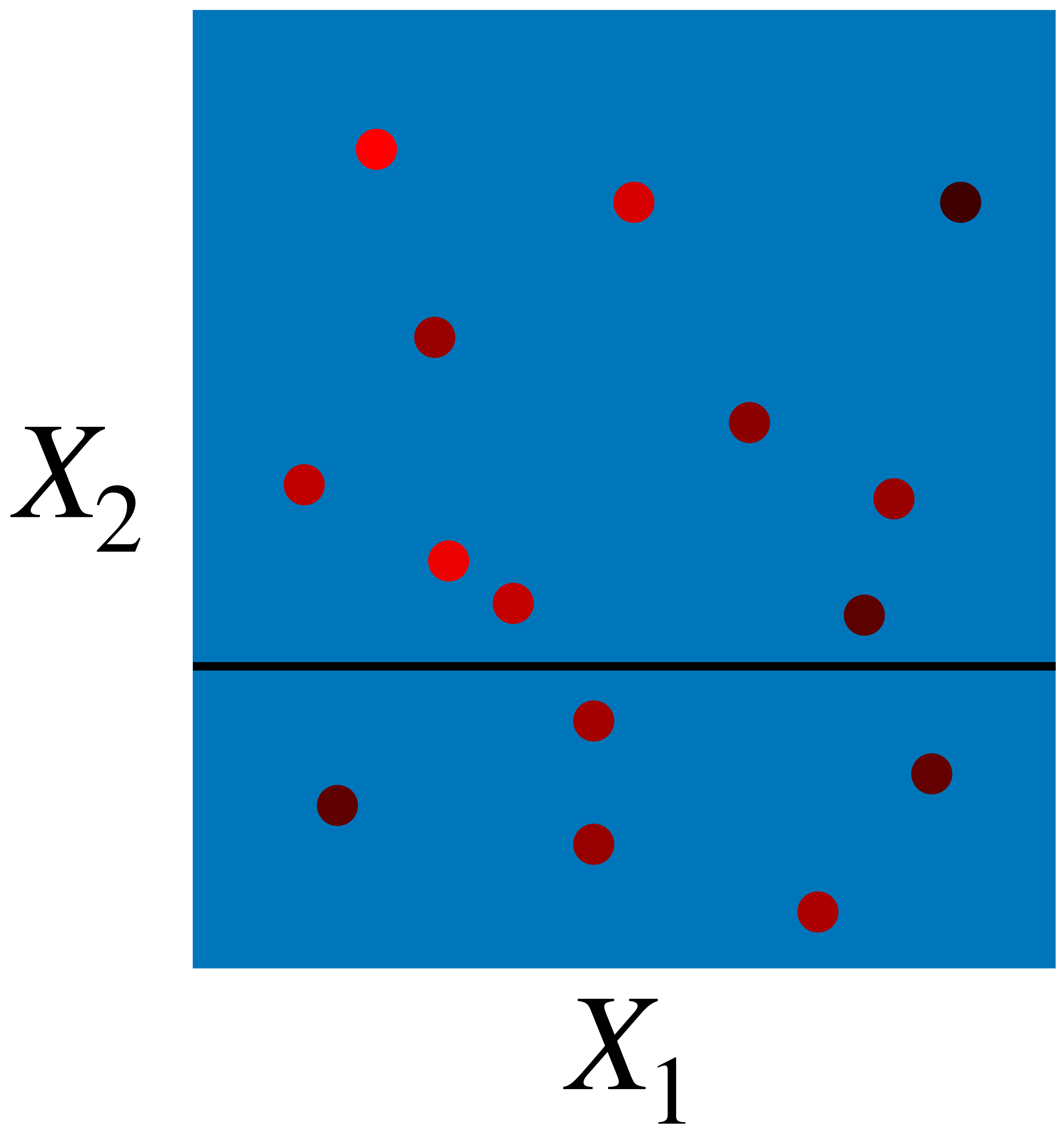
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

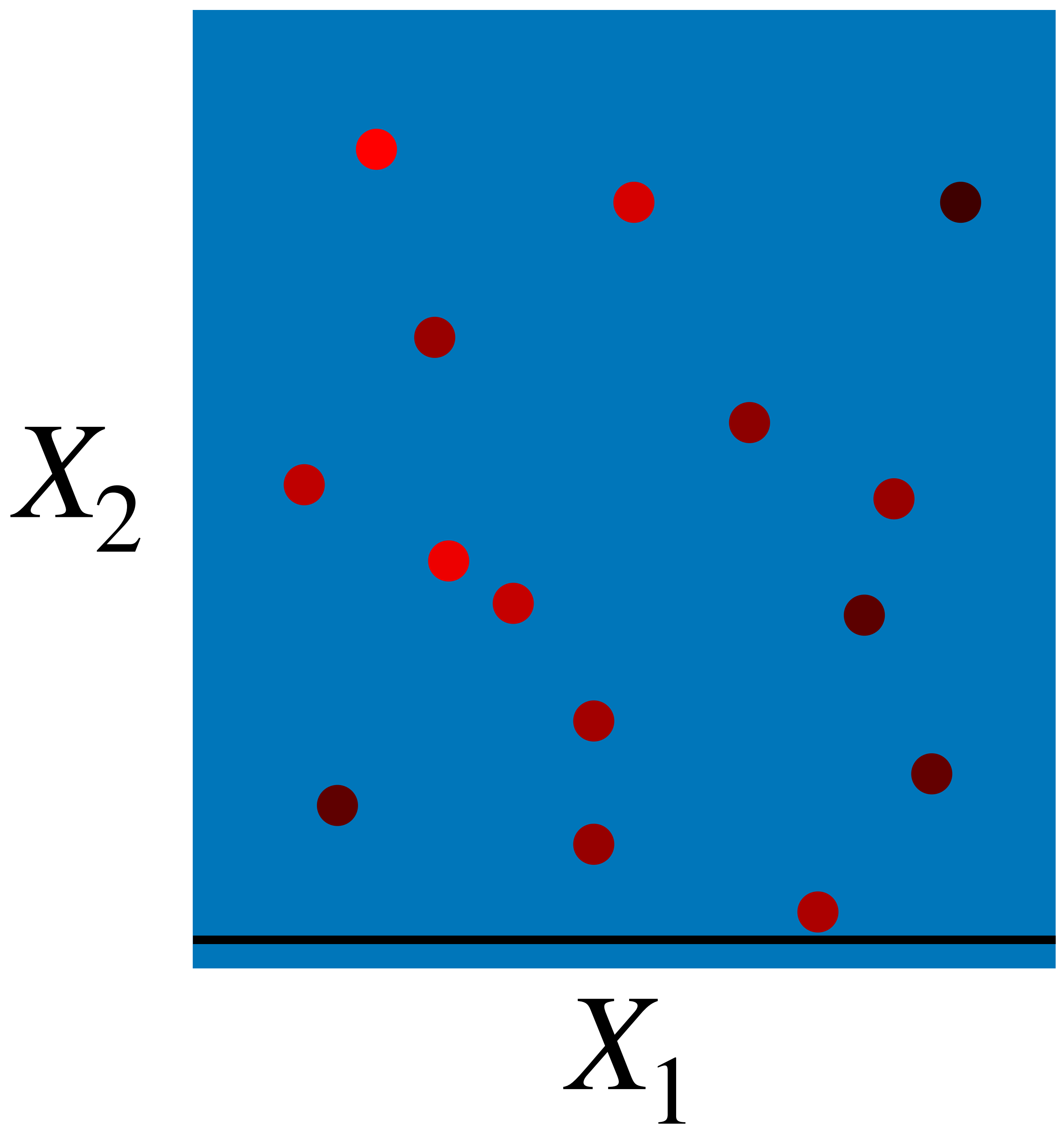
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.



Training a regression tree

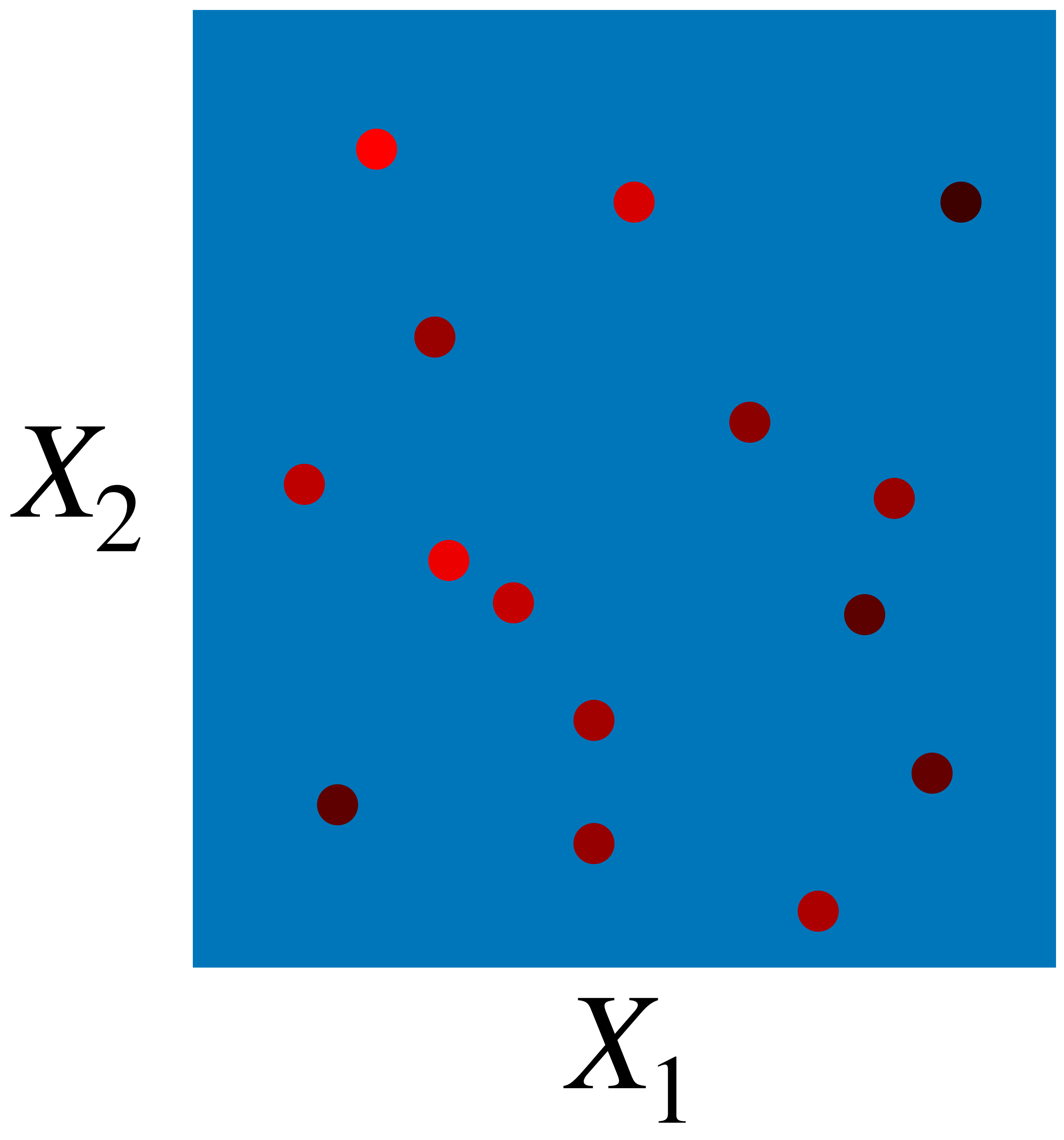
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i: X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i: X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

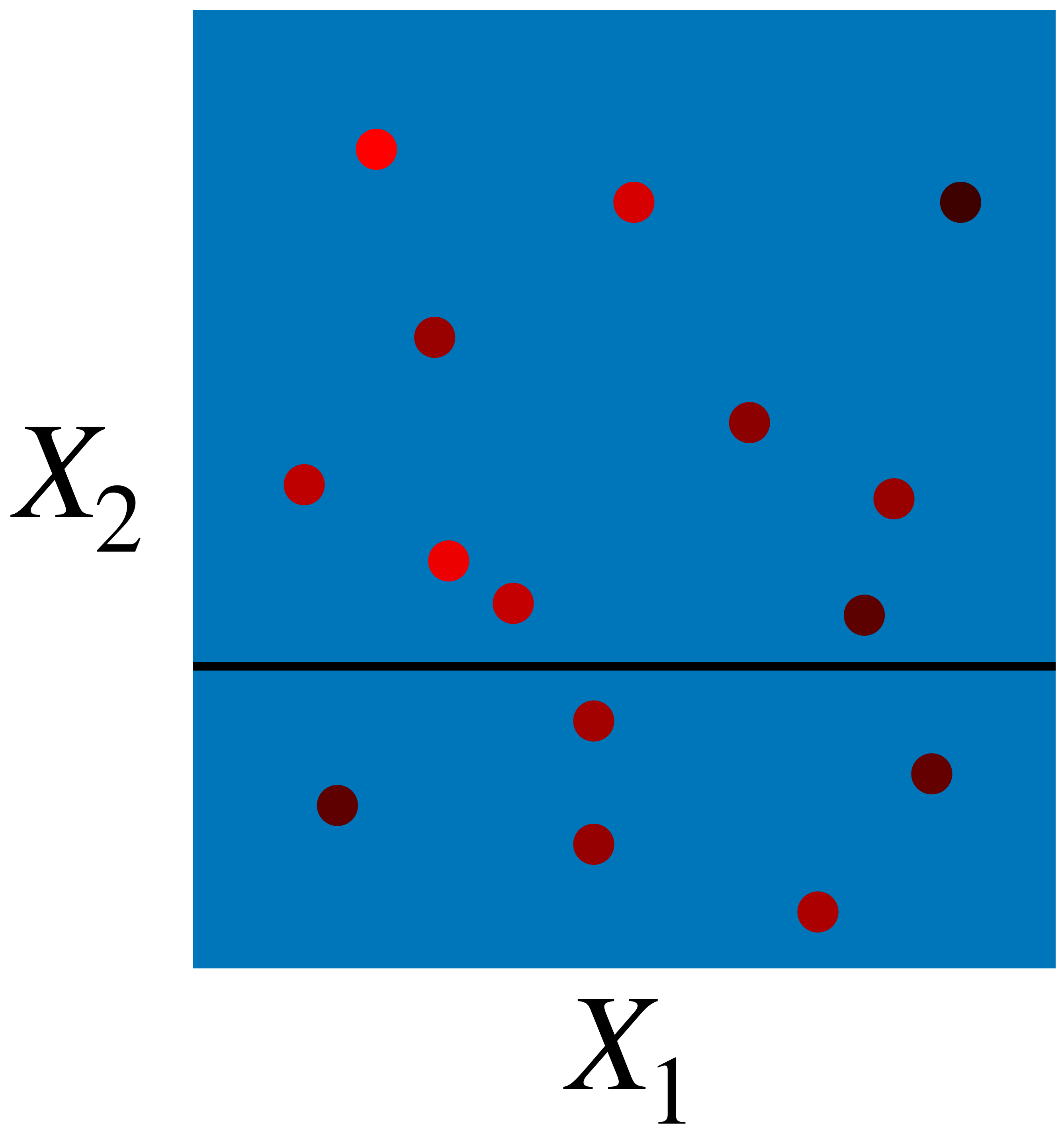
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$



Training a regression tree

Finding the rectangles $\widehat{R}_m$

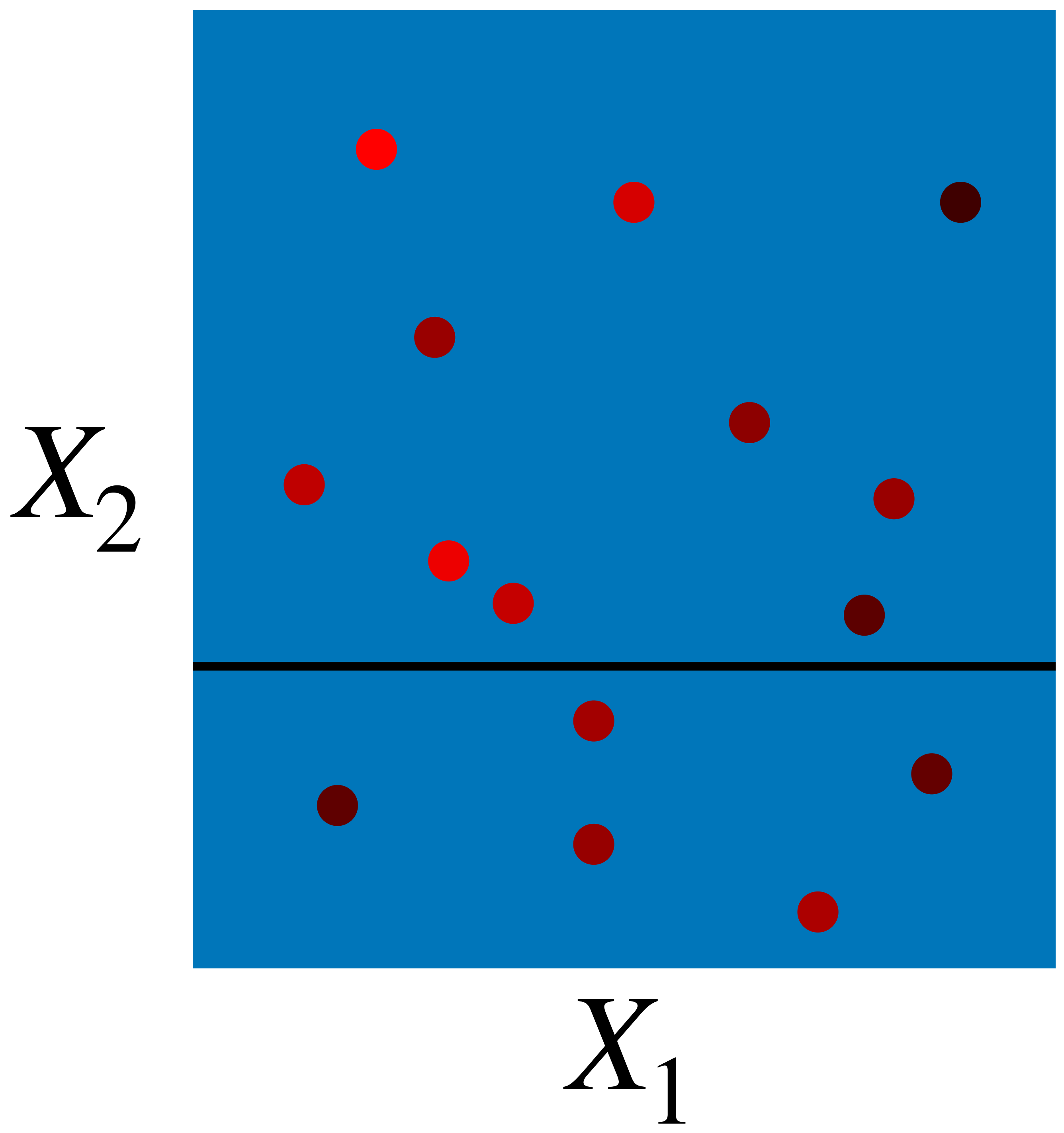
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

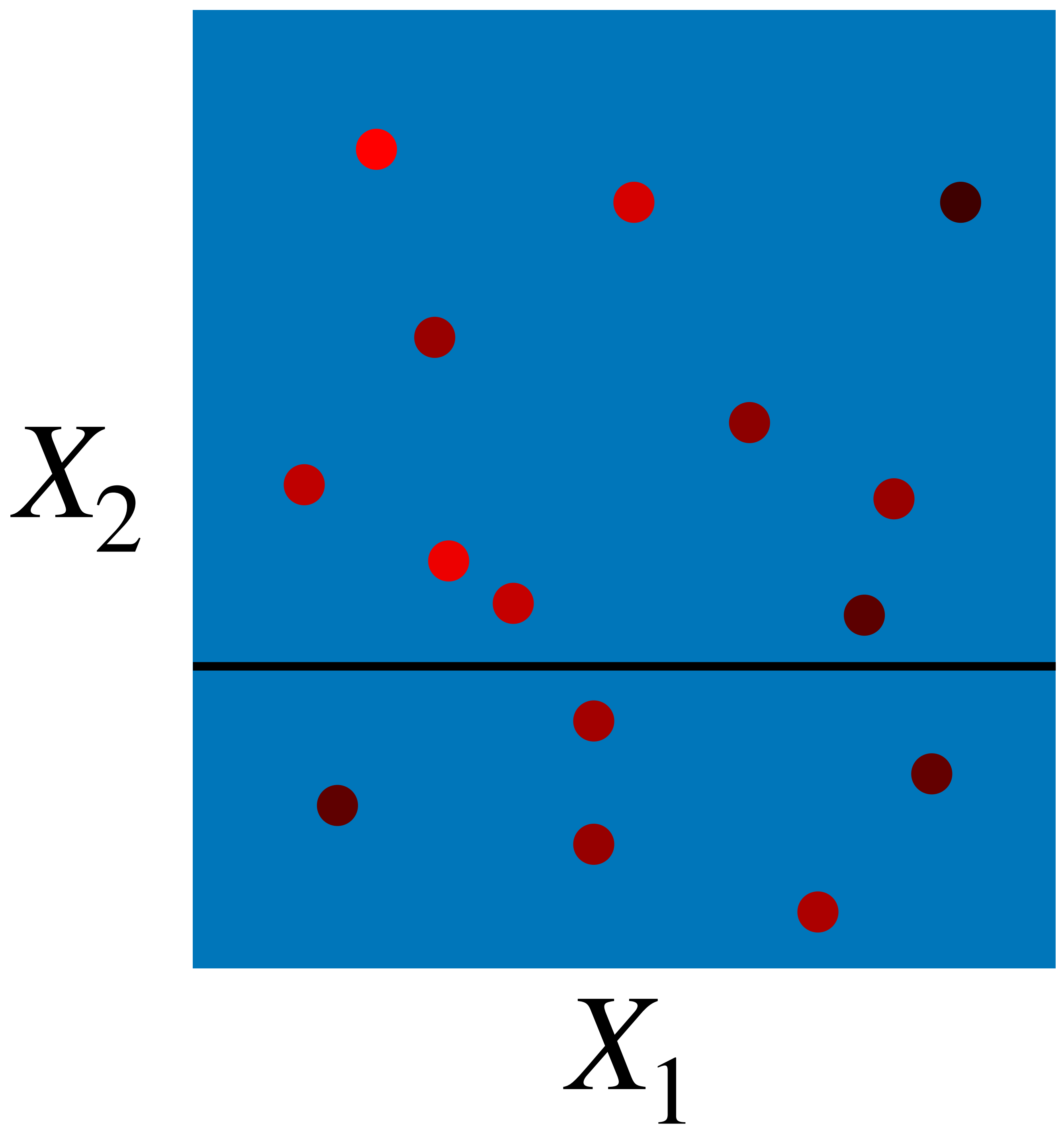
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

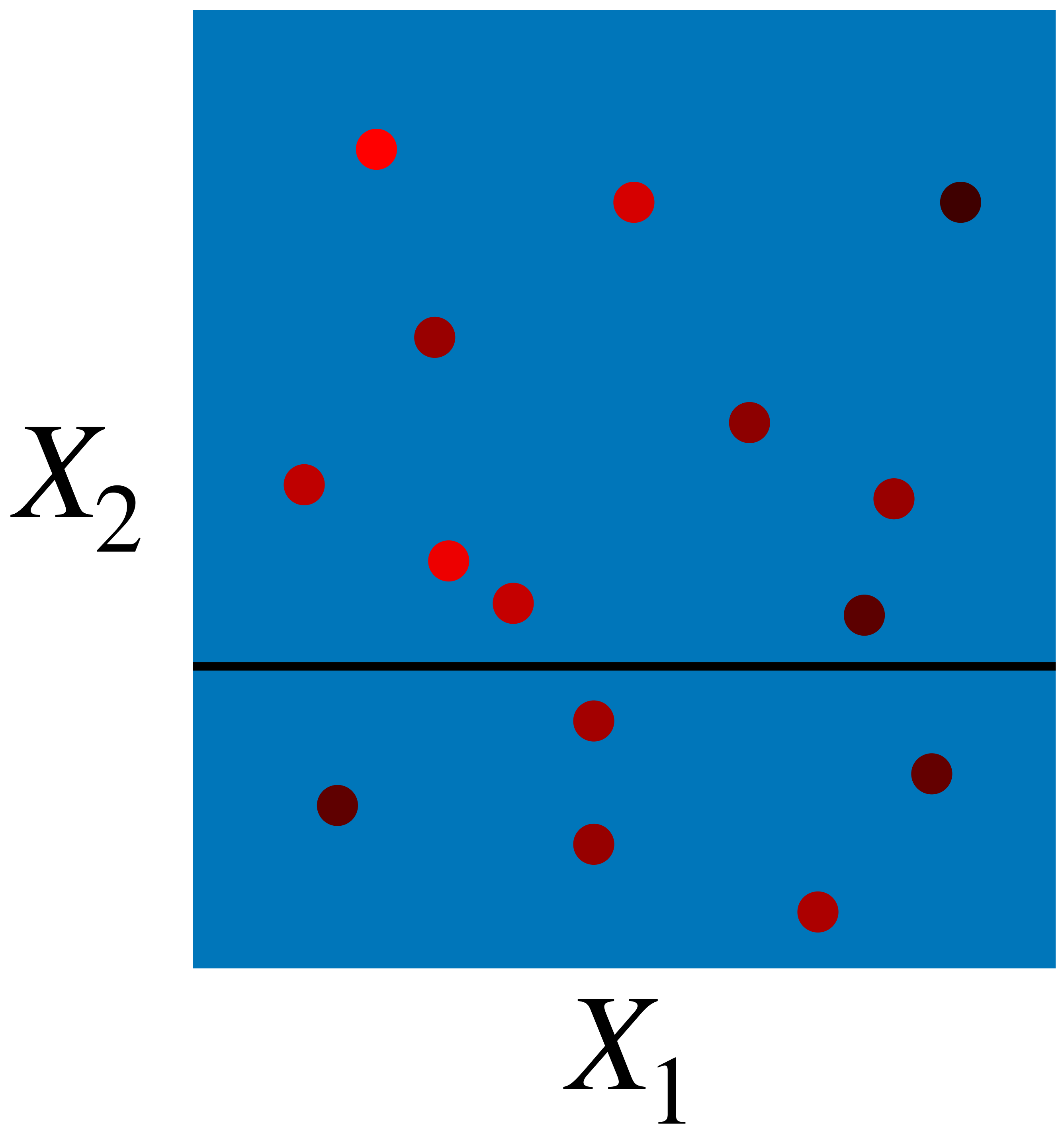
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

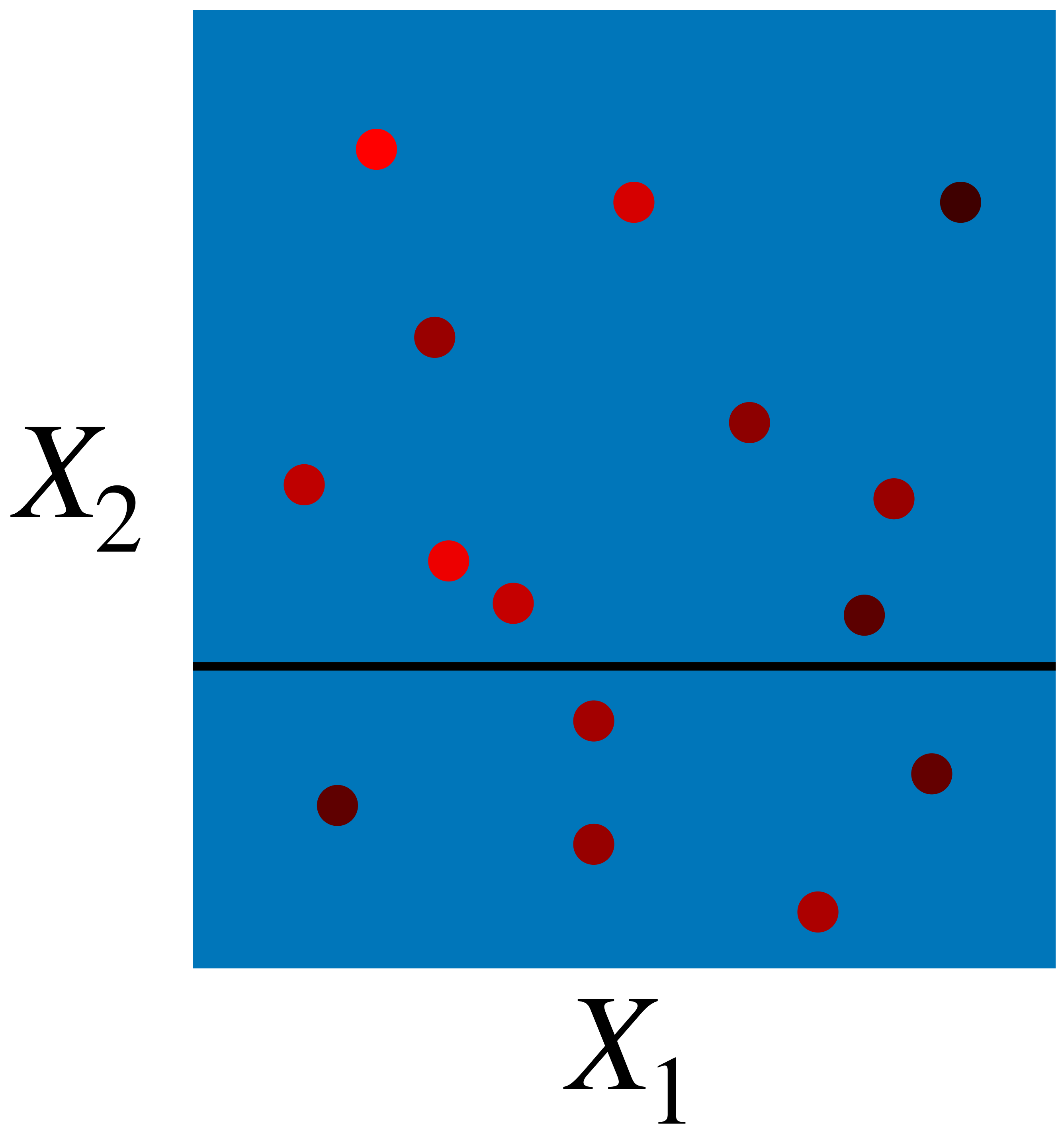
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

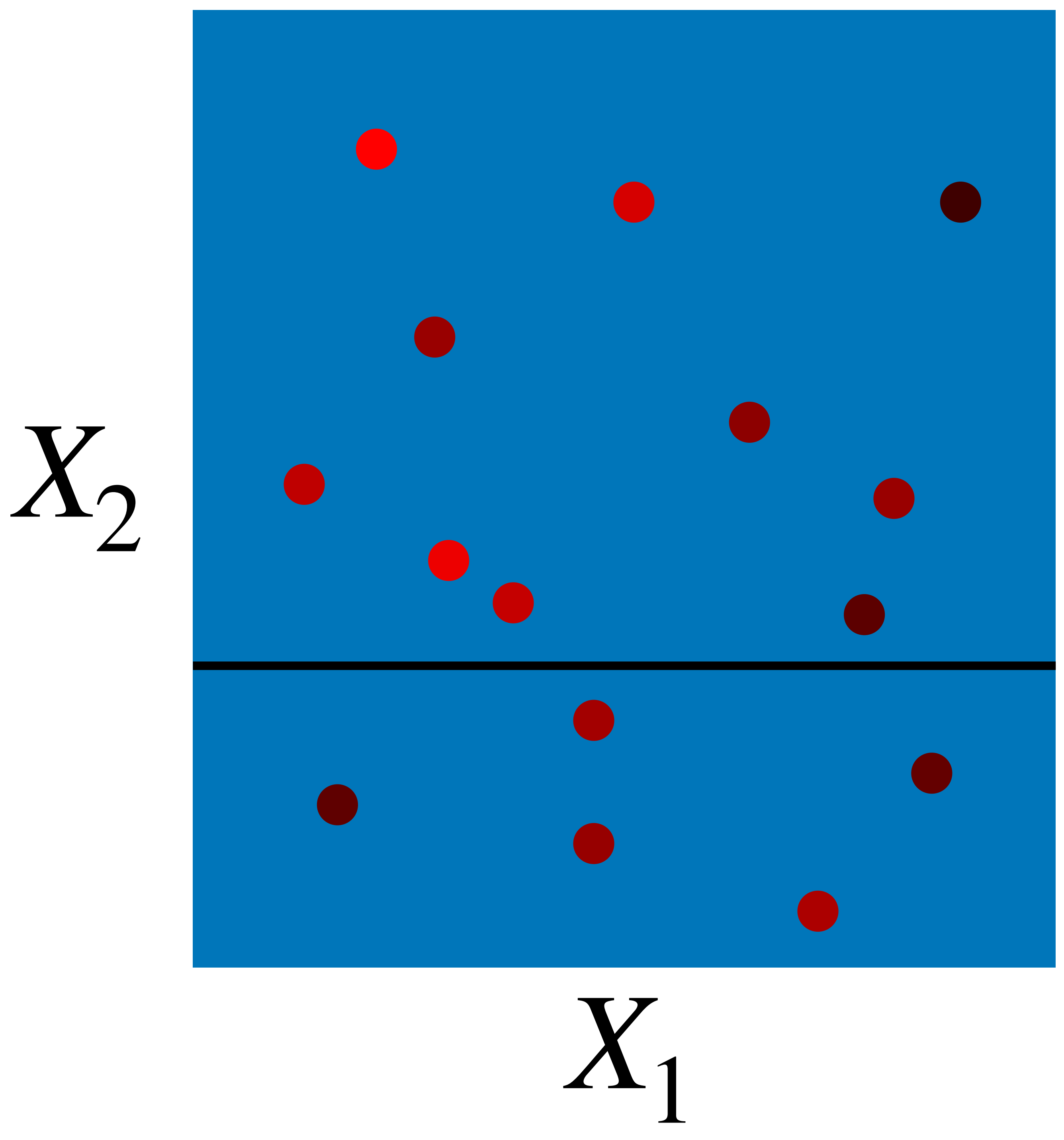
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

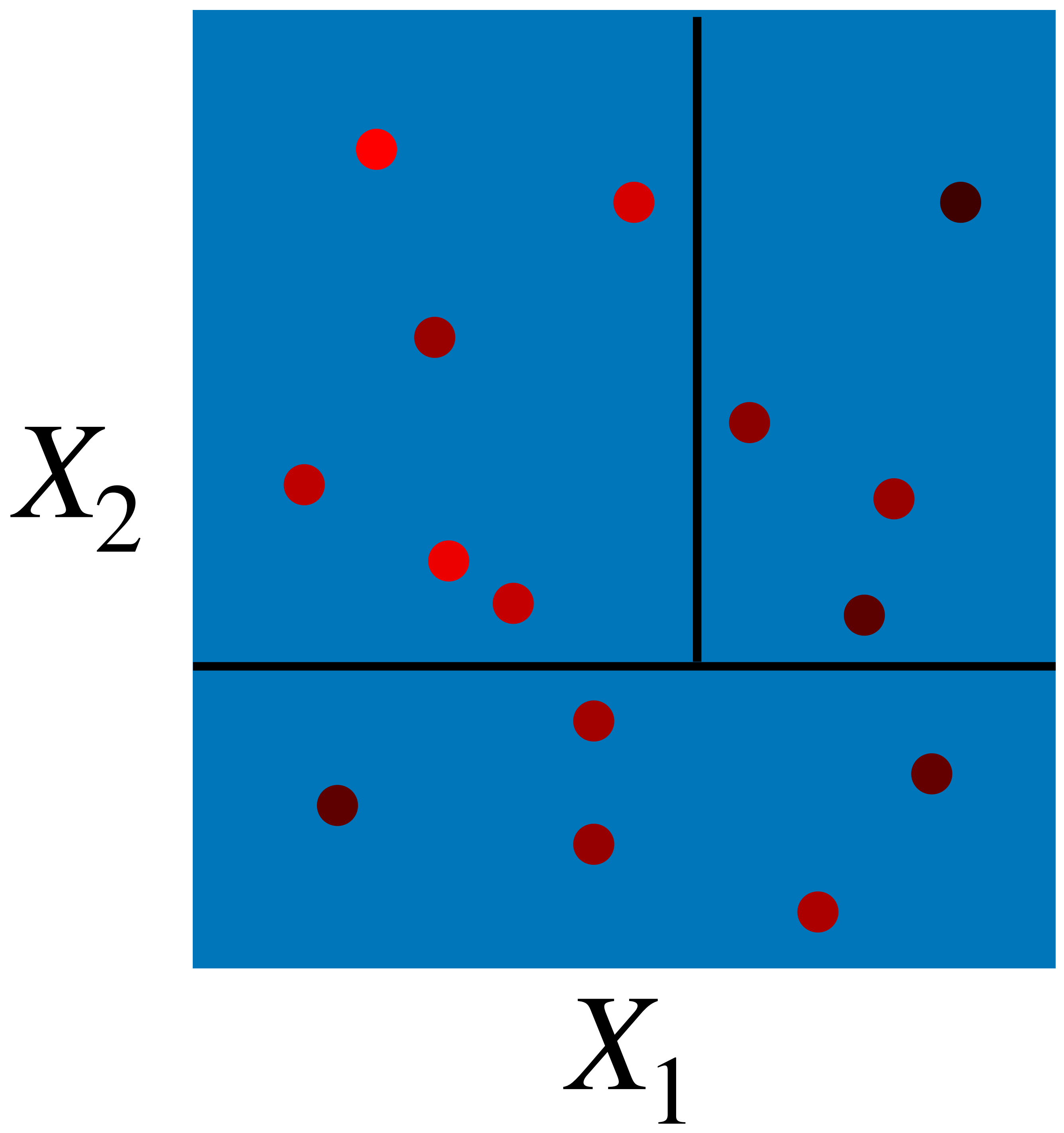
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

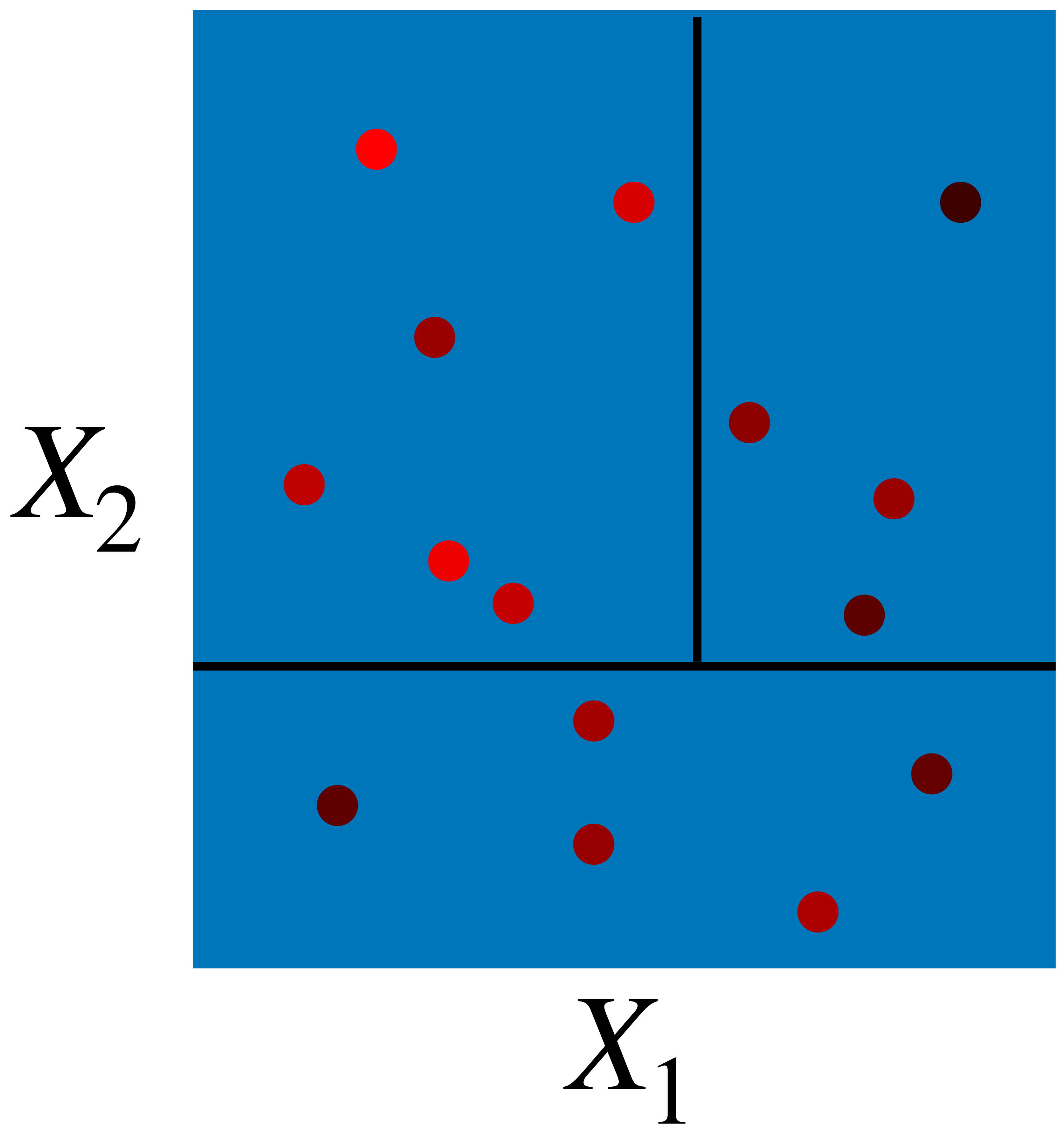
$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.

4. Repeat until there are M regions.



Training a regression tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm:

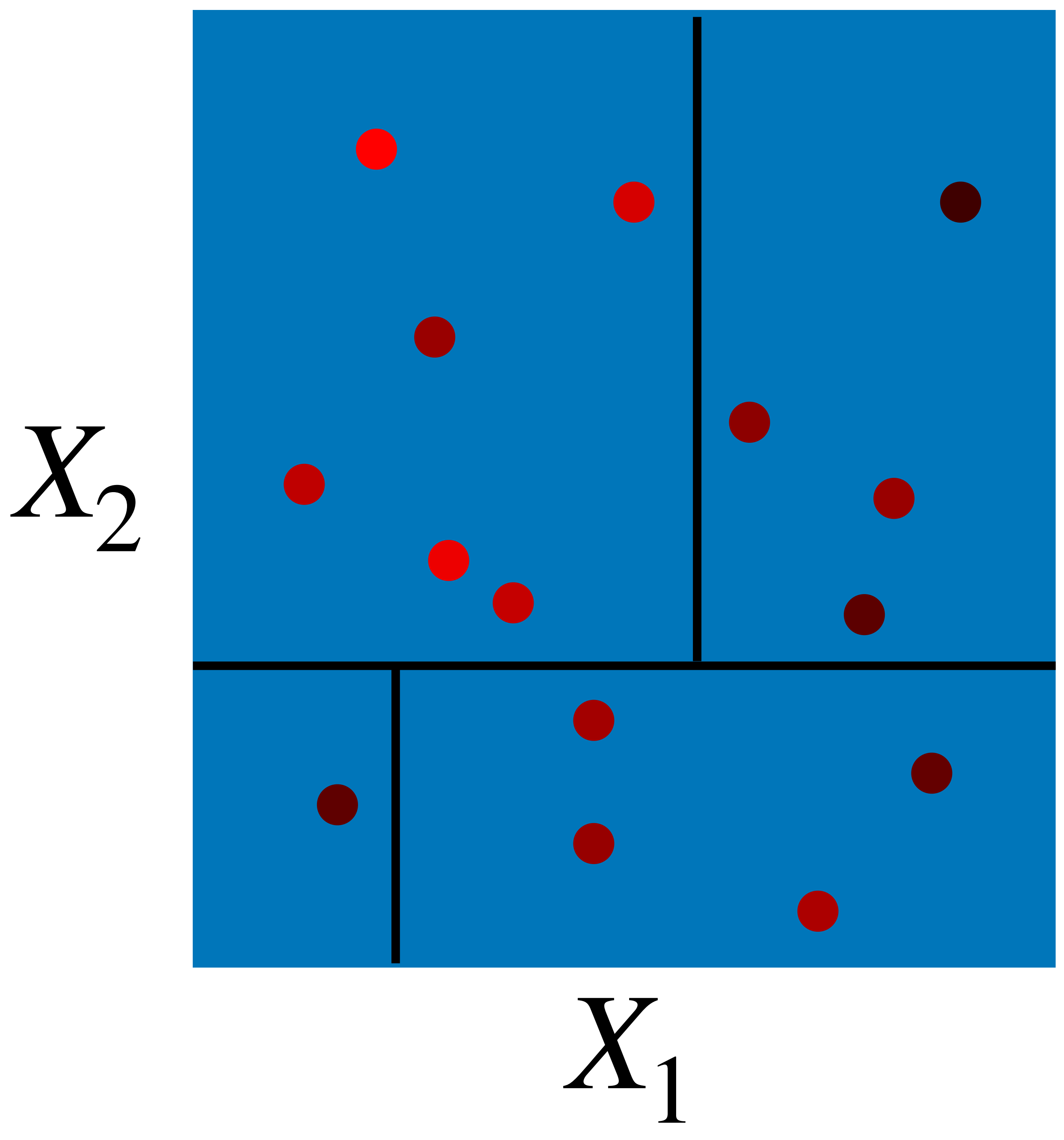
$$\text{RSS} = \sum_{i:X_i \in \widehat{R}_1} (Y_i - c_1)^2 + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i - c_M)^2$$

1. Fit constant model to the entire space and calculate RSS.

2. Find split of the whole region that decreases RSS the most.

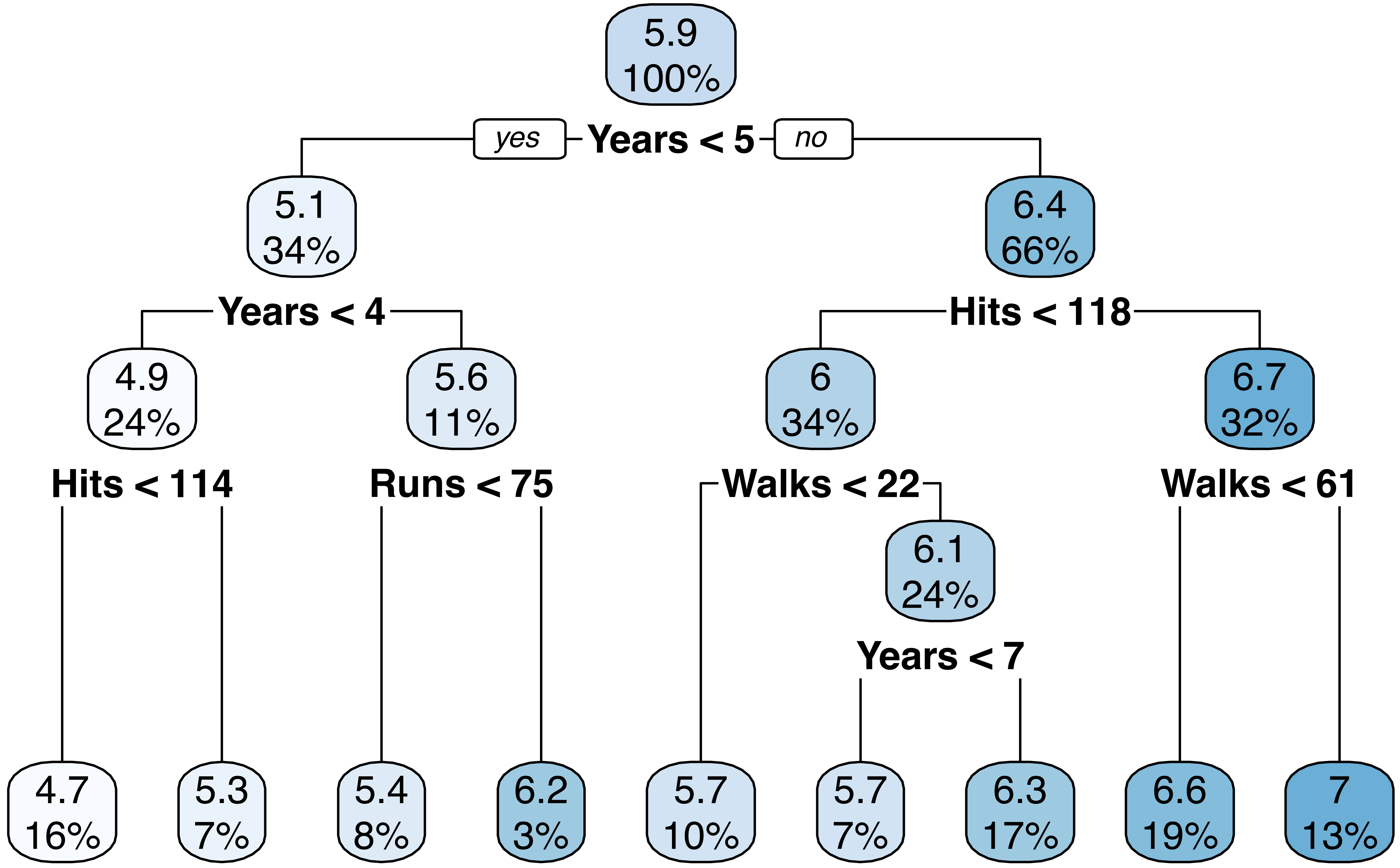
3. Find the split for each rectangle that decreases its RSS the most. Among these, choose best.

4. Repeat until there are M regions.



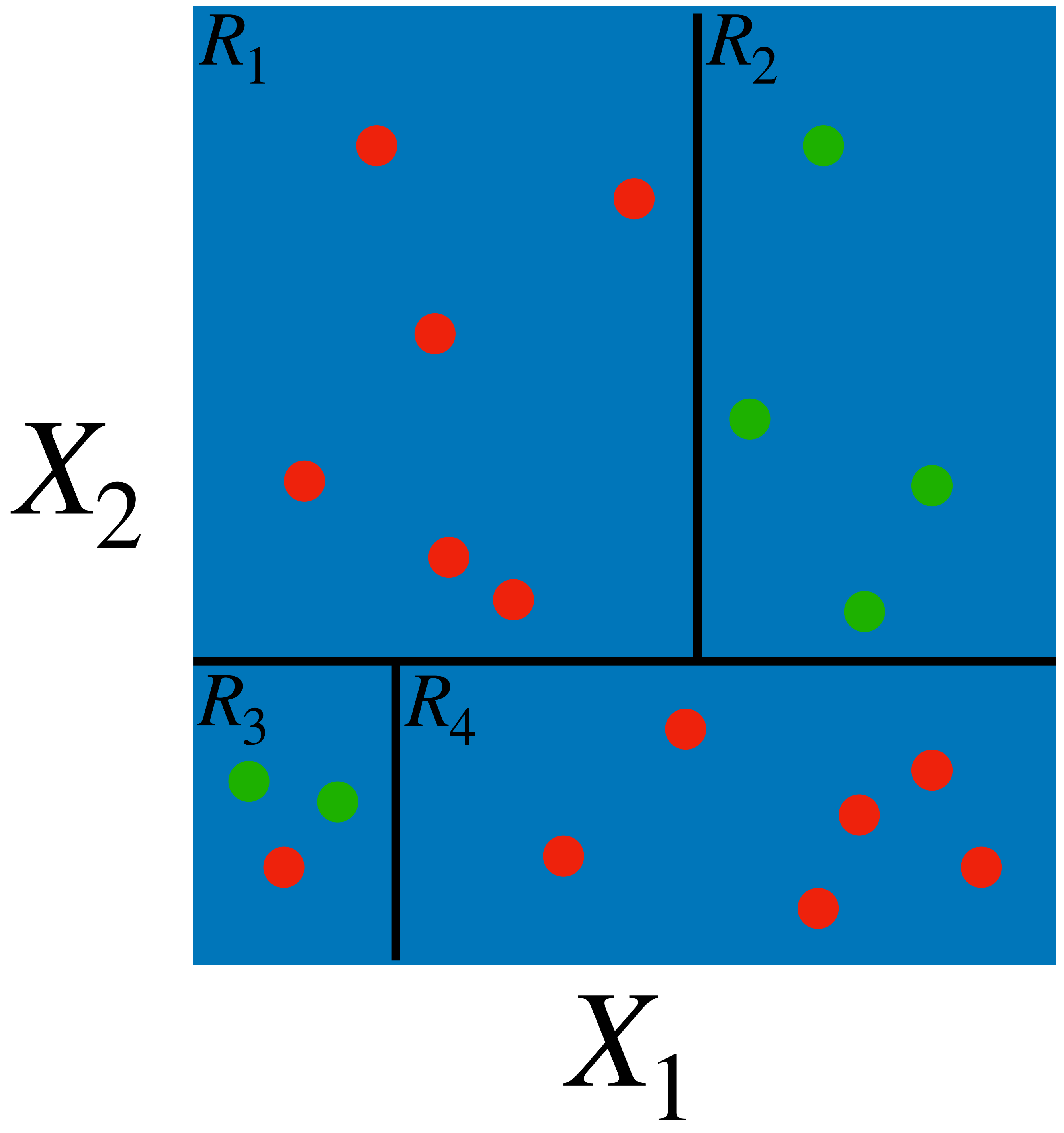
Training a regression tree

Final output



Training a classification tree

The misclassification error objective



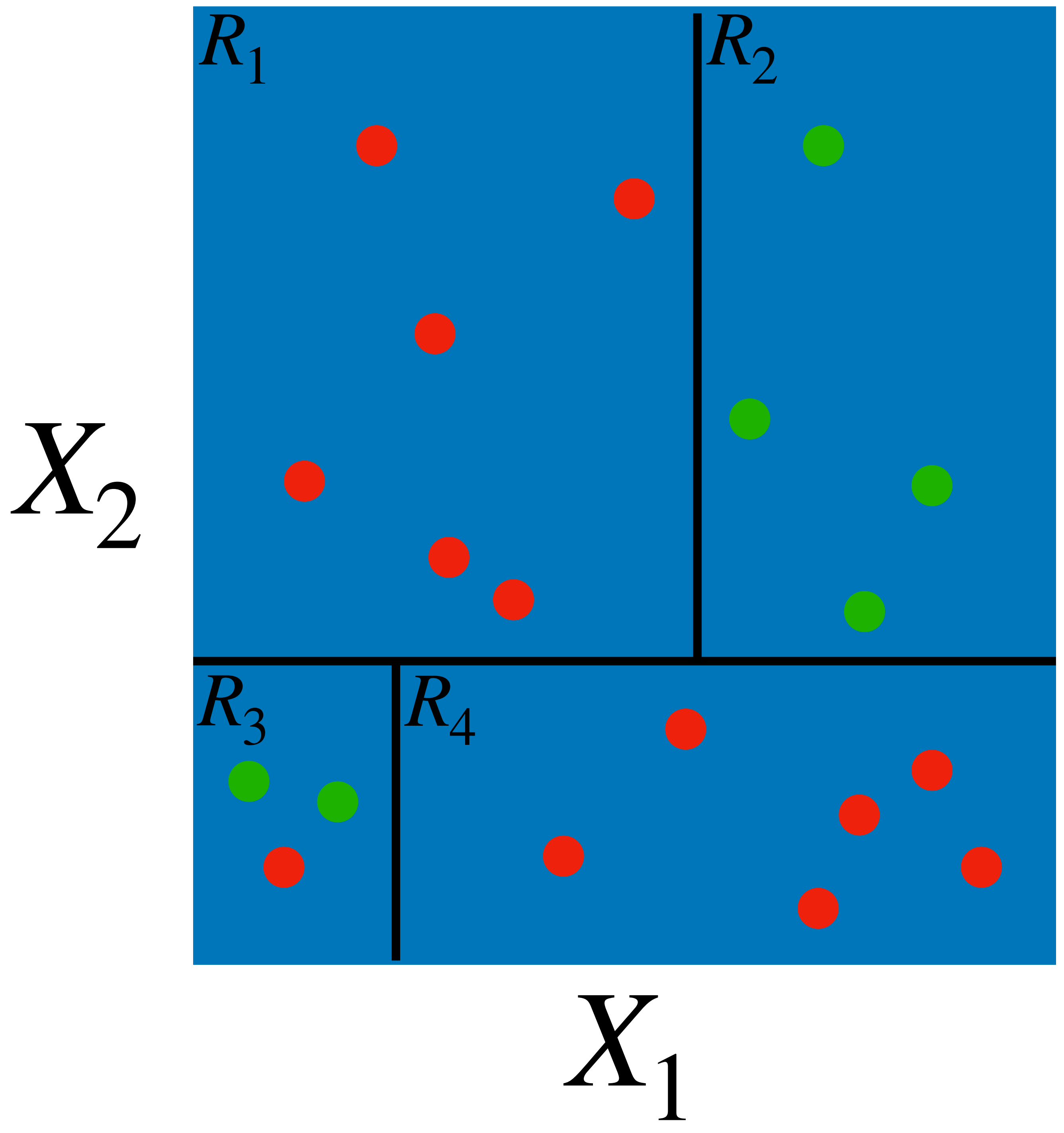
Training a classification tree

The misclassification error objective

As usual, we are given a training dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the response Y is binary.

For a fixed M, we seek rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ and values $\widehat{c}_1, \ldots, \widehat{c}_M$ to minimize the misclassification error:

$$\widehat{R}_1, \ldots, \widehat{R}_M, \widehat{c}_1, \ldots, \widehat{c}_M = \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \widehat{Y}_i)$$



Training a classification tree

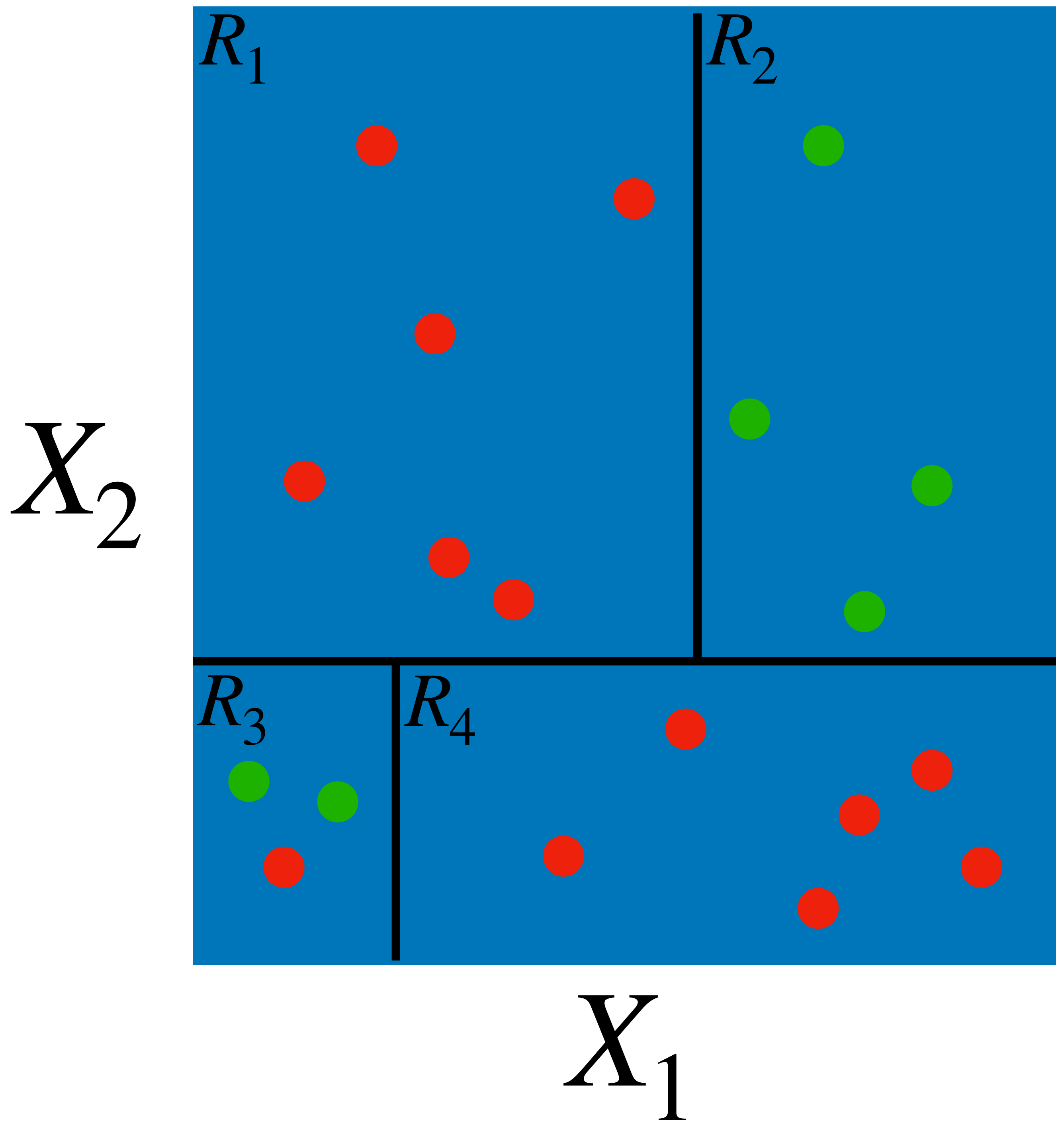
The misclassification error objective

As usual, we are given a training dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the response Y is binary.

For a fixed M, we seek rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ and values $\widehat{c}_1, \ldots, \widehat{c}_M$ to minimize the misclassification error:

$$\widehat{R}_1, \ldots, \widehat{R}_M, \widehat{c}_1, \ldots, \widehat{c}_M = \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \widehat{Y}_i)$$

$$= \underset{R_1,\ldots,R_M,c_1,\ldots,c_M}{\arg\min} \frac{1}{n}\left\{ \sum_{i:X_i\in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i\in R_M} (Y_i \neq c_M) \right\}$$

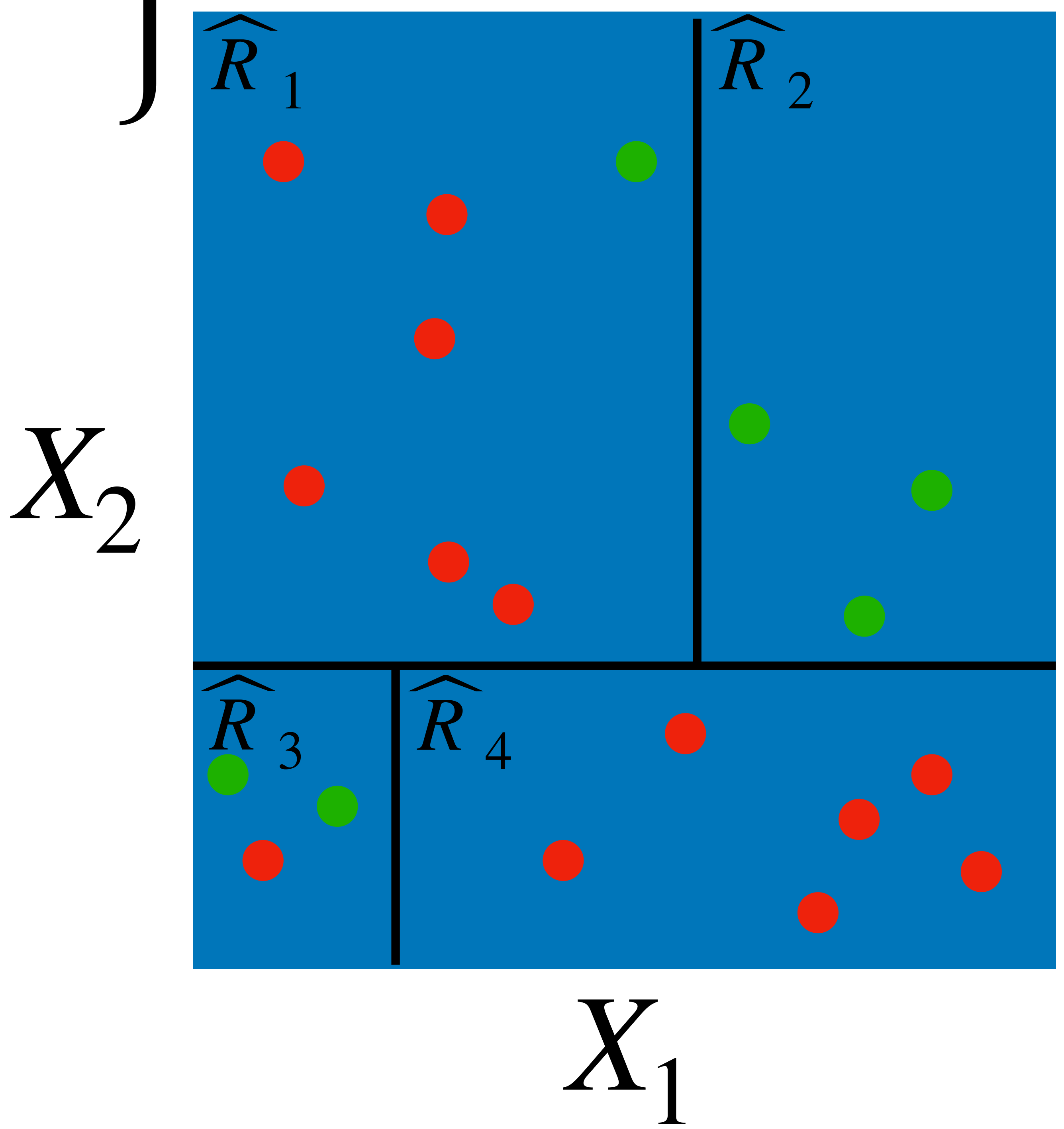


Training a classification tree

Optimal $\widehat{c}_m$ given $\widehat{R}_m$

First let's consider a simpler problem, where rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ are given:

$$\widehat{c}_1, \ldots, \widehat{c}_M = \arg\min_{c_1,\ldots,c_M} \frac{1}{n} \left\{ \sum_{i:X_i \in \widehat{R}_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i \in \widehat{R}_M} (Y_i \neq c_M) \right\}$$



Training a classification tree

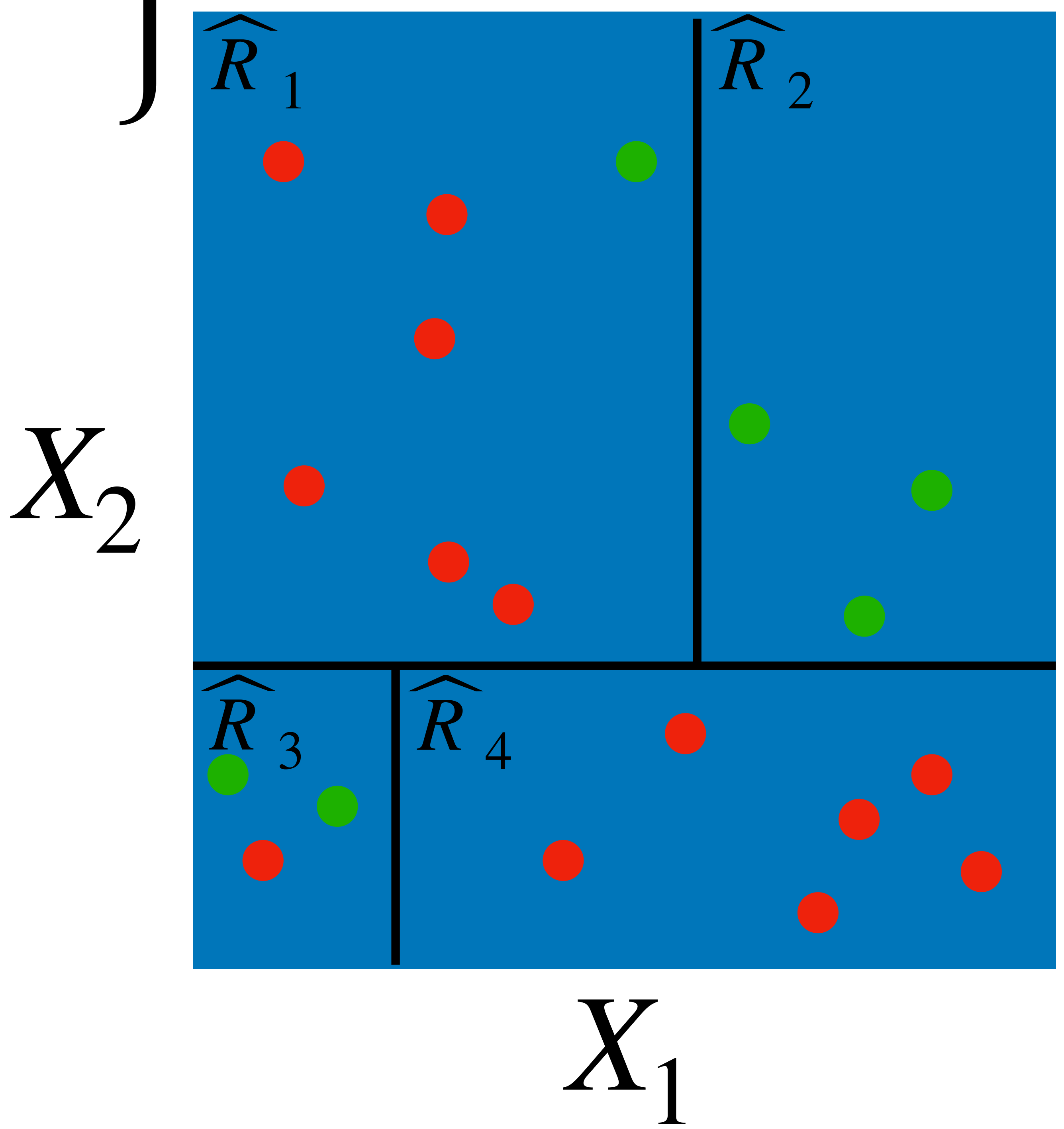
Optimal $\widehat{c}_m$ given $\widehat{R}_m$

First let's consider a simpler problem, where rectangles $\widehat{R}_1, \ldots, \widehat{R}_M$ are given:

$$\widehat{c}_1, \ldots, \widehat{c}_M = \underset{c_1, \ldots, c_M}{\arg\min} \frac{1}{n} \left\{ \sum_{i : X_i \in \widehat{R}_1} I(Y_i \neq c_1) + \cdots + \sum_{i : X_i \in \widehat{R}_M} (Y_i \neq c_M) \right\}$$

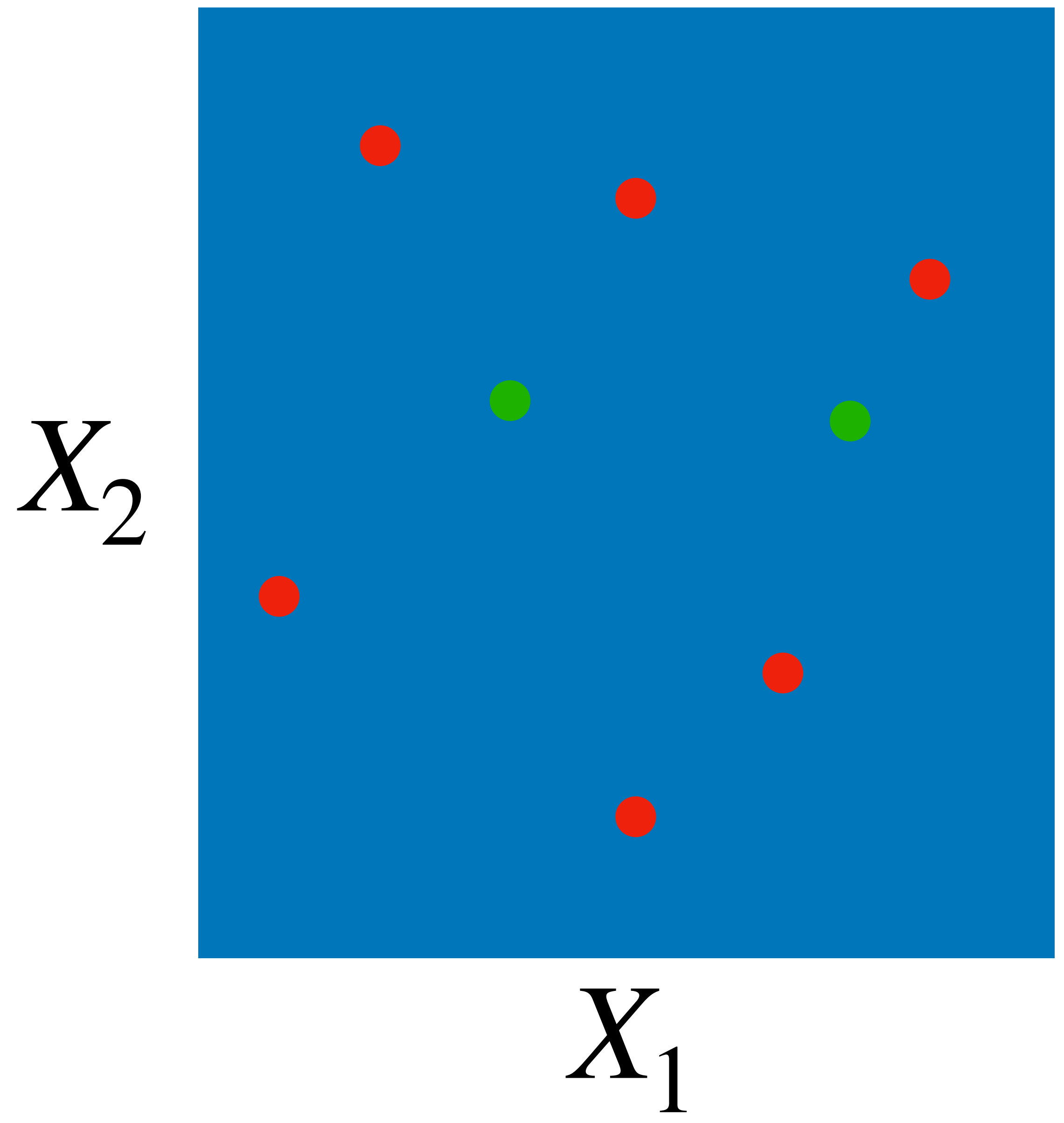
We're fitting the same category to each region, so the solution is the majority vote:

$$\widehat{c}_m = \text{mode}\left(\{ Y_i : X_i \in \widehat{R}_m \} \right).$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$: Inadequacy of misclassification error

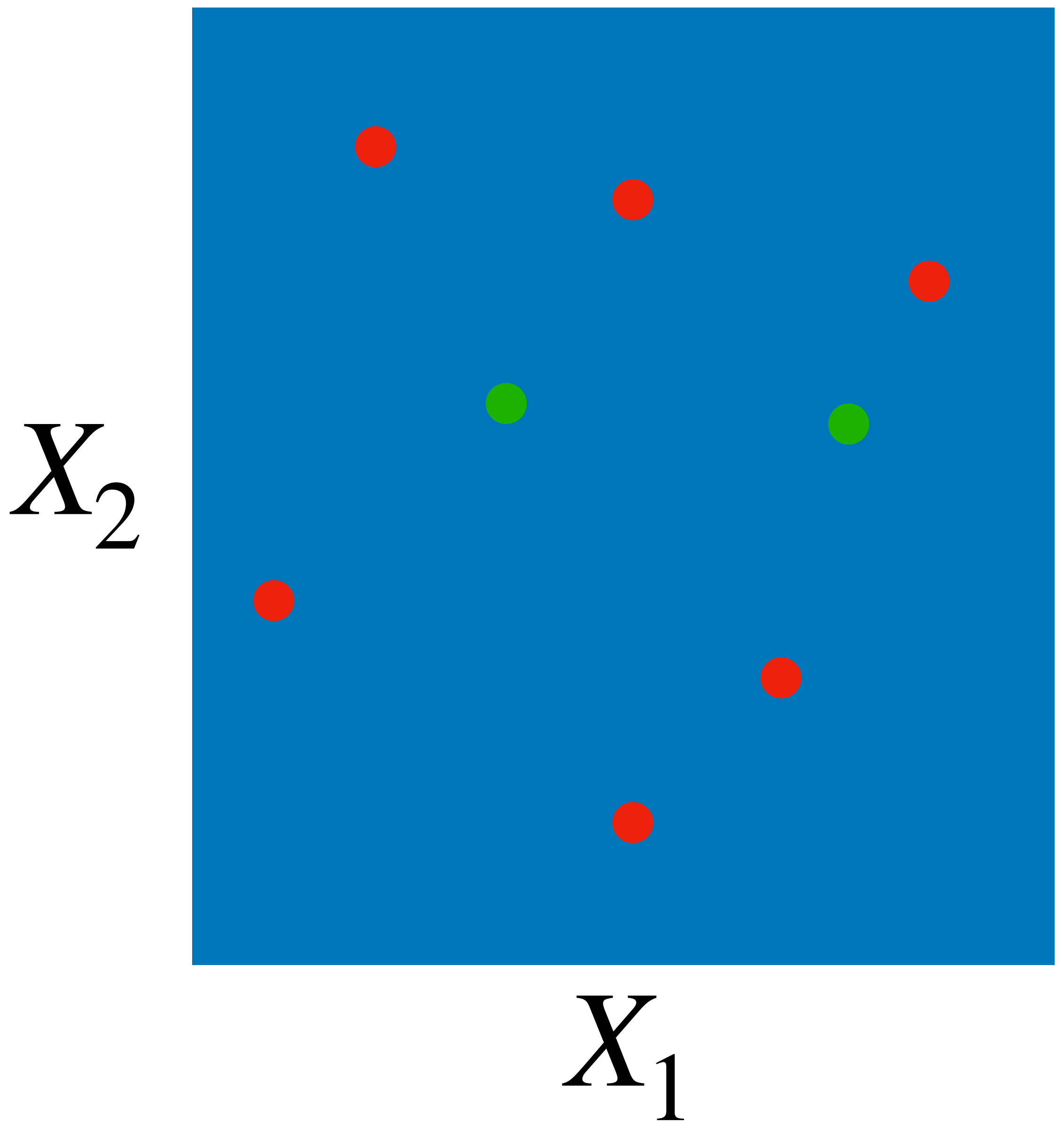


Training a classification tree

Finding the rectangles $\widehat{R}_m$: Inadequacy of misclassification error

Misclassification error:

$$\frac{1}{n}\left\{\sum_{i:X_i\in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i\in R_M} (Y_i \neq c_M)\right\}$$

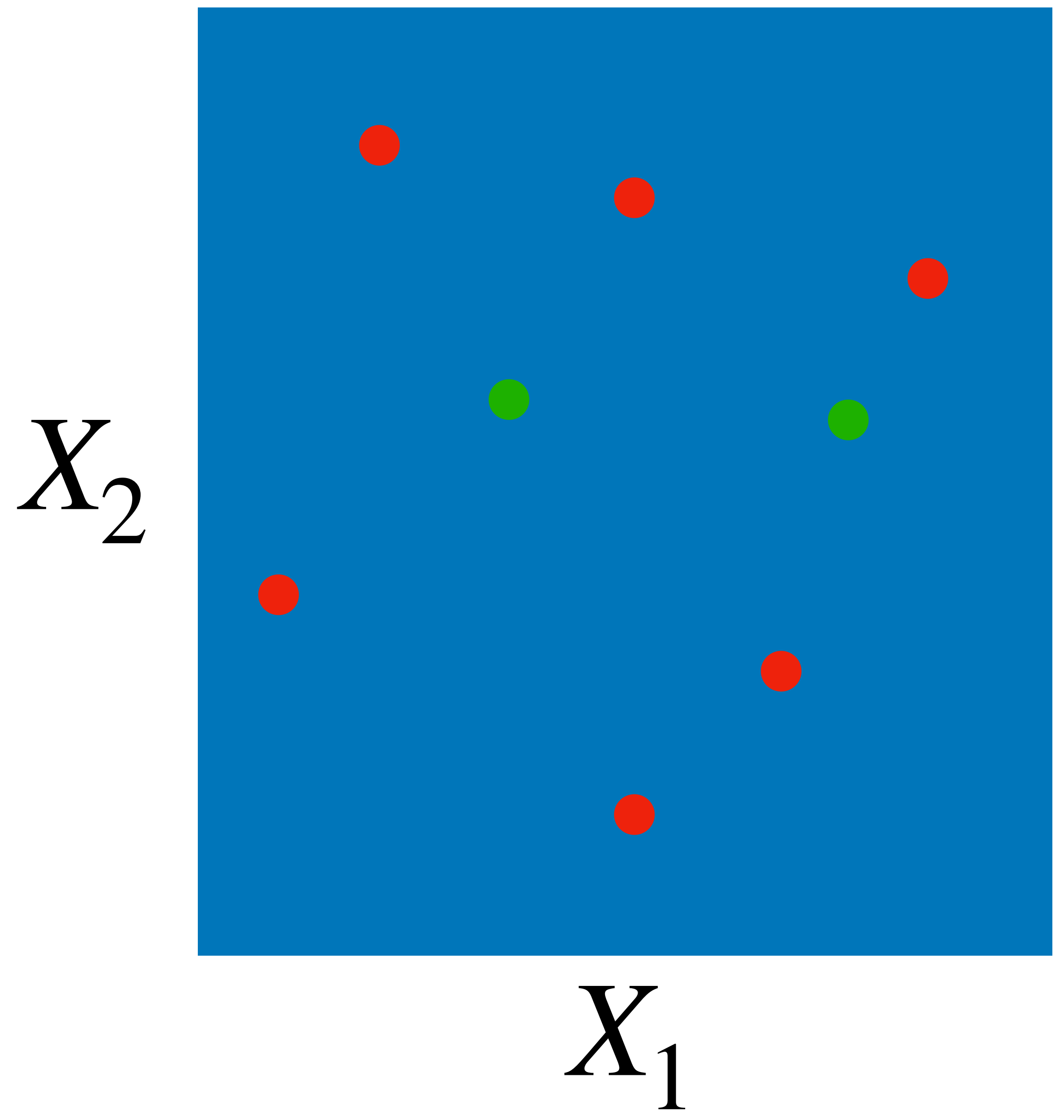


Training a classification tree

Finding the rectangles $\widehat{R}_m$: Inadequacy of misclassification error

Misclassification error:

$$\frac{1}{n}\left\{\sum_{i:X_i\in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i\in R_M} (Y_i \neq c_M)\right\}$$



In example at right, no choice of split point decreases misclassification error.

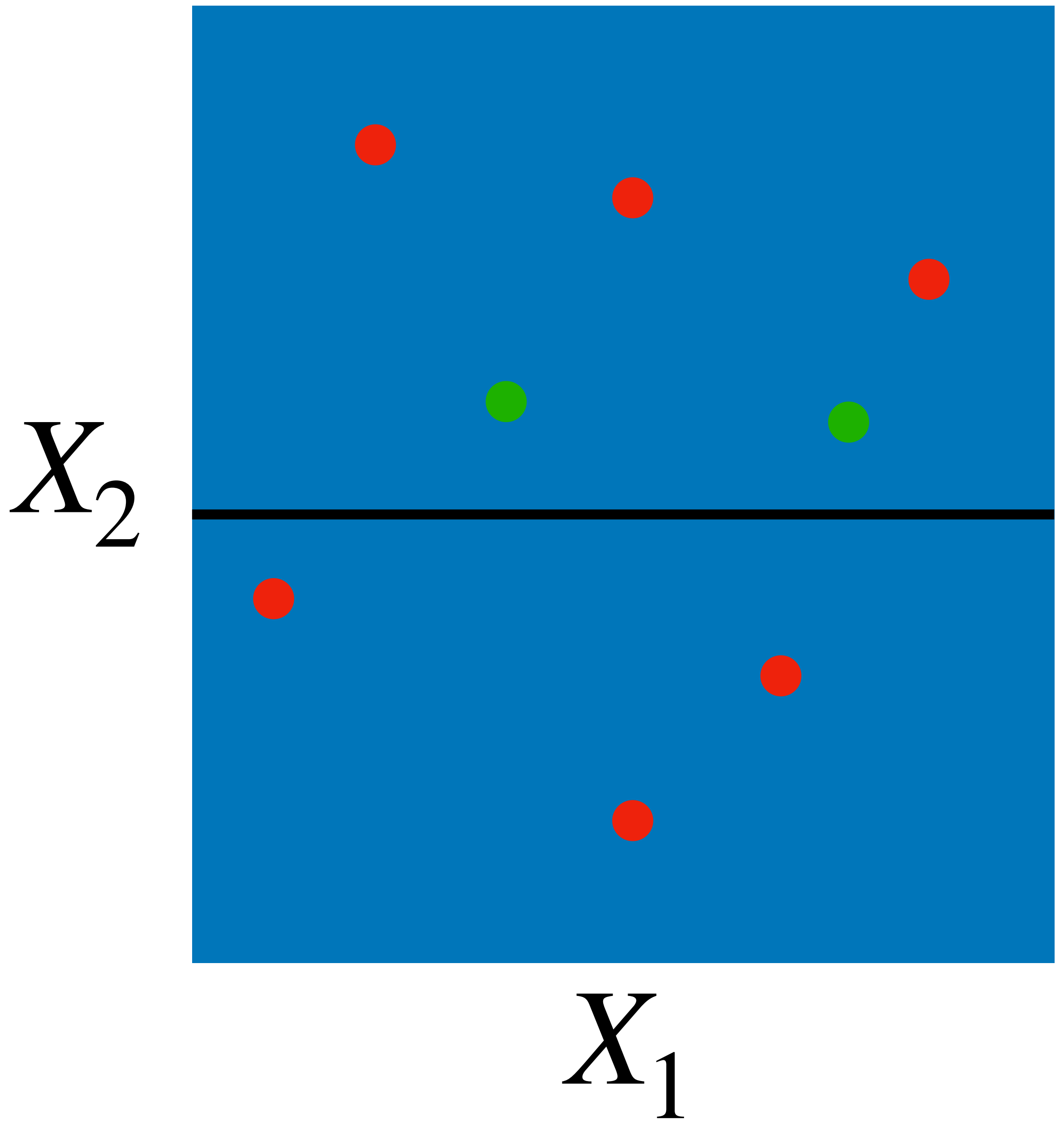
Training a classification tree

Finding the rectangles $\widehat{R}_m$: Inadequacy of misclassification error

Misclassification error:

$$\frac{1}{n}\left\{\sum_{i:X_i\in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i\in R_M} (Y_i \neq c_M)\right\}$$

In example at right, no choice of split point decreases misclassification error.

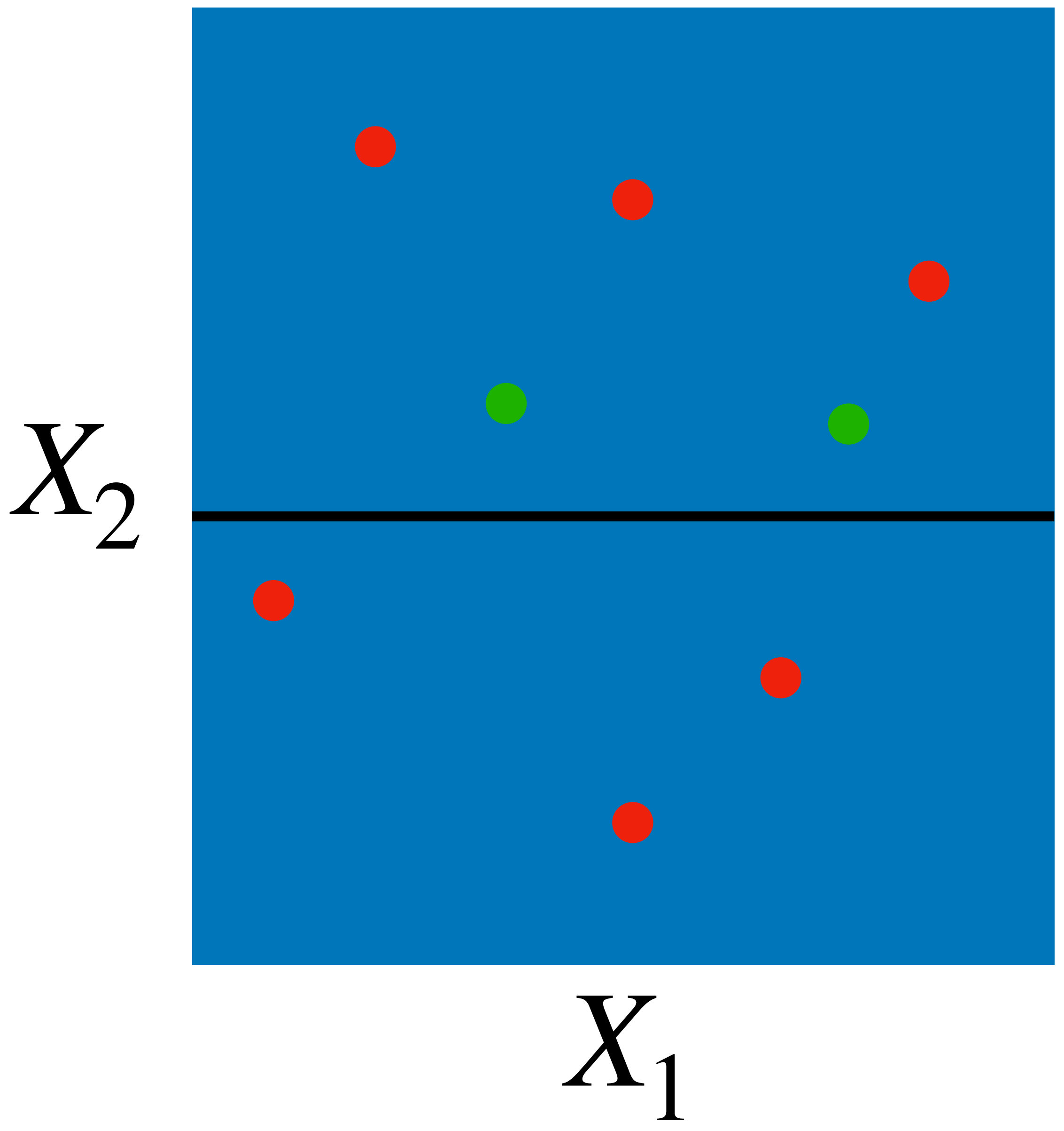


Training a classification tree

Finding the rectangles $\widehat{R}_m$: Inadequacy of misclassification error

Misclassification error:

$$\frac{1}{n}\left\{\sum_{i:X_i\in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i\in R_M} (Y_i \neq c_M)\right\}$$



In example at right, no choice of split point decreases misclassification error.

Misclassification error not sensitive enough to find good split points at each step.

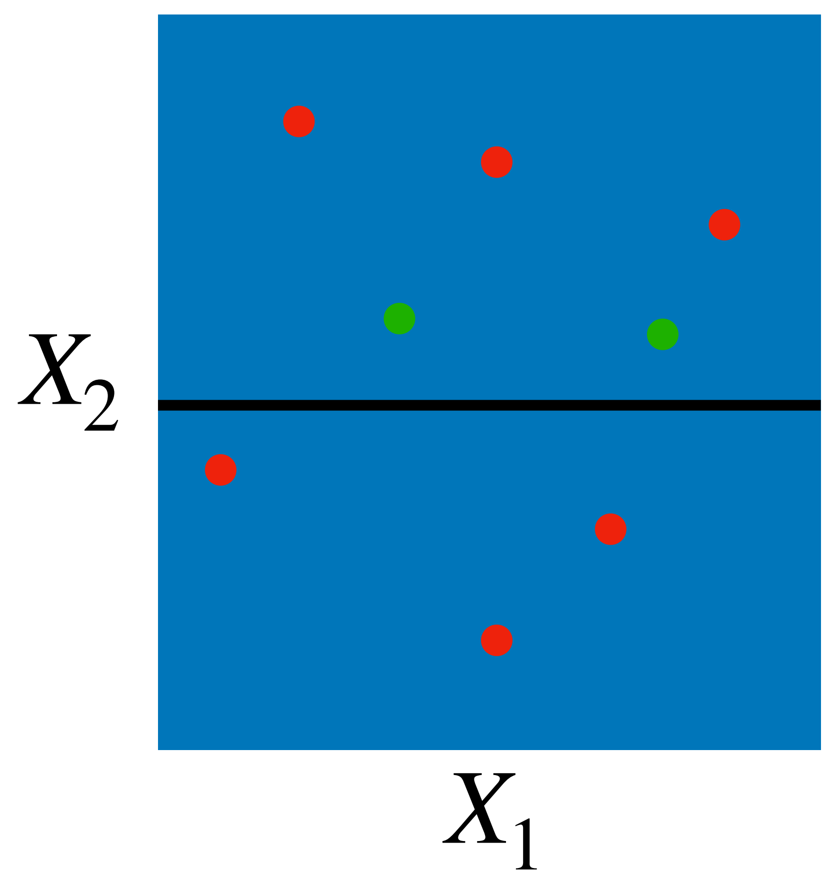
Training a classification tree

Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

Training a classification tree

Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

If $\hat{p}_m$ is region m class 1 proportion, misclassification error in that region is $\min(\hat{p}_m, 1 - \hat{p}_m)$.



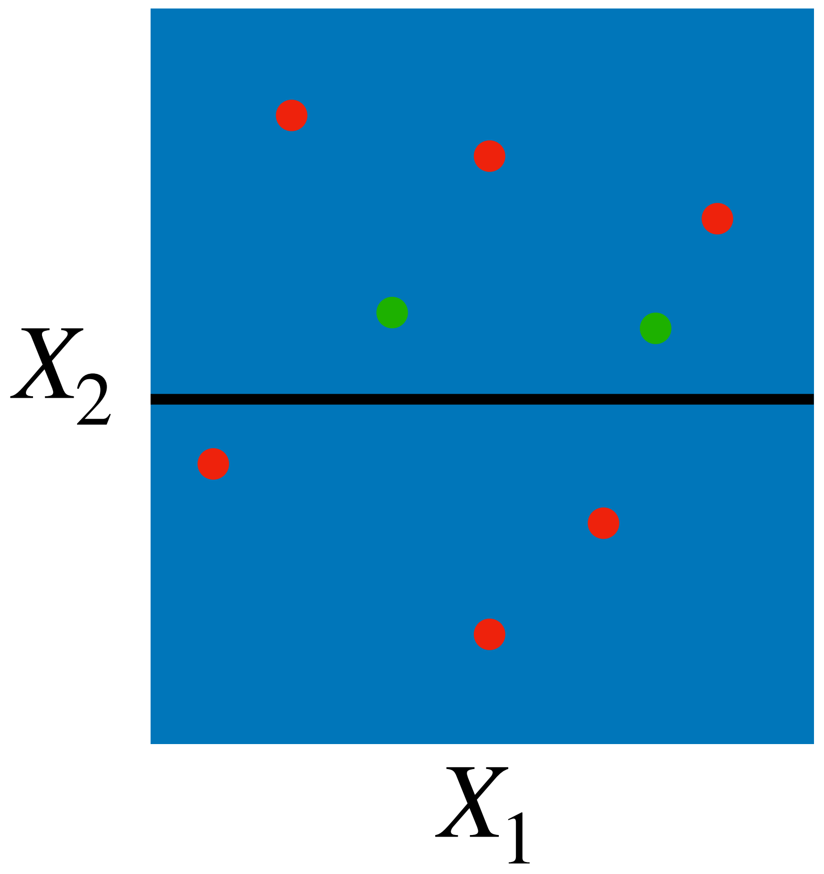
Training a classification tree

Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

If $\widehat{p}_m$ is region m class 1 proportion, misclassification error in that region is $\min(\widehat{p}_m, 1 - \widehat{p}_m)$.

If n_m is the number of training observations in that region, then

$$\text{Total misclassification error} = \frac{1}{n}\left\{ \sum_{i:X_i \in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i \in R_M} (Y_i \neq c_M) \right\}$$



Training a classification tree

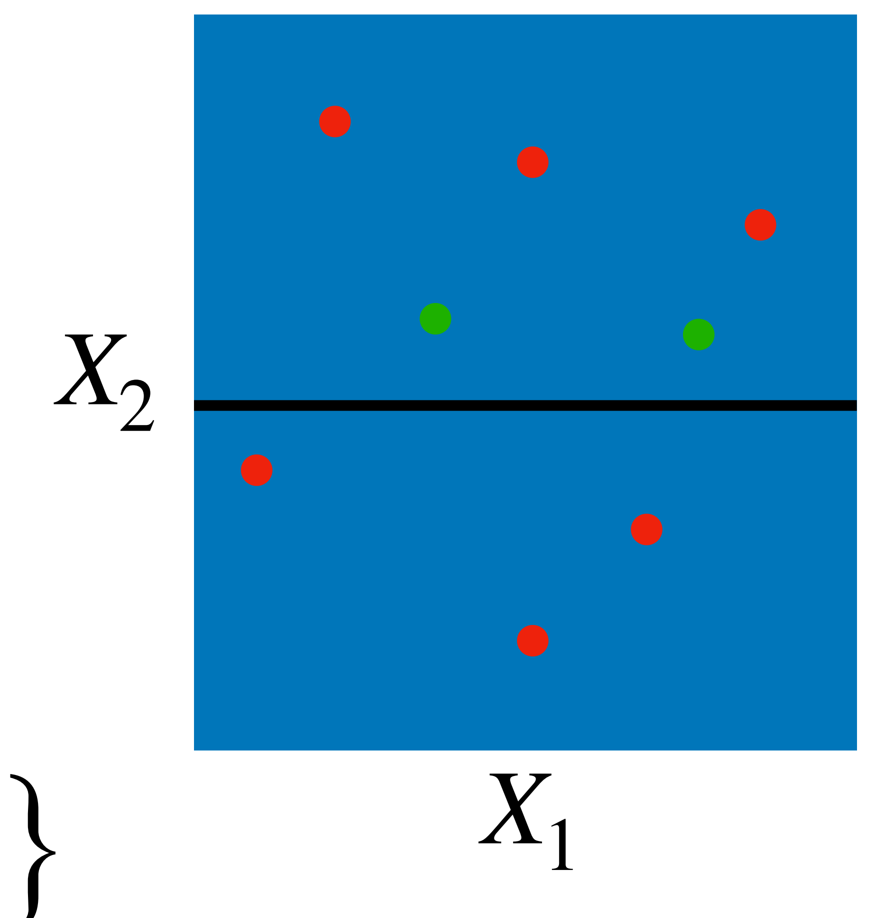
Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

If $\widehat{p}_m$ is region m class 1 proportion, misclassification error in that region is $\min(\widehat{p}_m, 1 - \widehat{p}_m)$.

If n_m is the number of training observations in that region, then

$$\text{Total misclassification error} = \frac{1}{n}\left\{ \sum_{i:X_i \in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i \in R_M} (Y_i \neq c_M) \right\}$$

$$= \frac{1}{n}\left\{ n_1 \min(\hat{p}_1, 1 - \hat{p}_1) + \cdots + n_M \min(\hat{p}_M, 1 - \hat{p}_M) \right\}$$



Training a classification tree

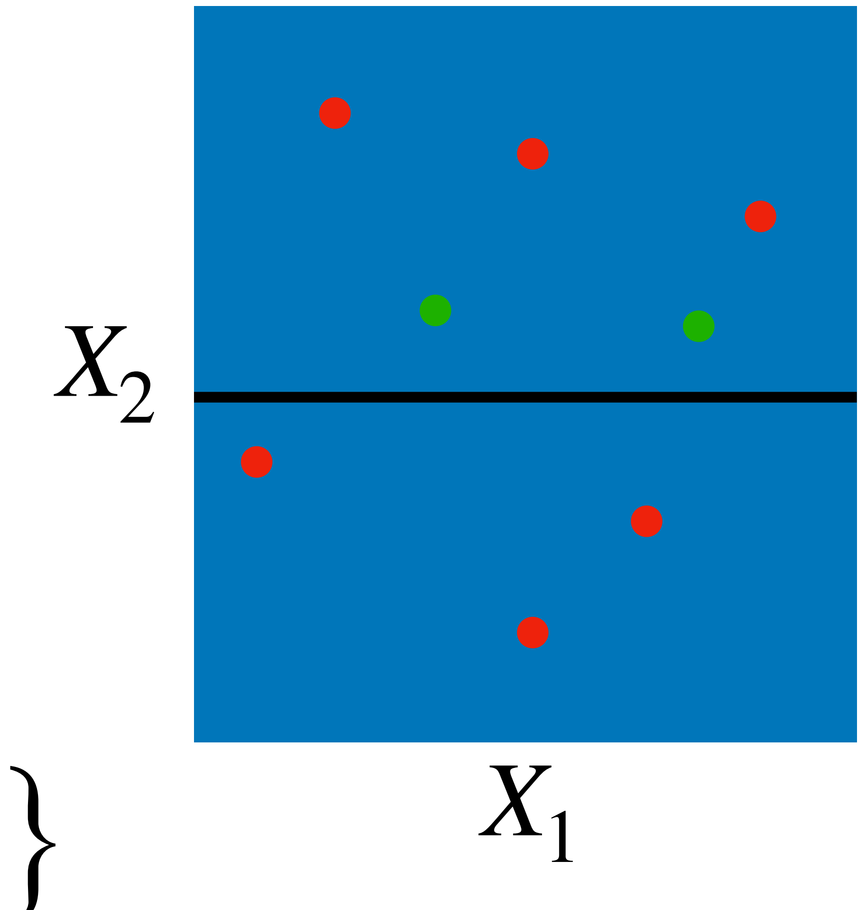
Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

If $\widehat{p}_m$ is region m class 1 proportion, misclassification error in that region is $\min(\widehat{p}_m, 1 - \widehat{p}_m)$.

If n_m is the number of training observations in that region, then

$$\text{Total misclassification error} = \frac{1}{n}\left\{ \sum_{i:X_i \in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i \in R_M} (Y_i \neq c_M) \right\}$$

$$= \frac{1}{n}\left\{ n_1 \min(\hat{p}_1, 1 - \hat{p}_1) + \cdots + n_M \min(\hat{p}_M, 1 - \hat{p}_M) \right\}$$



Replace misclassification error $\min(\widehat{p}_m, 1 - \widehat{p}_m)$ by the Gini index $= 2\widehat{p}_m(1 - \widehat{p}_m)$:

Training a classification tree

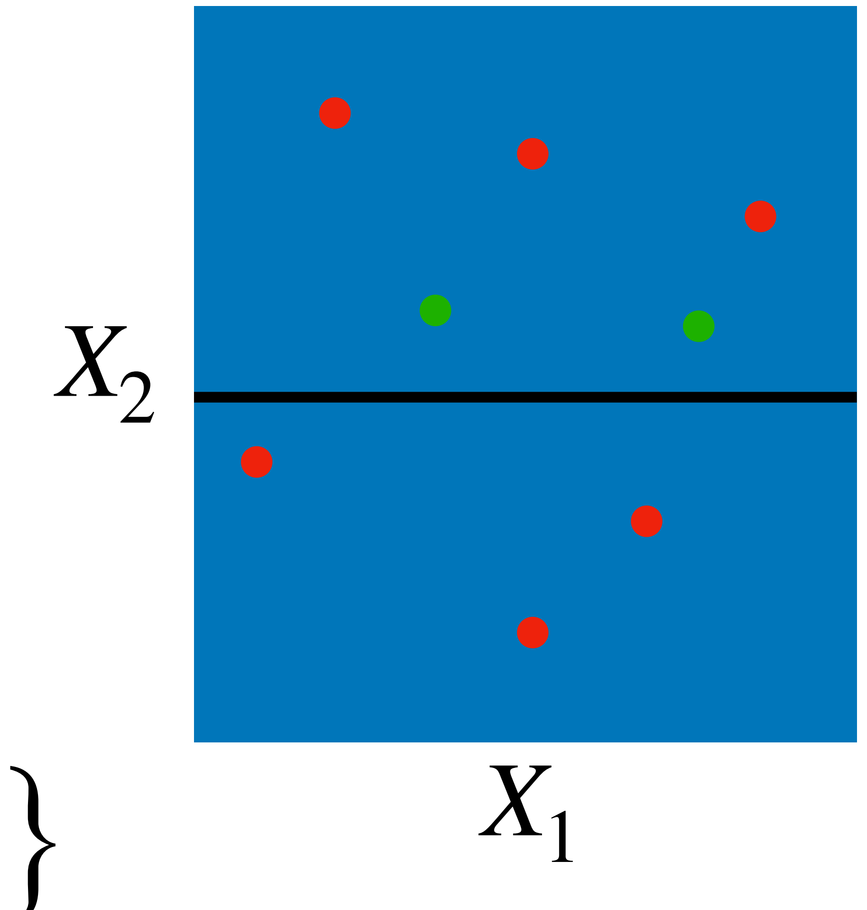
Finding the rectangles $\widehat{R}_m$: The Gini index (measure of node impurity)

If $\hat{p}_m$ is region m class 1 proportion, misclassification error in that region is $\min(\hat{p}_m, 1 - \hat{p}_m)$.

If n_m is the number of training observations in that region, then

$$\text{Total misclassification error} = \frac{1}{n}\left\{ \sum_{i:X_i \in R_1} I(Y_i \neq c_1) + \cdots + \sum_{i:X_i \in R_M} (Y_i \neq c_M) \right\}$$

$$= \frac{1}{n}\left\{ n_1 \min(\hat{p}_1, 1 - \hat{p}_1) + \cdots + n_M \min(\hat{p}_M, 1 - \hat{p}_M) \right\}$$



Replace misclassification error $\min(\hat{p}_m, 1 - \hat{p}_m)$ by the Gini index $= 2\hat{p}_m(1 - \hat{p}_m)$:

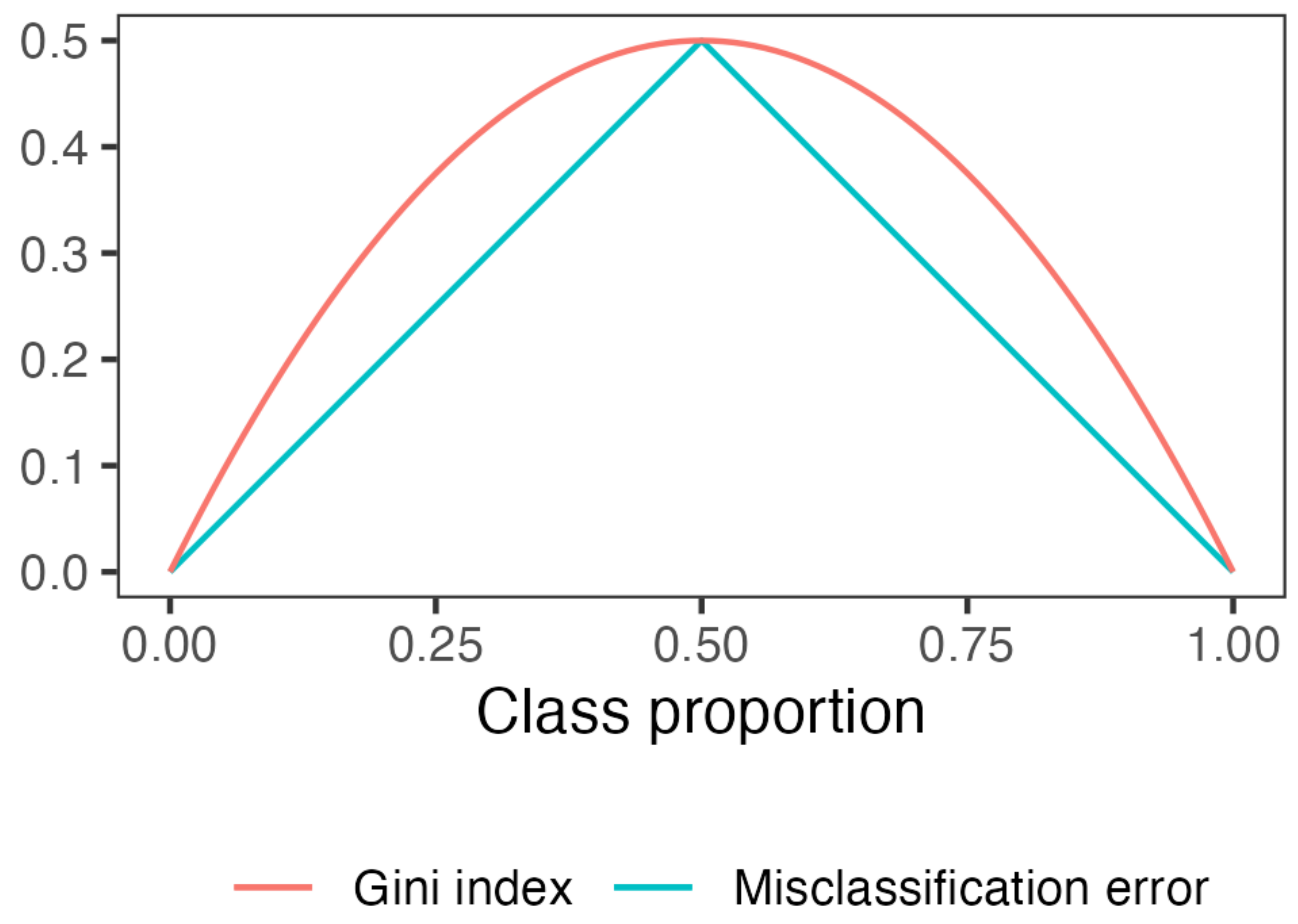
$$\text{Total Gini index} = \frac{1}{n}\left\{ n_1 \cdot 2\hat{p}_1(1 - \hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1 - \hat{p}_M) \right\}.$$

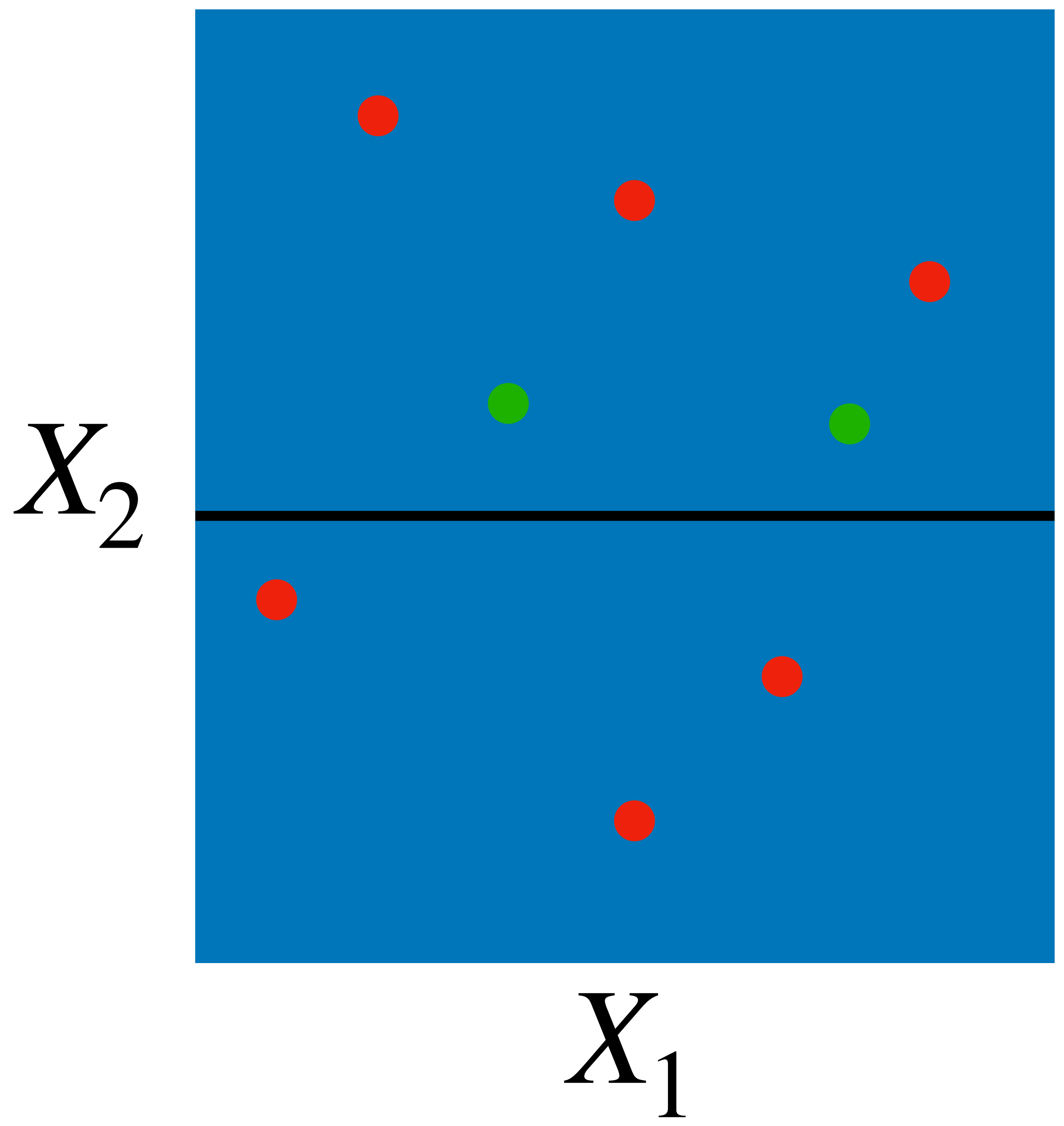
Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error

Training a classification tree

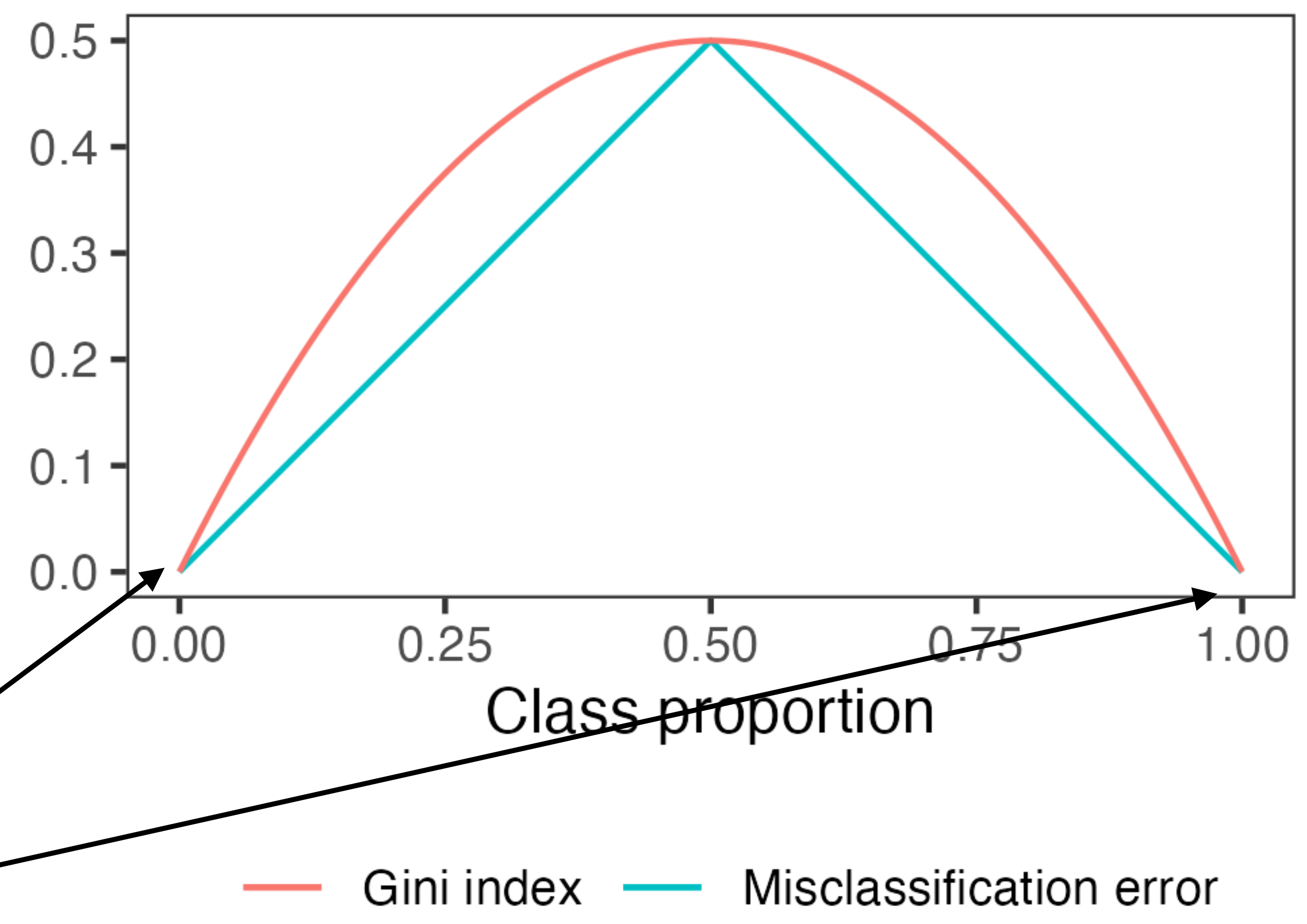
Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



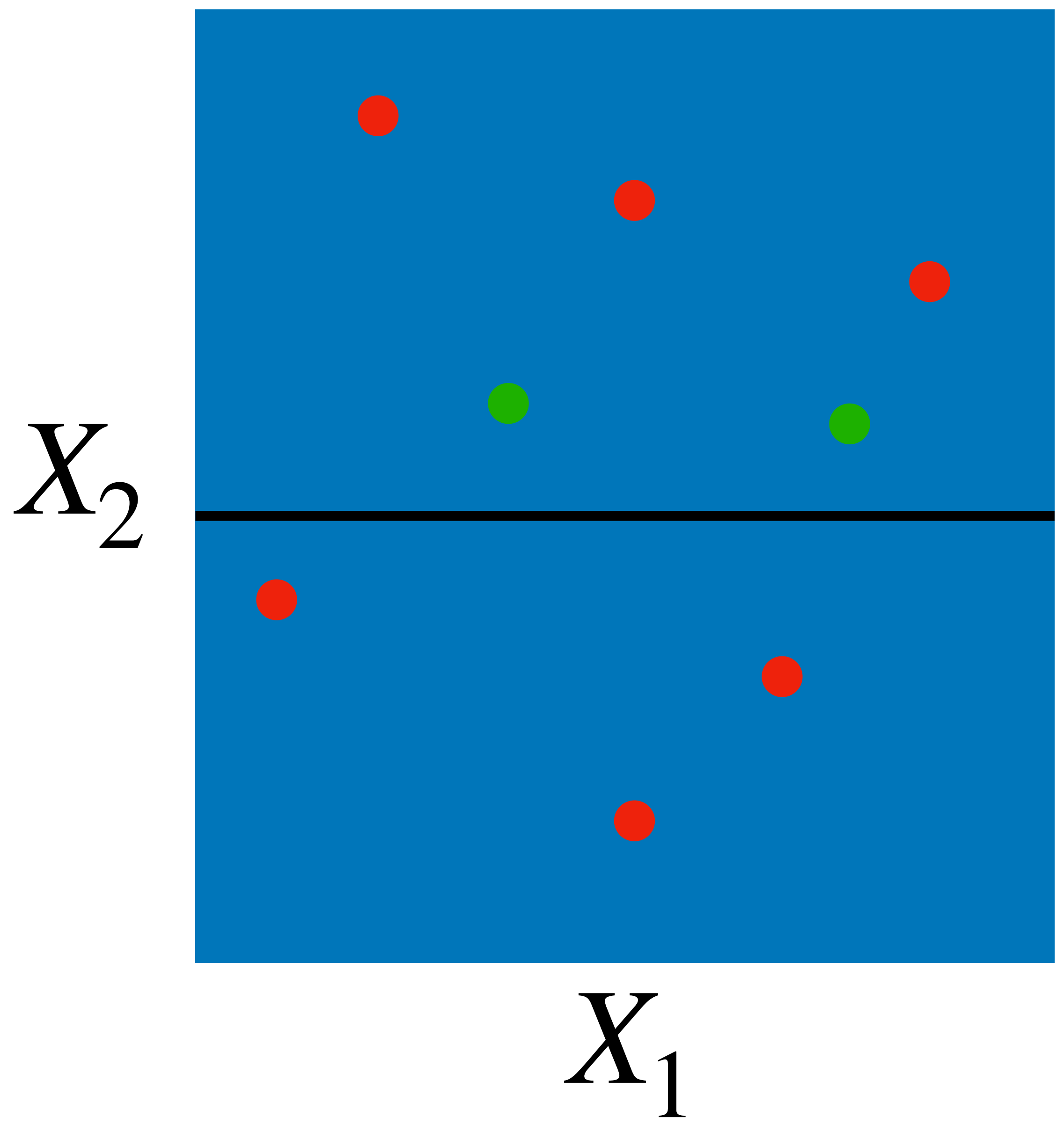


Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error

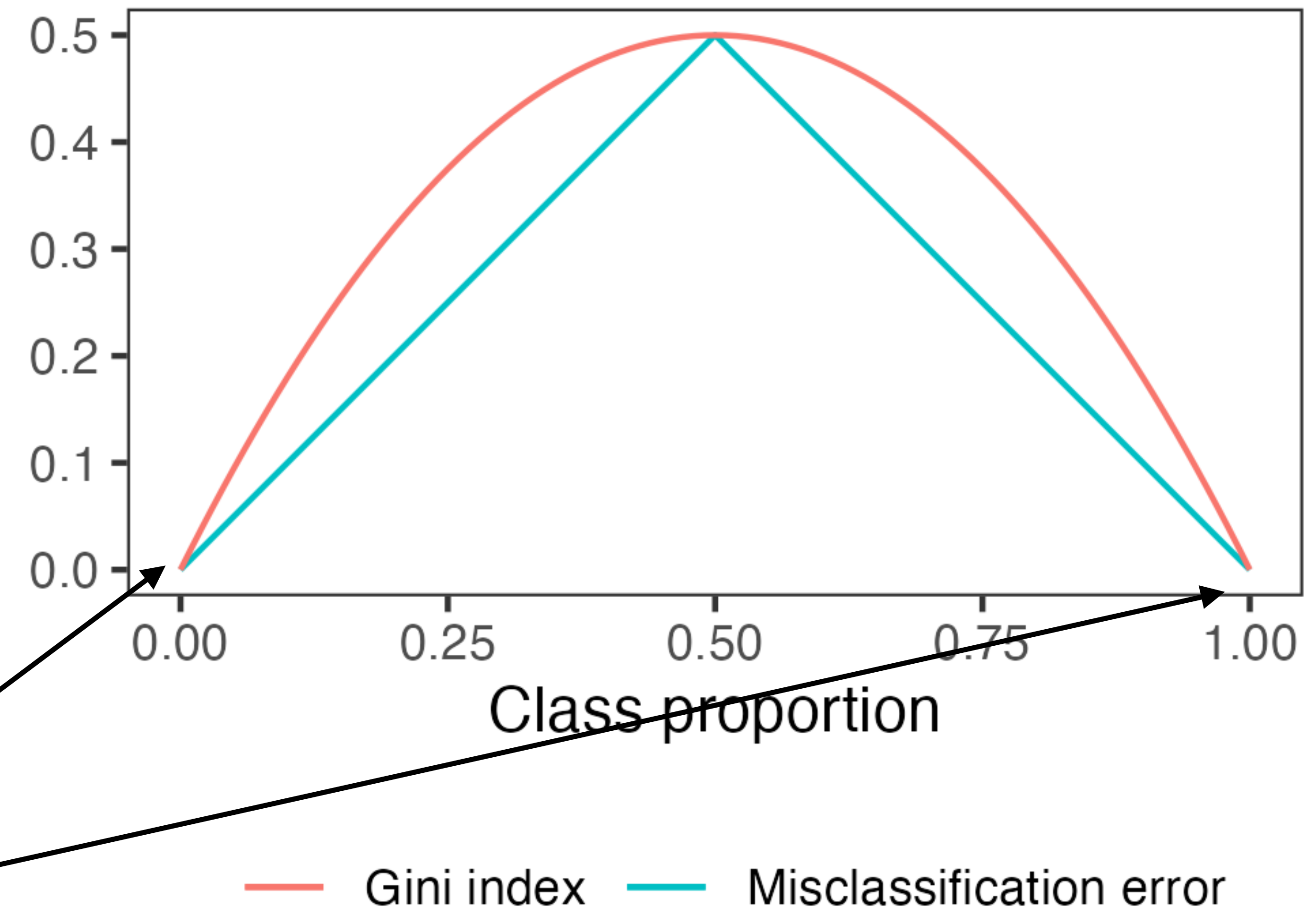


Rises more sharply away from 0 and 1, promoting node purity

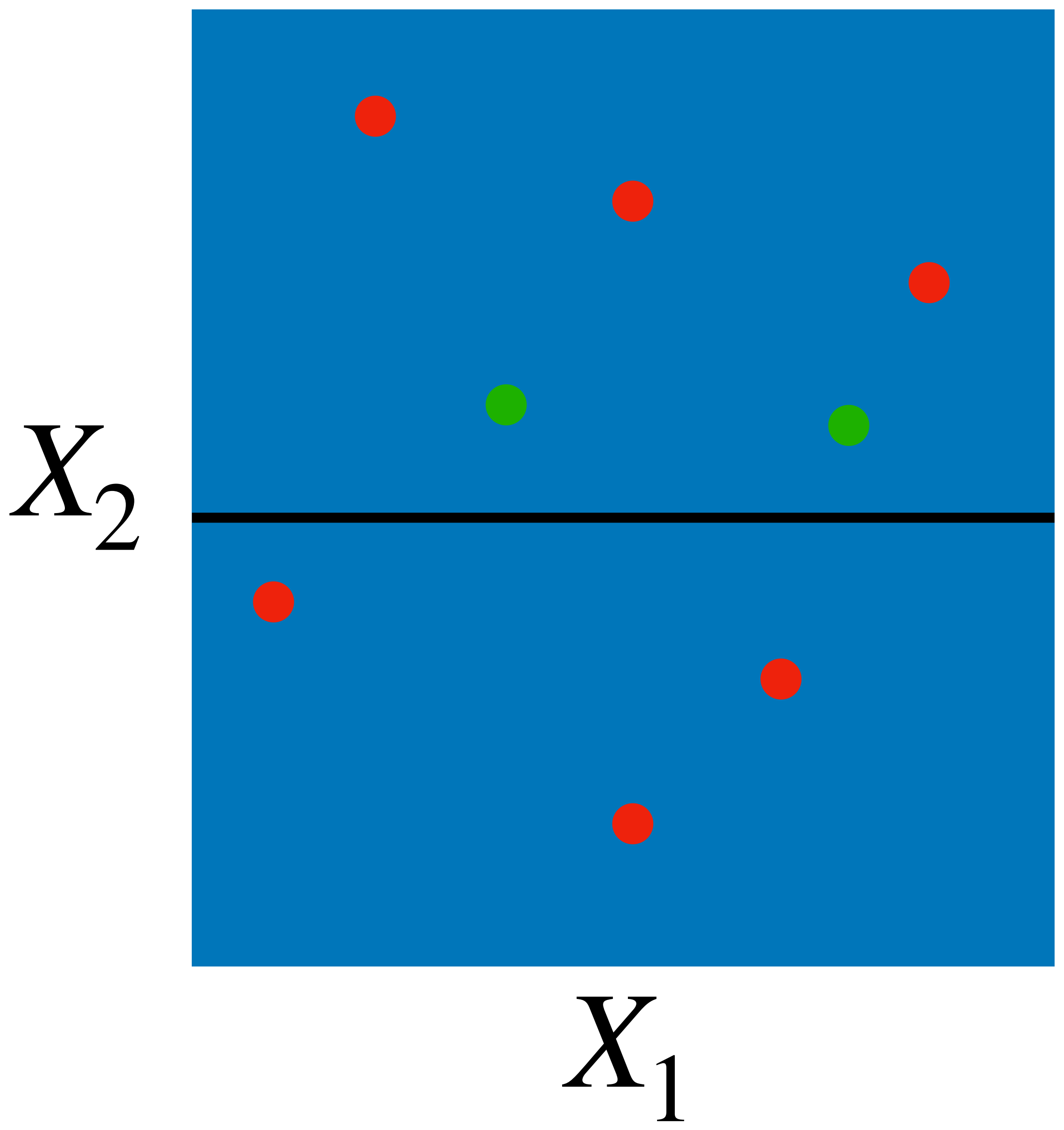


Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



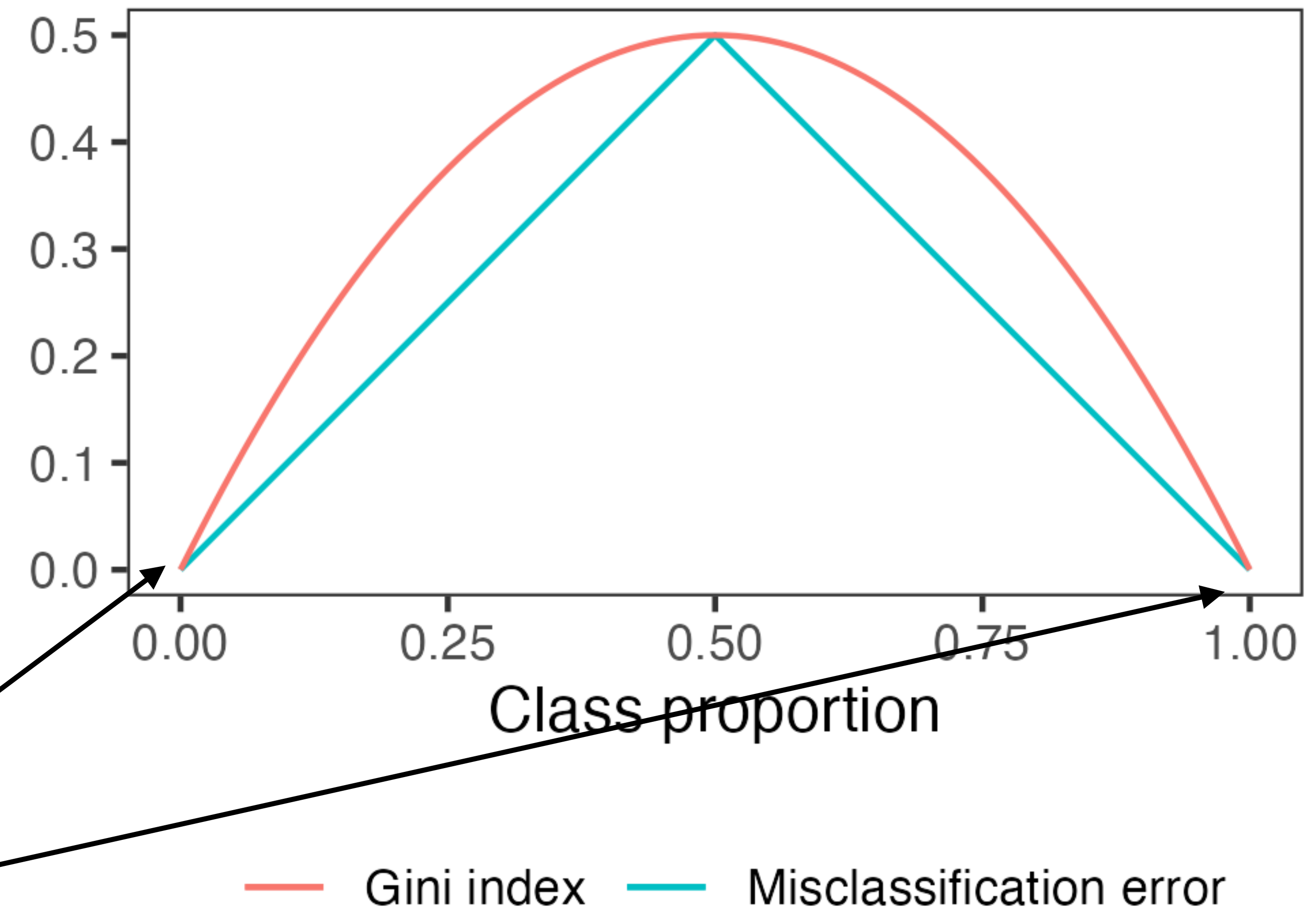
Rises more sharply away from 0 and 1, promoting node purity



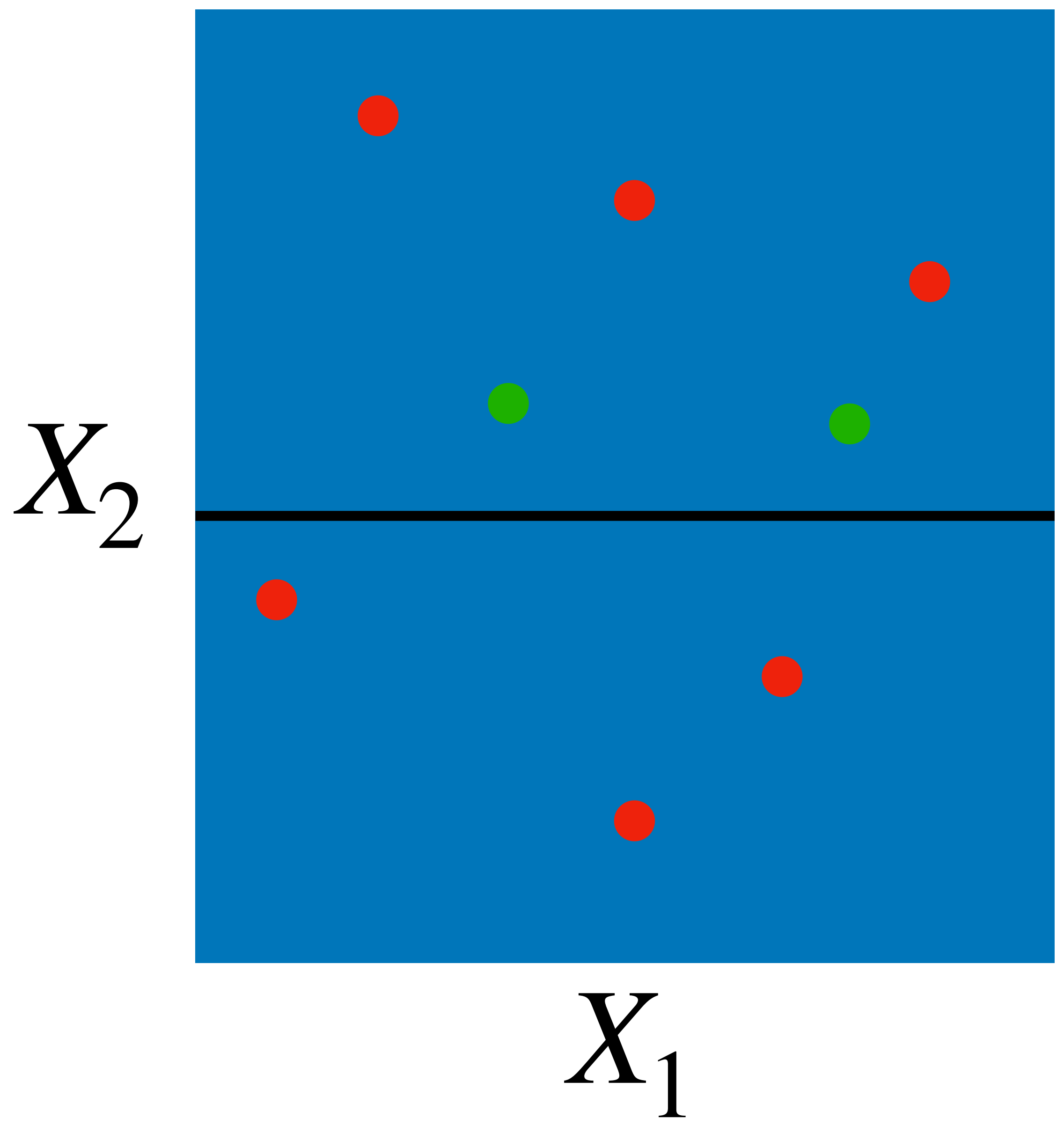
Tree	Total Misclass. error

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



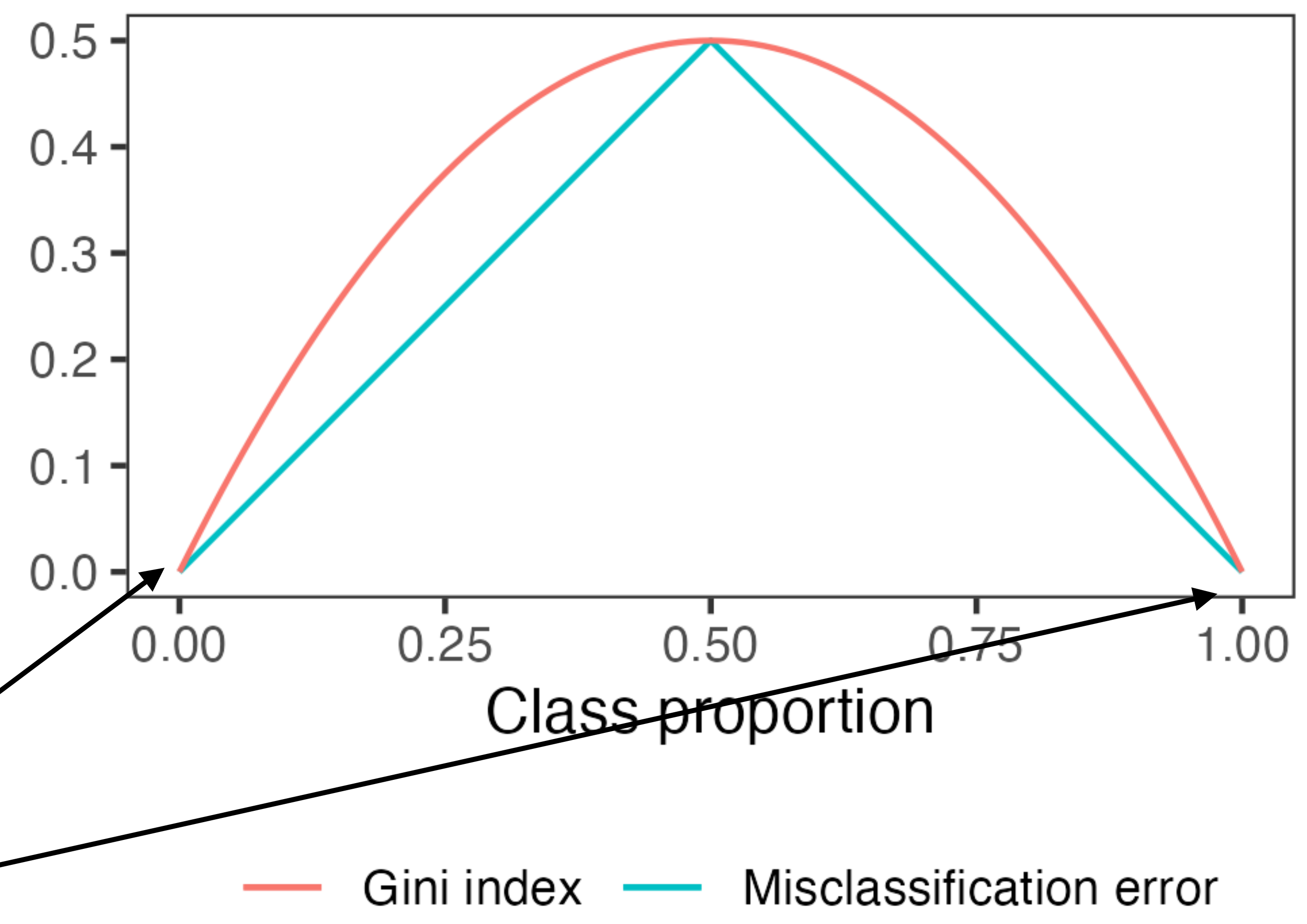
Rises more sharply away from 0 and 1, promoting node purity



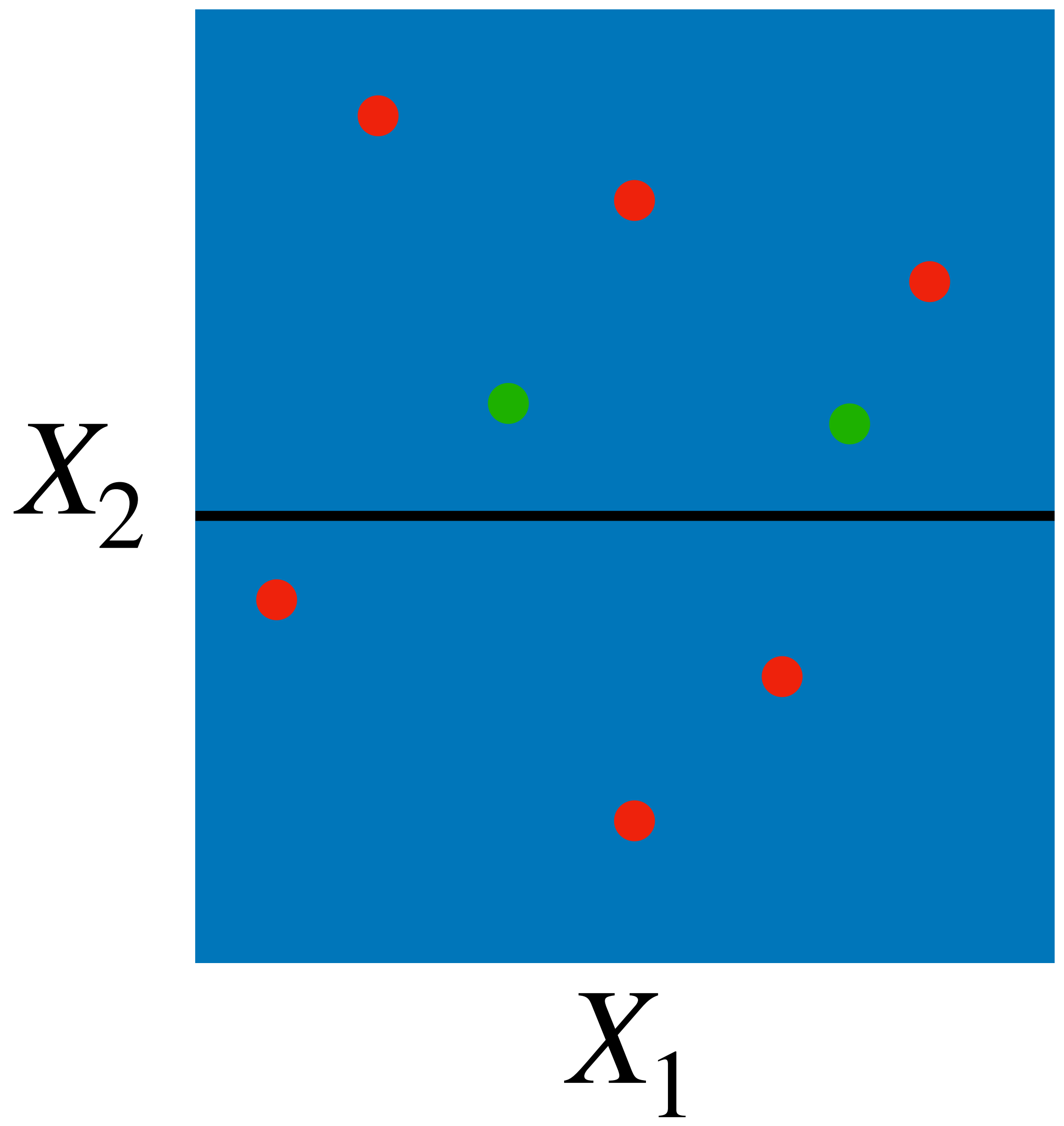
Tree	Total Misclass. error
No splits	

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



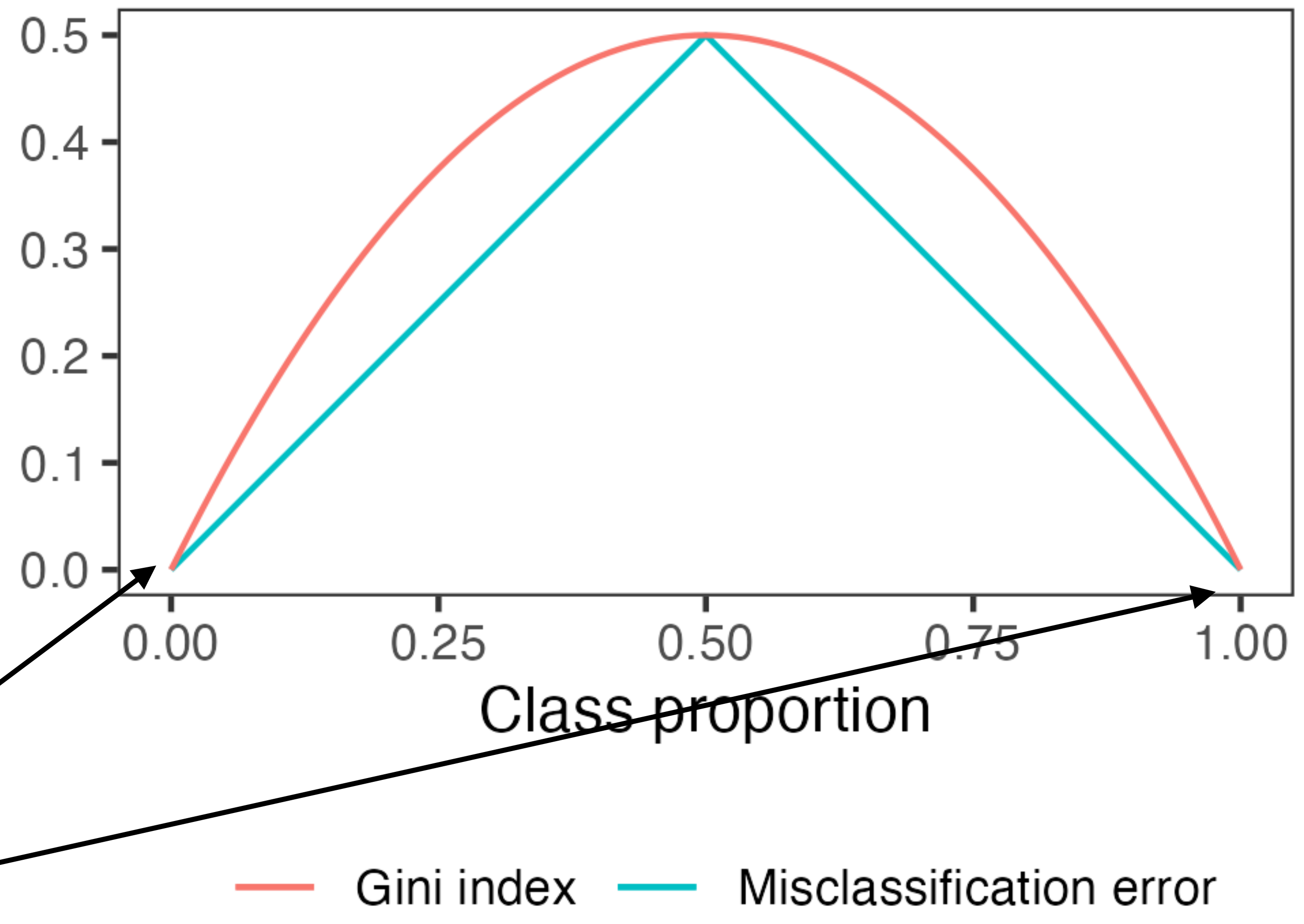
Rises more sharply away from 0 and 1, promoting node purity



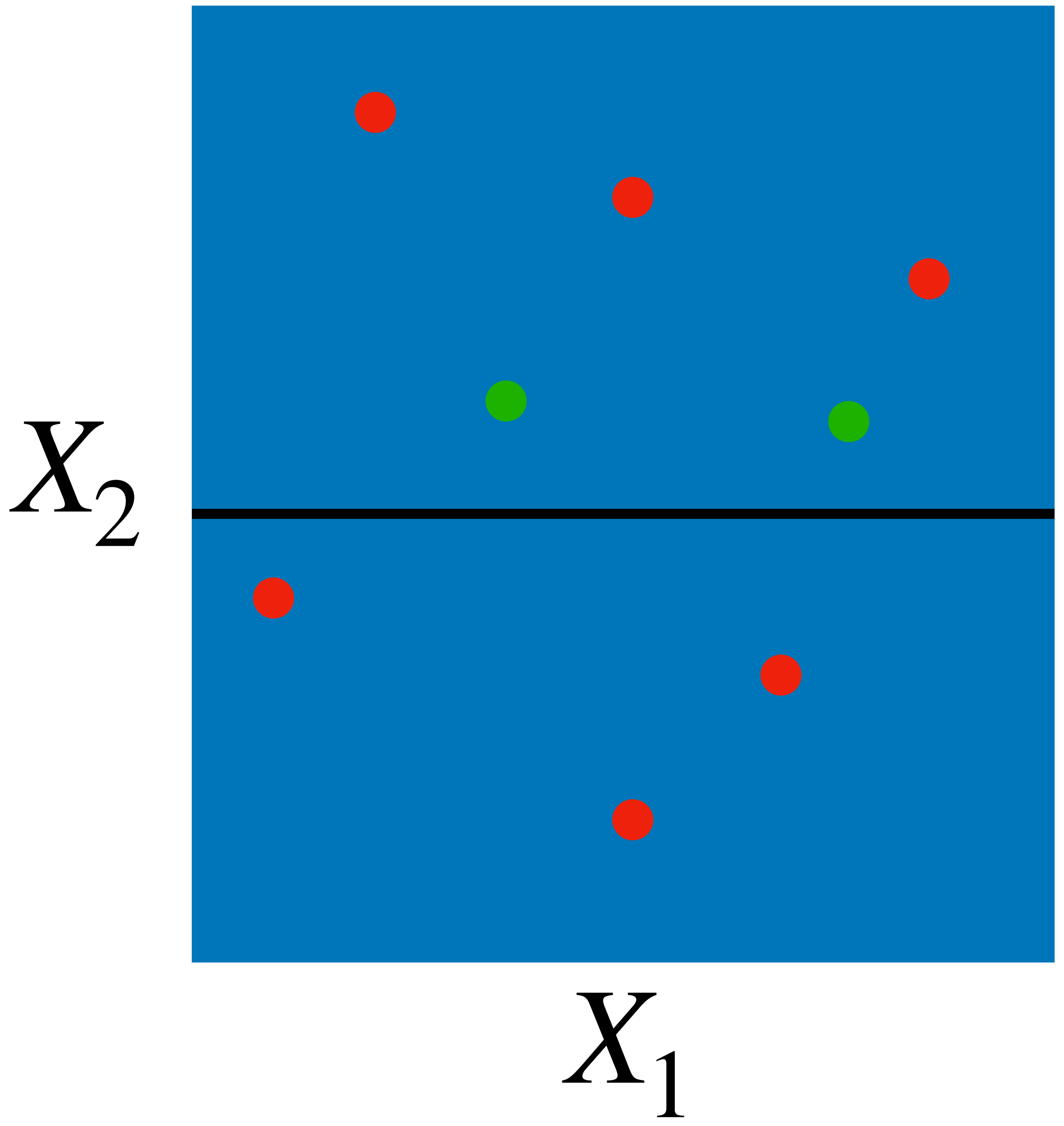
Tree	Total Misclass. error
No splits	$\frac{1}{4}$

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



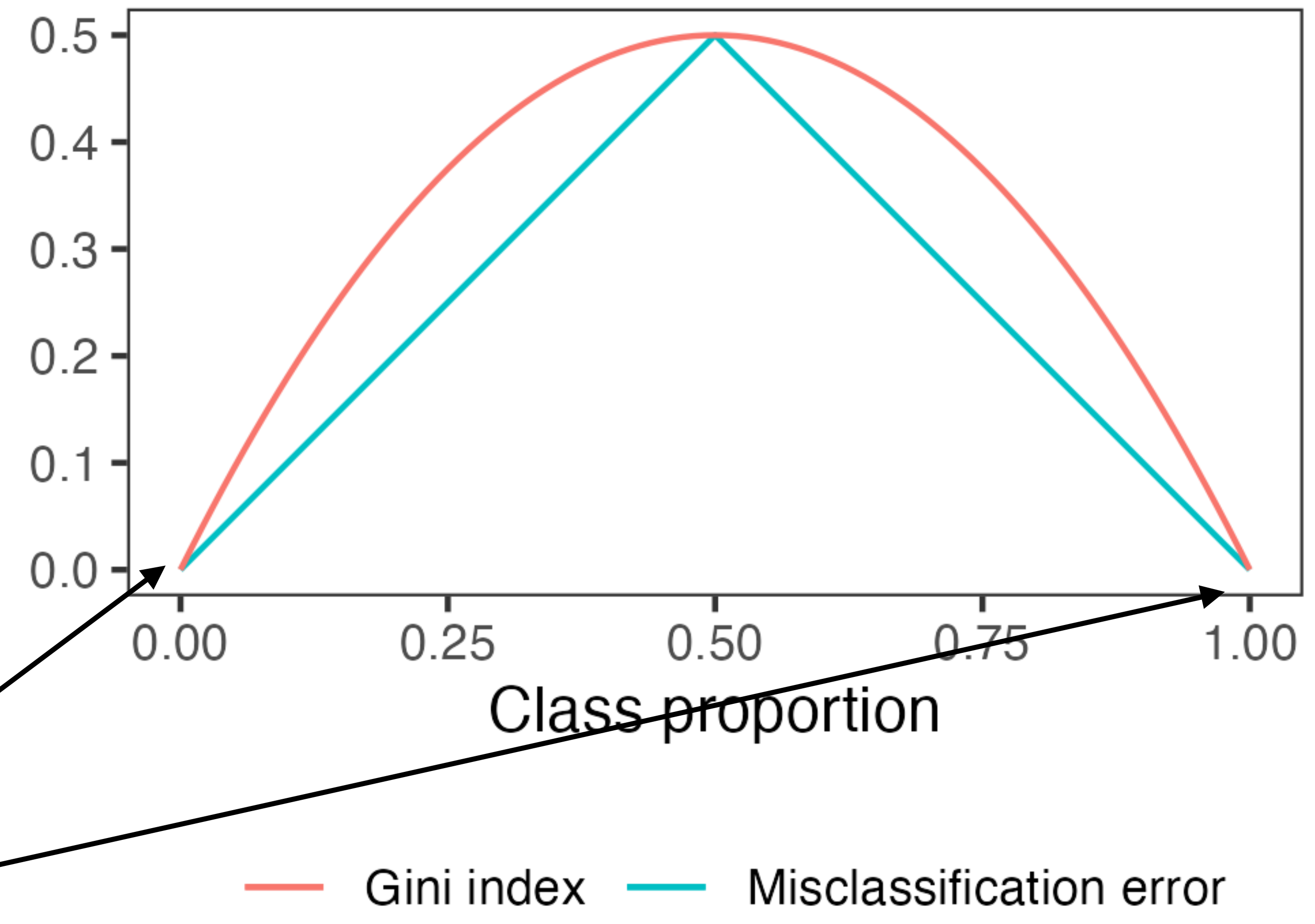
Rises more sharply away from 0 and 1, promoting node purity



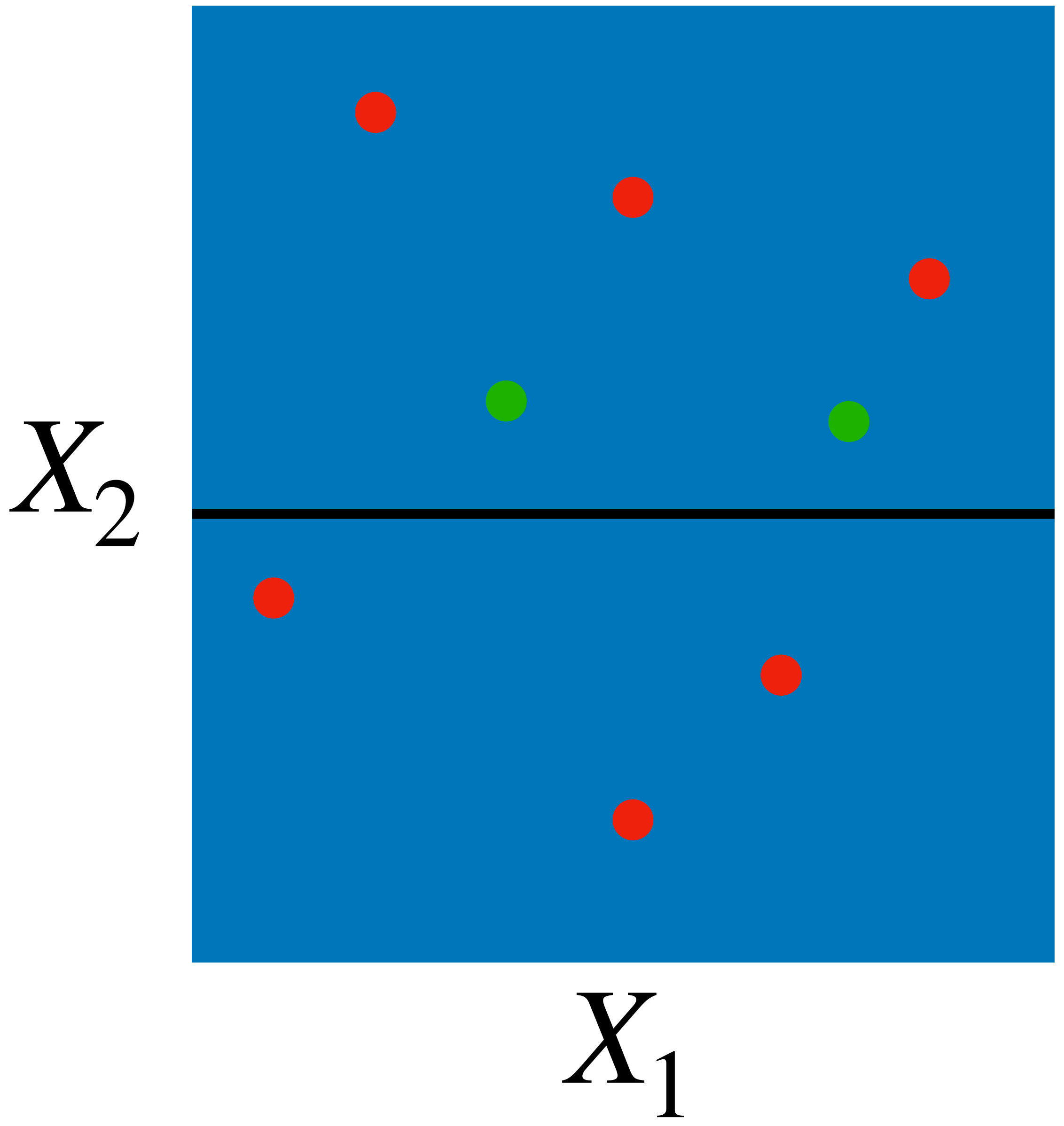
Tree	Total Misclass. error
No splits	$\frac{1}{4}$
One split	

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



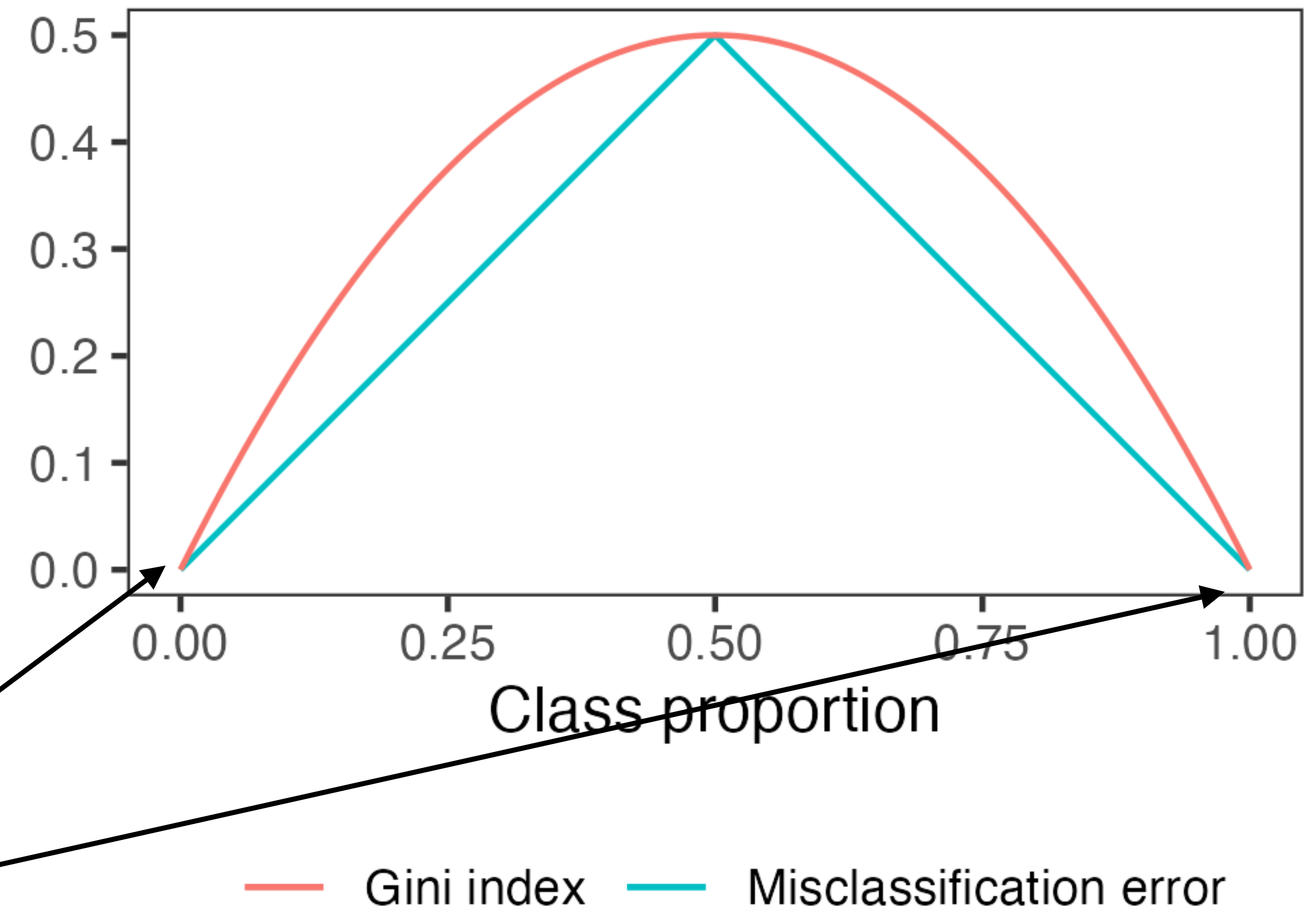
Rises more sharply away from 0 and 1, promoting node purity



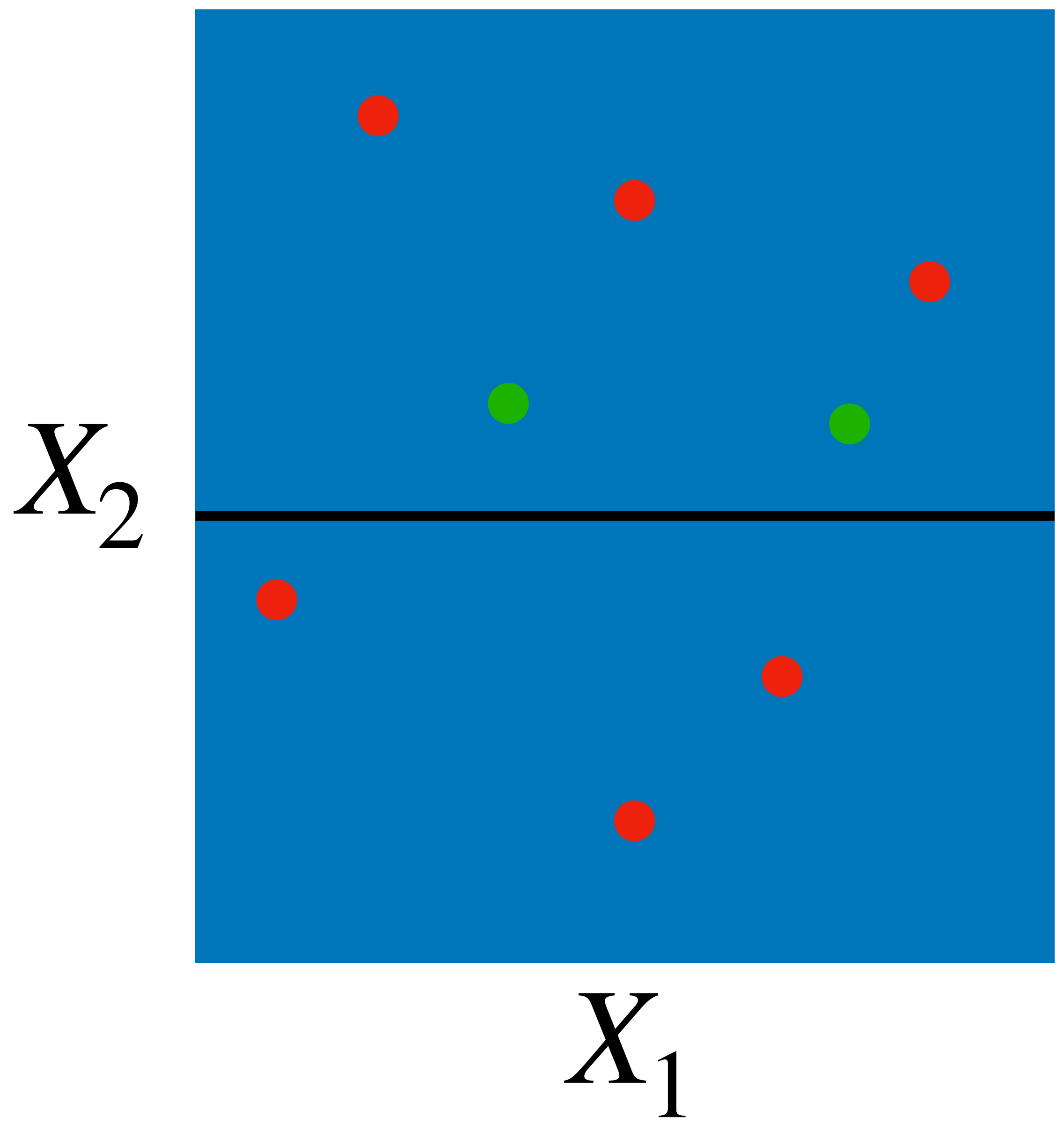
Tree	Total Misclass. error
No splits	$\frac{1}{4}$
One split	$\frac{1}{4}$

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



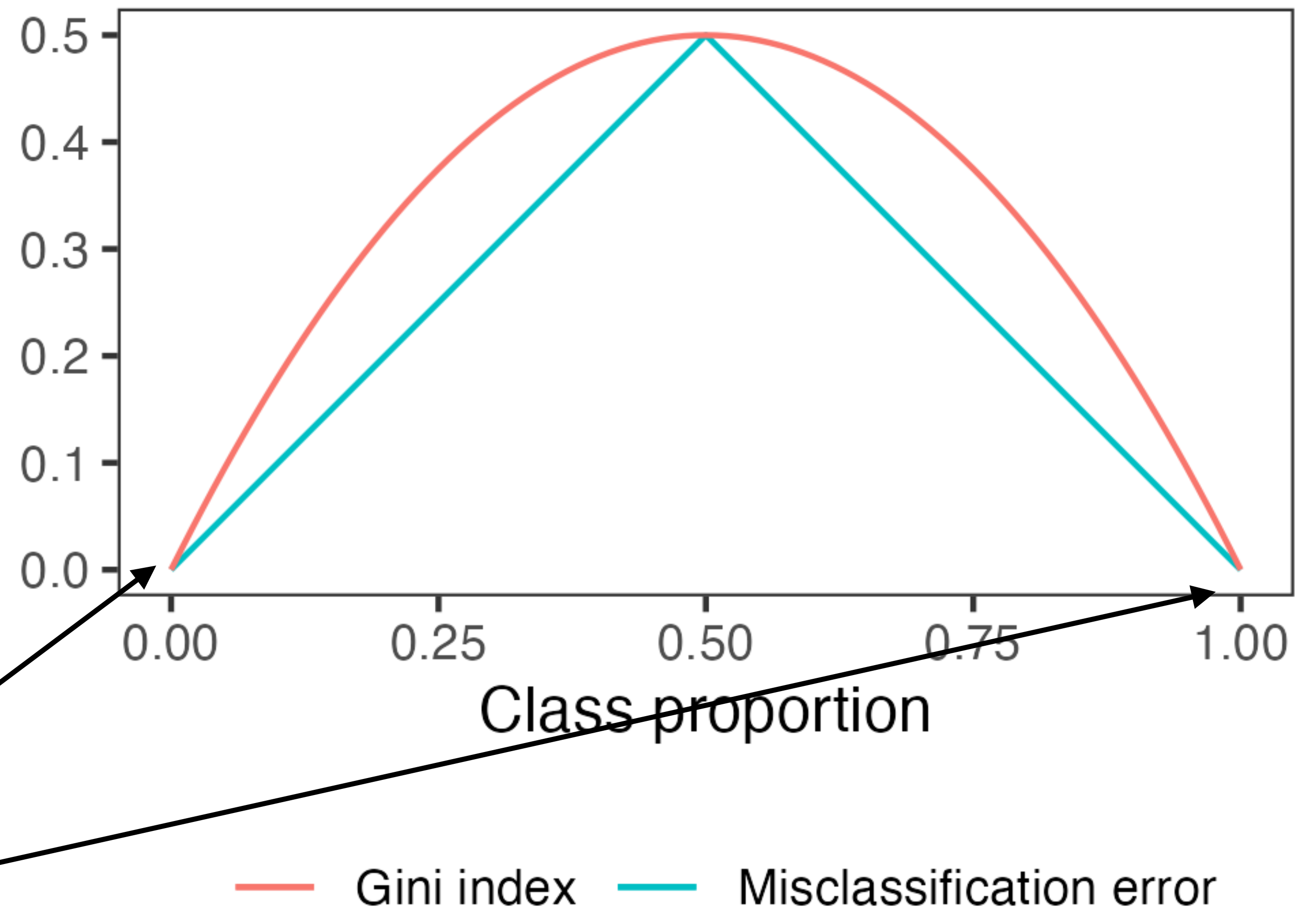
Rises more sharply away from 0 and 1, promoting node purity



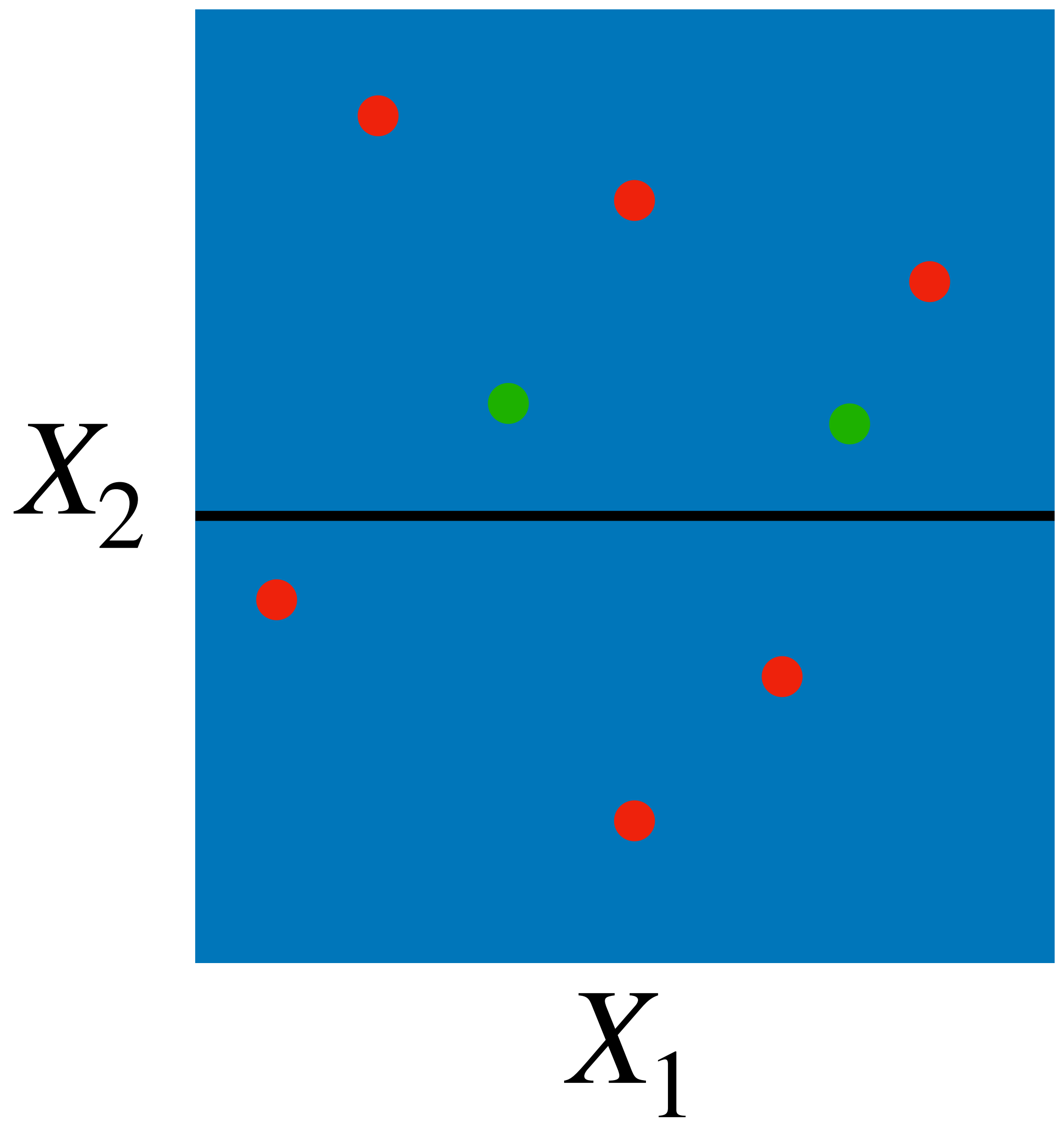
Tree	Total Misclass. error	Total Gini index
No splits	$\frac{1}{4}$	
One split	$\frac{1}{4}$	

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



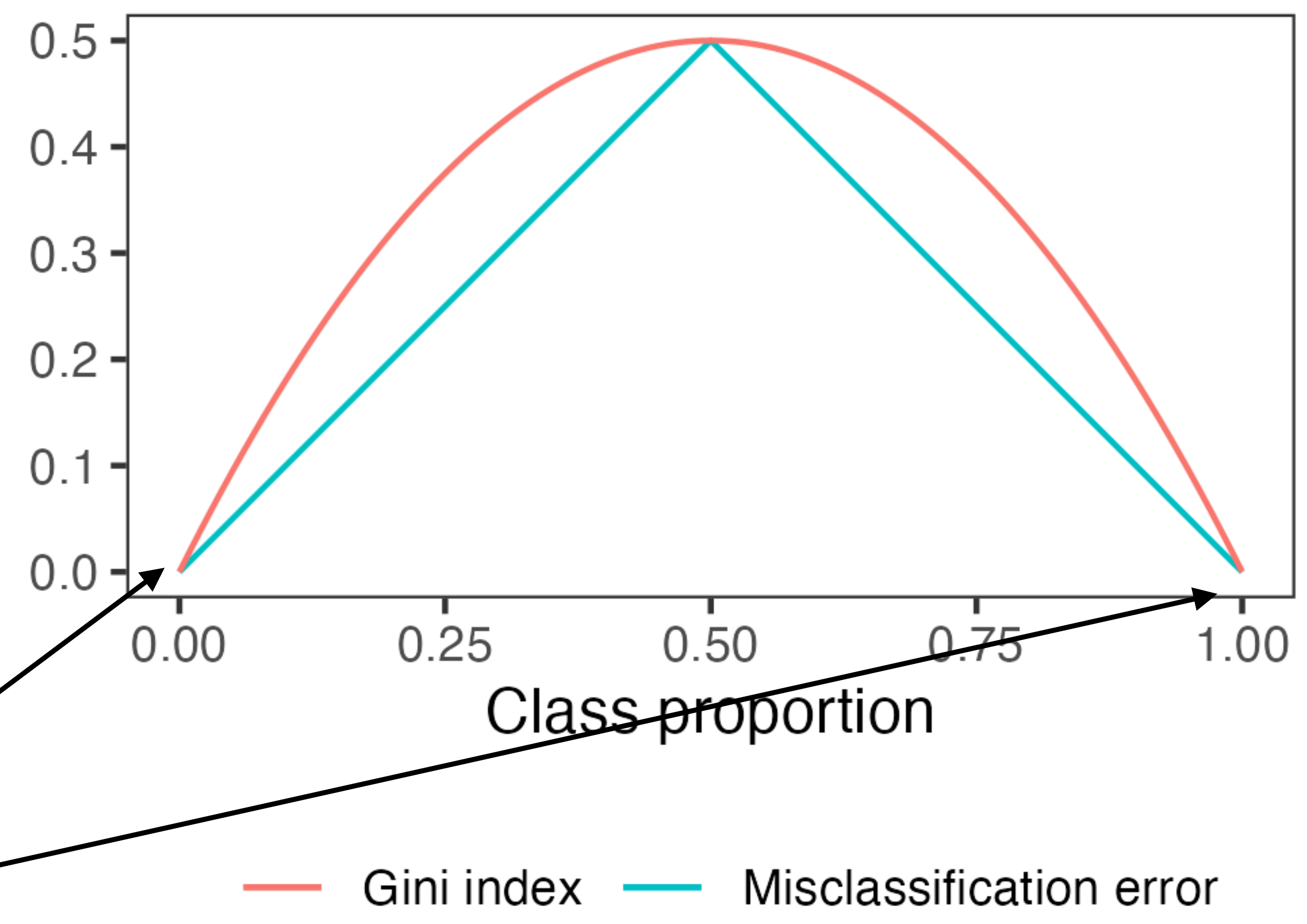
Rises more sharply away from 0 and 1, promoting node purity



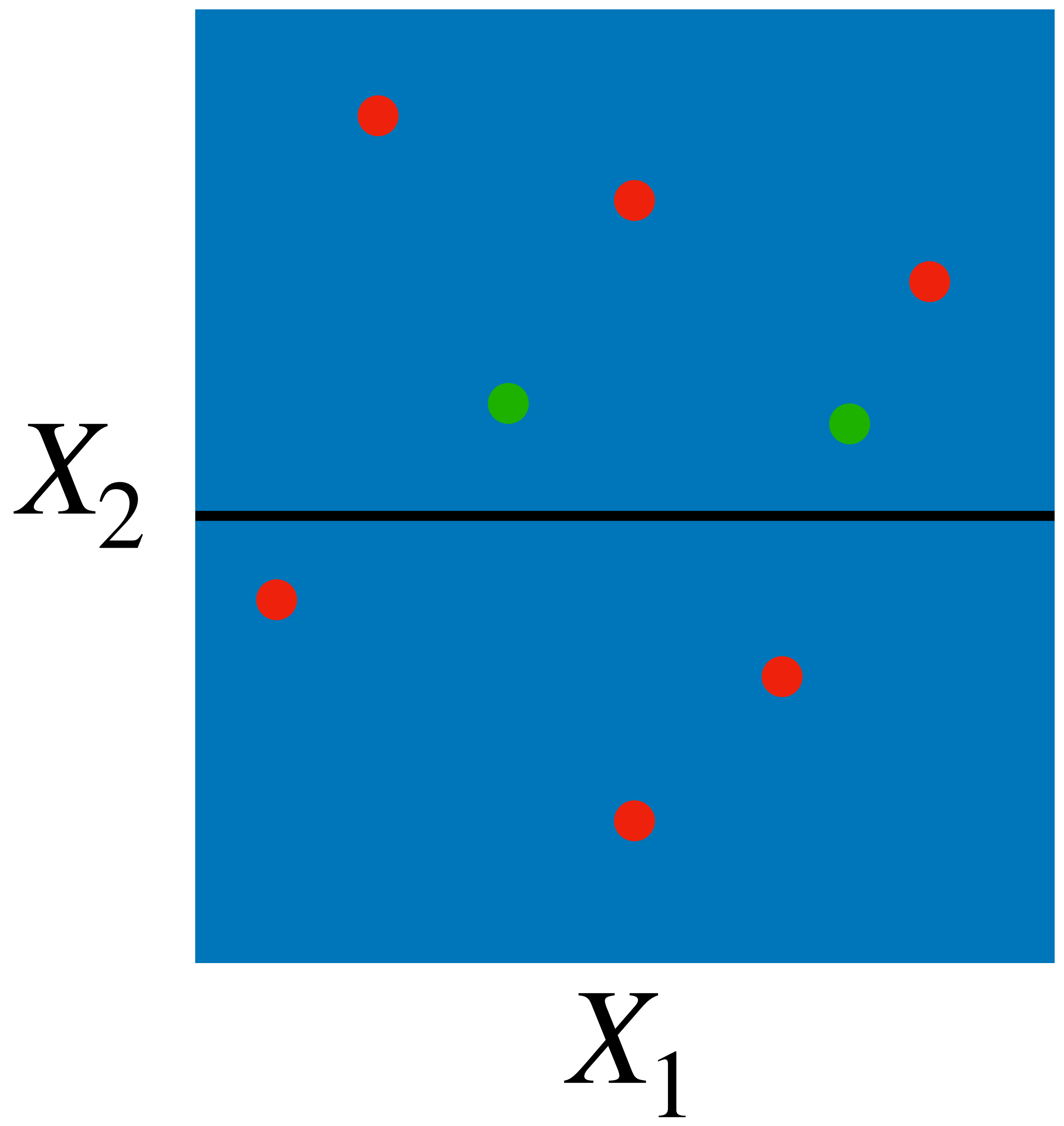
Tree	Total Misclass. error	Total Gini index
No splits	$\frac{1}{4}$	$\frac{1}{8}\left(8 \cdot 2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) = \frac{3}{8}$
One split	$\frac{1}{4}$	

Training a classification tree

Finding the rectangles $\widehat{R}_m$: Gini index versus misclassification error



Rises more sharply away from 0 and 1, promoting node purity



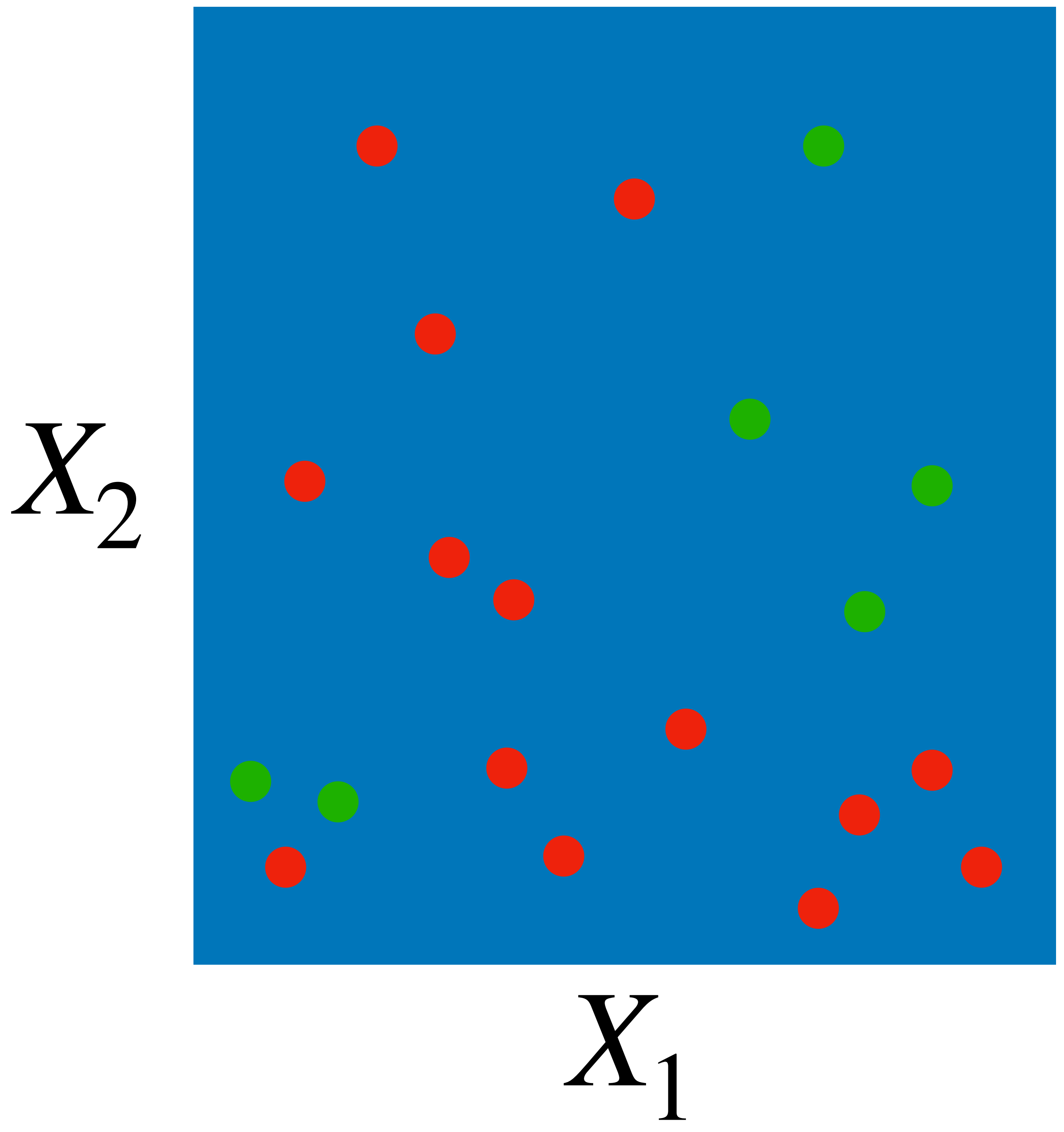
Tree	Total Misclass. error	Total Gini index
No splits	$\frac{1}{4}$	$\frac{1}{8}\left(8 \cdot 2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) = \frac{3}{8}$
One split	$\frac{1}{4}$	$\frac{1}{8}\left(5 \cdot 2 \cdot \frac{2}{5} \cdot \frac{3}{5} + 3 \cdot 0\right) = \frac{3}{10}$

Training a classification tree

Finding the rectangles $\widehat{R}_m$

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



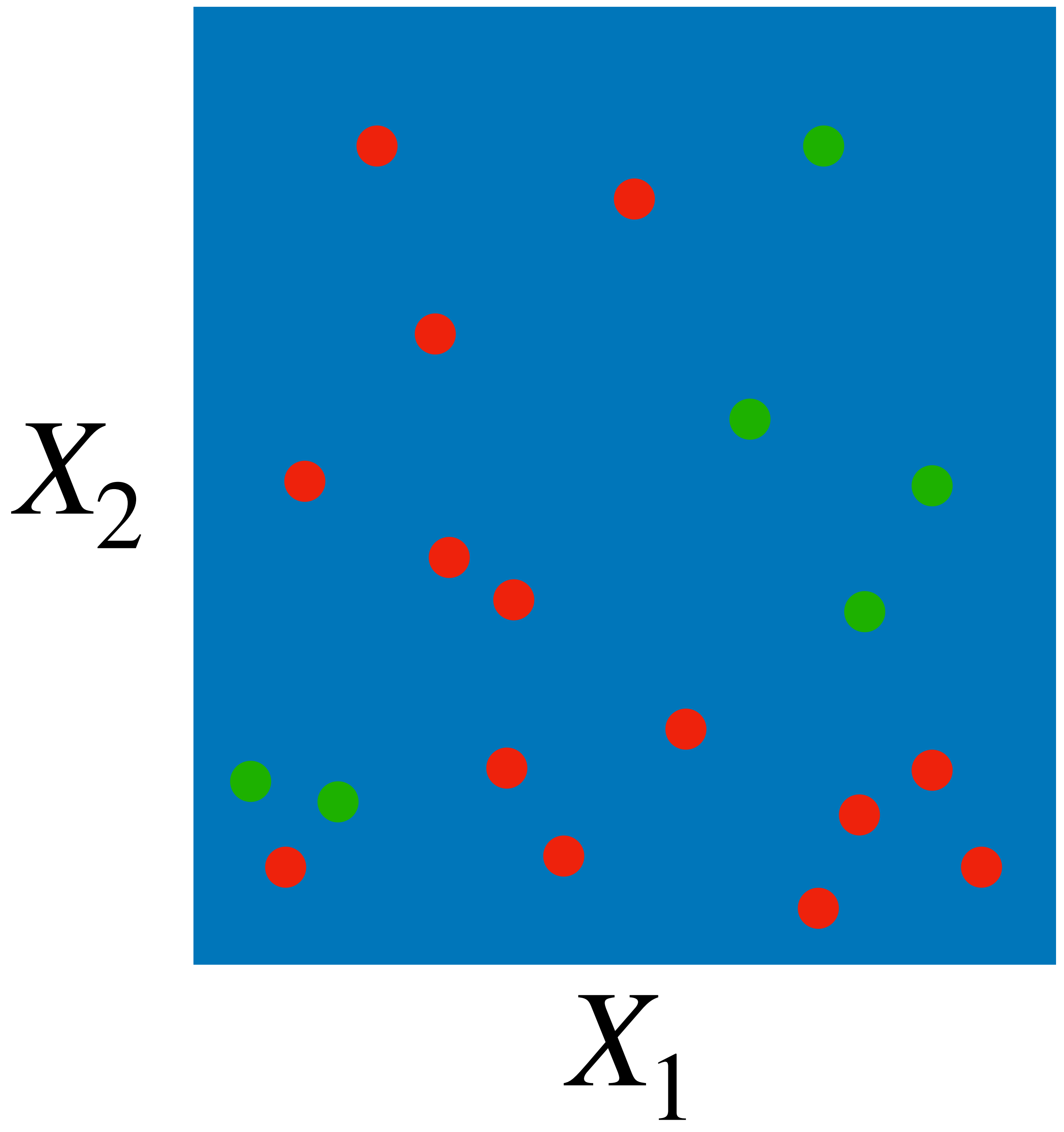
Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

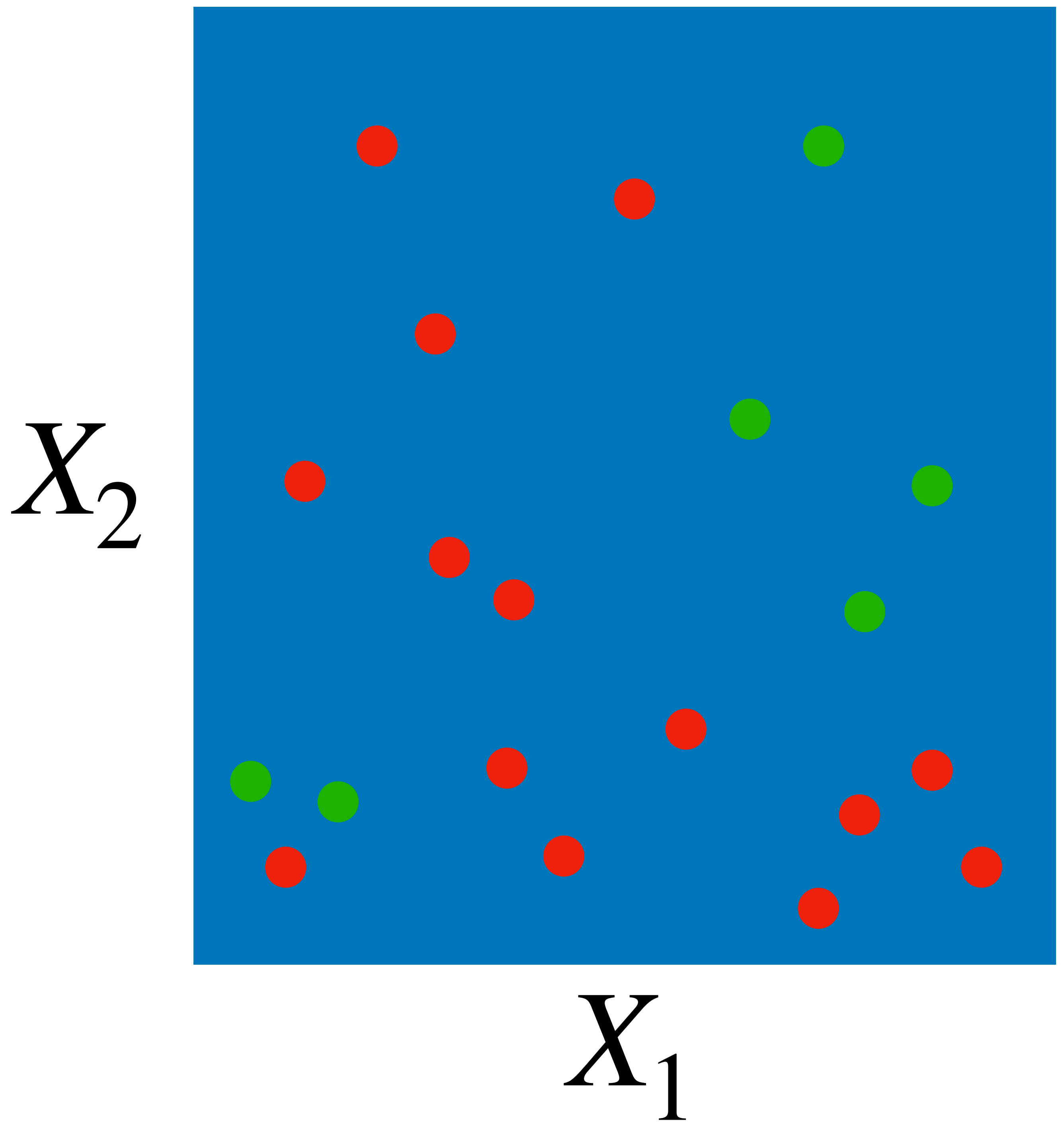
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

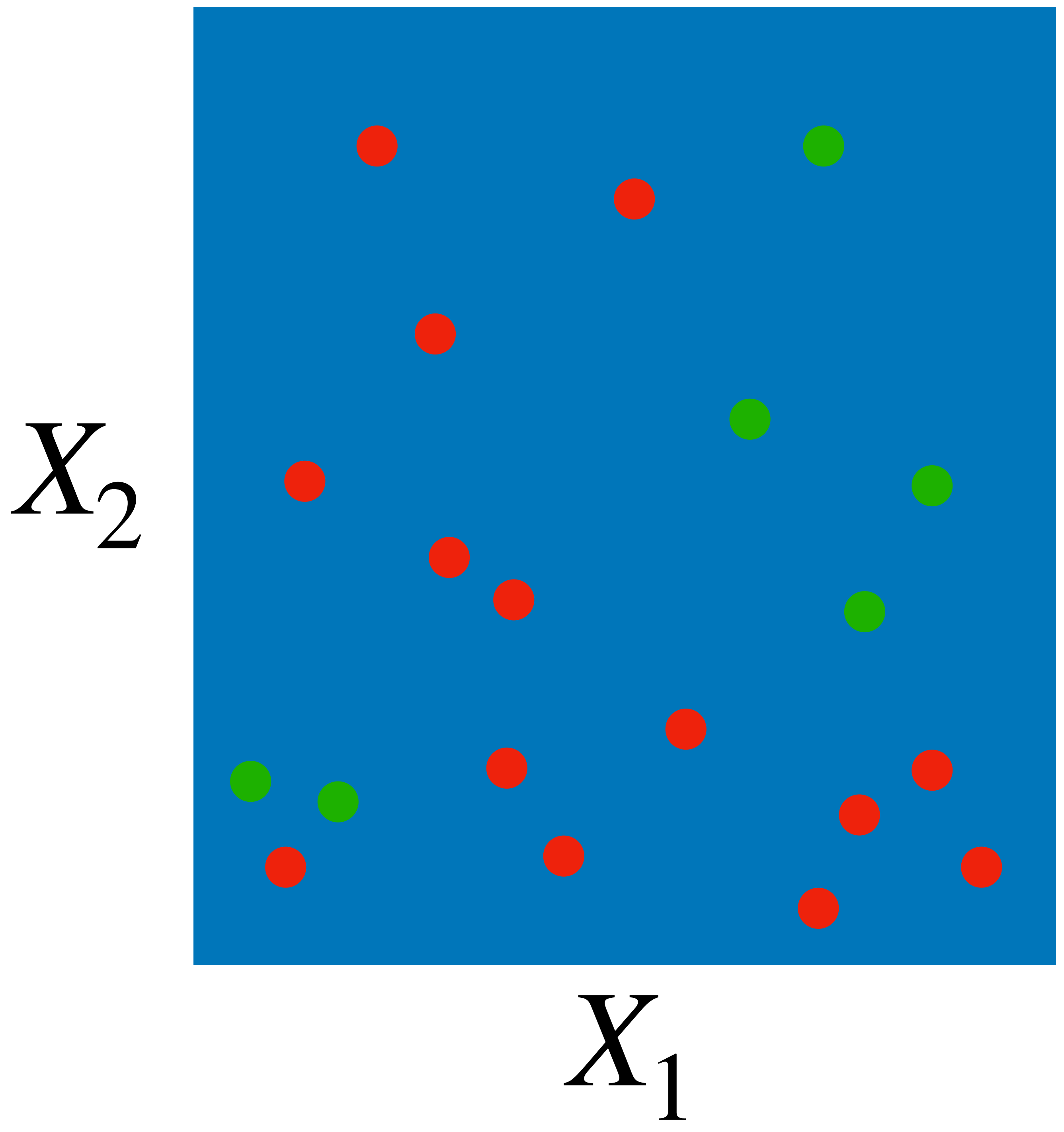
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

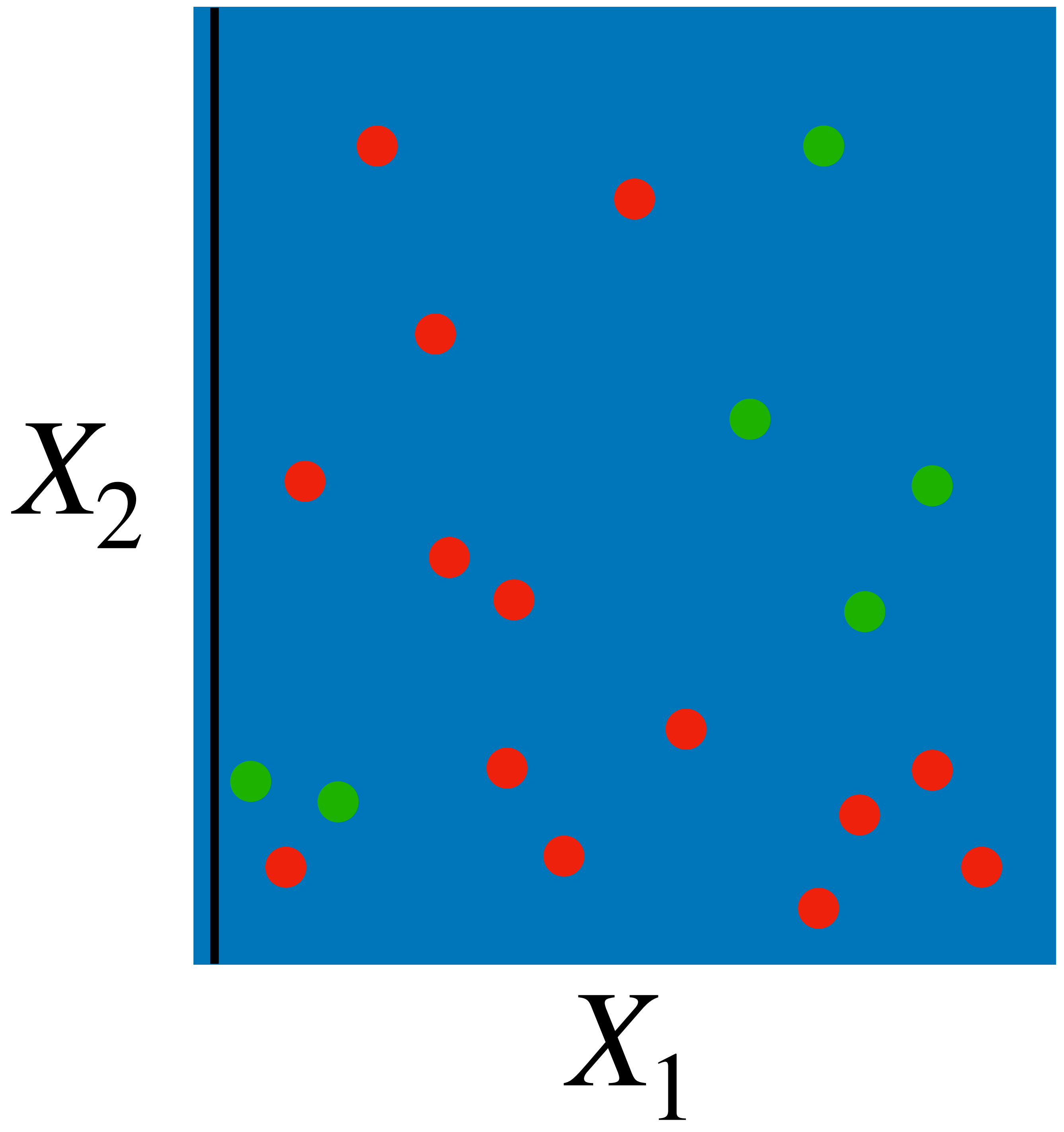
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

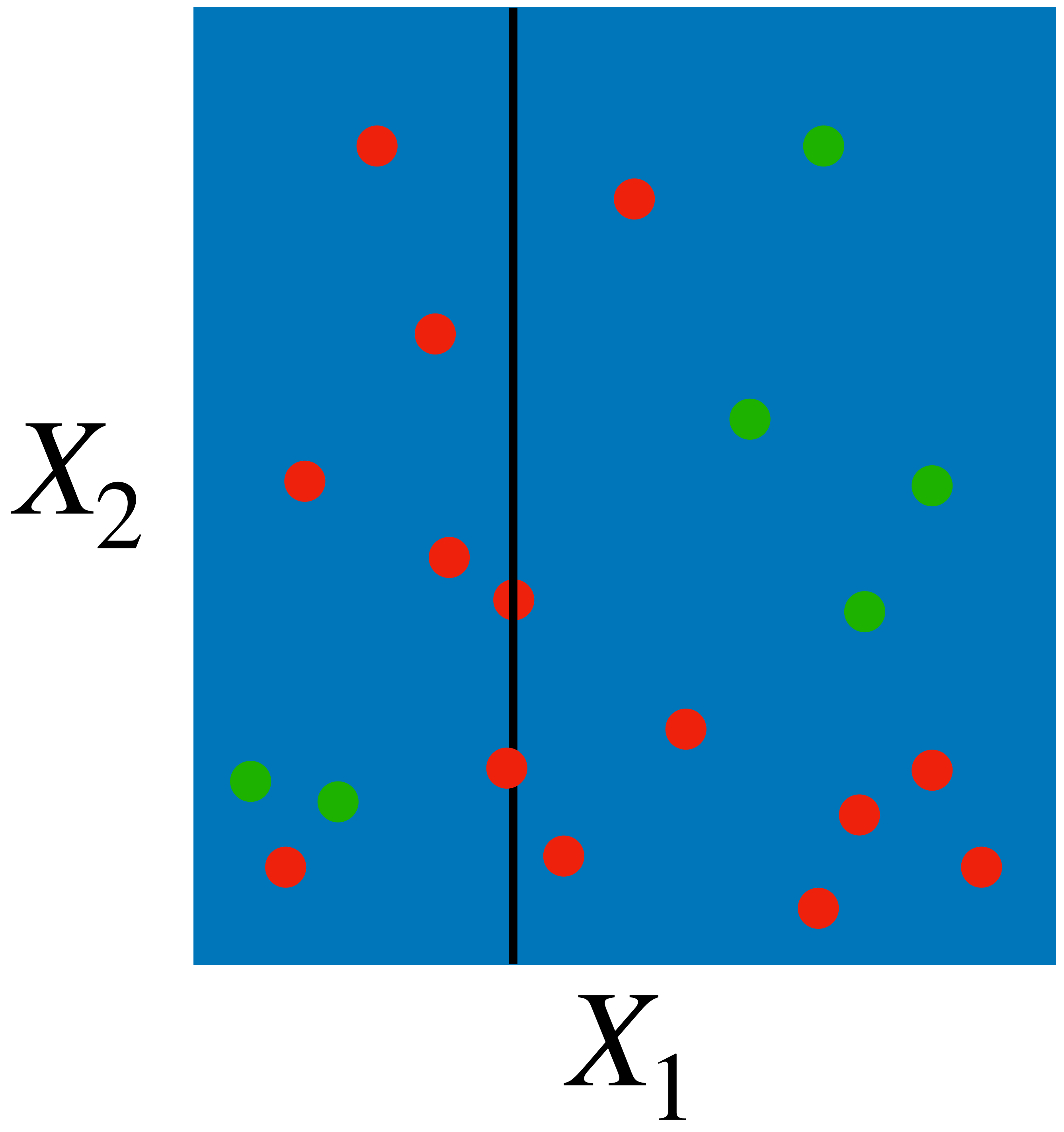
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

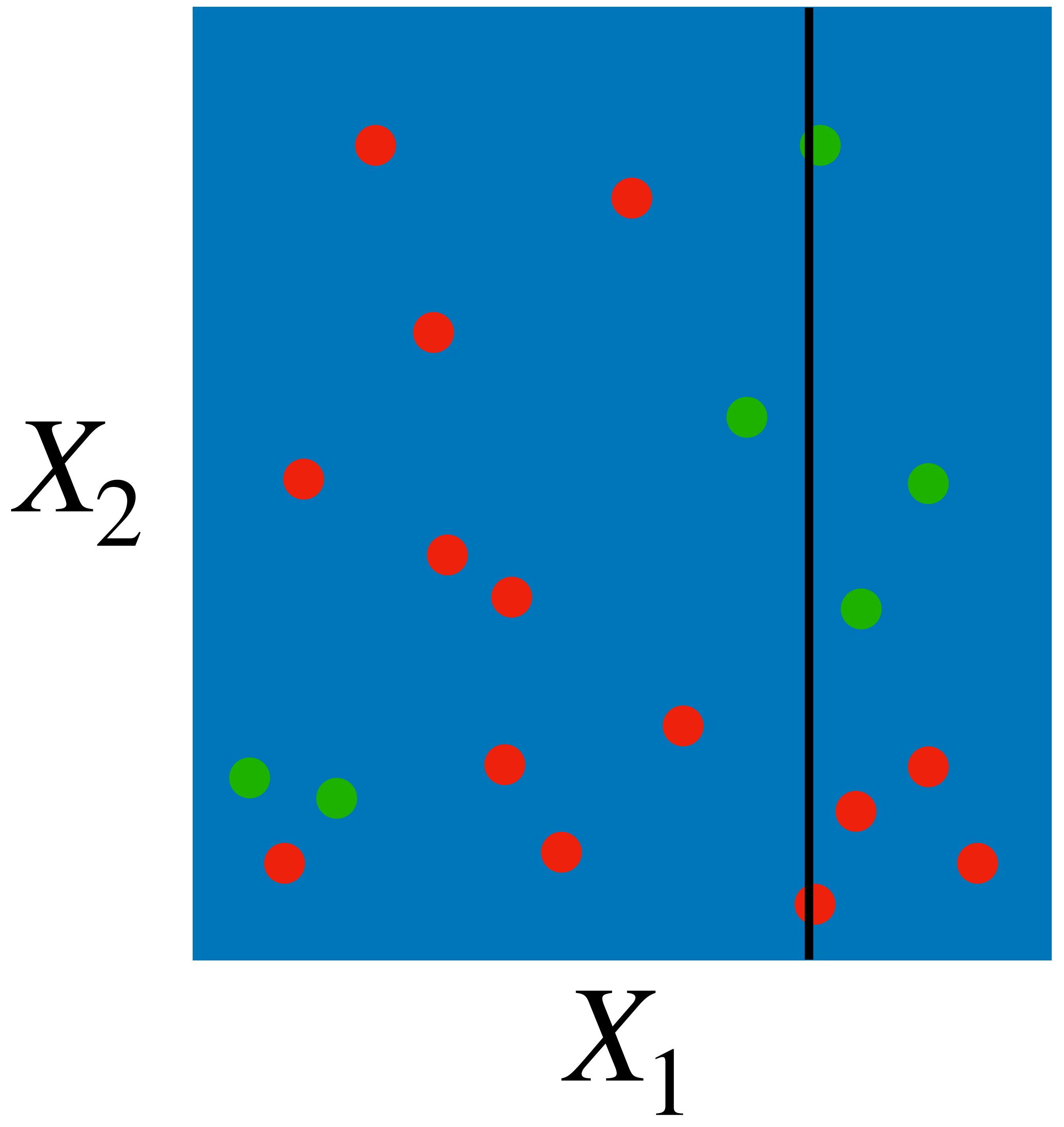
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

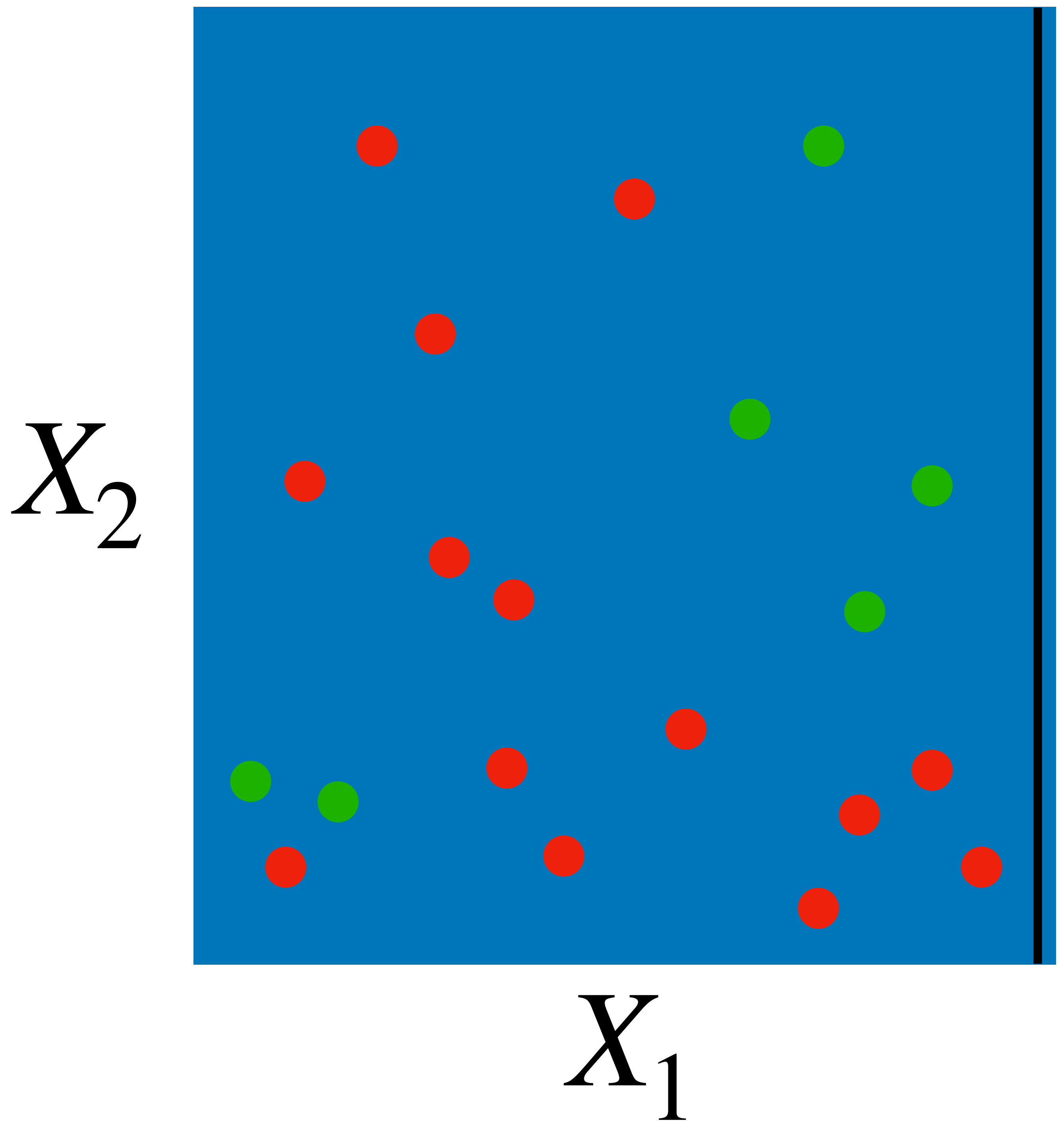
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

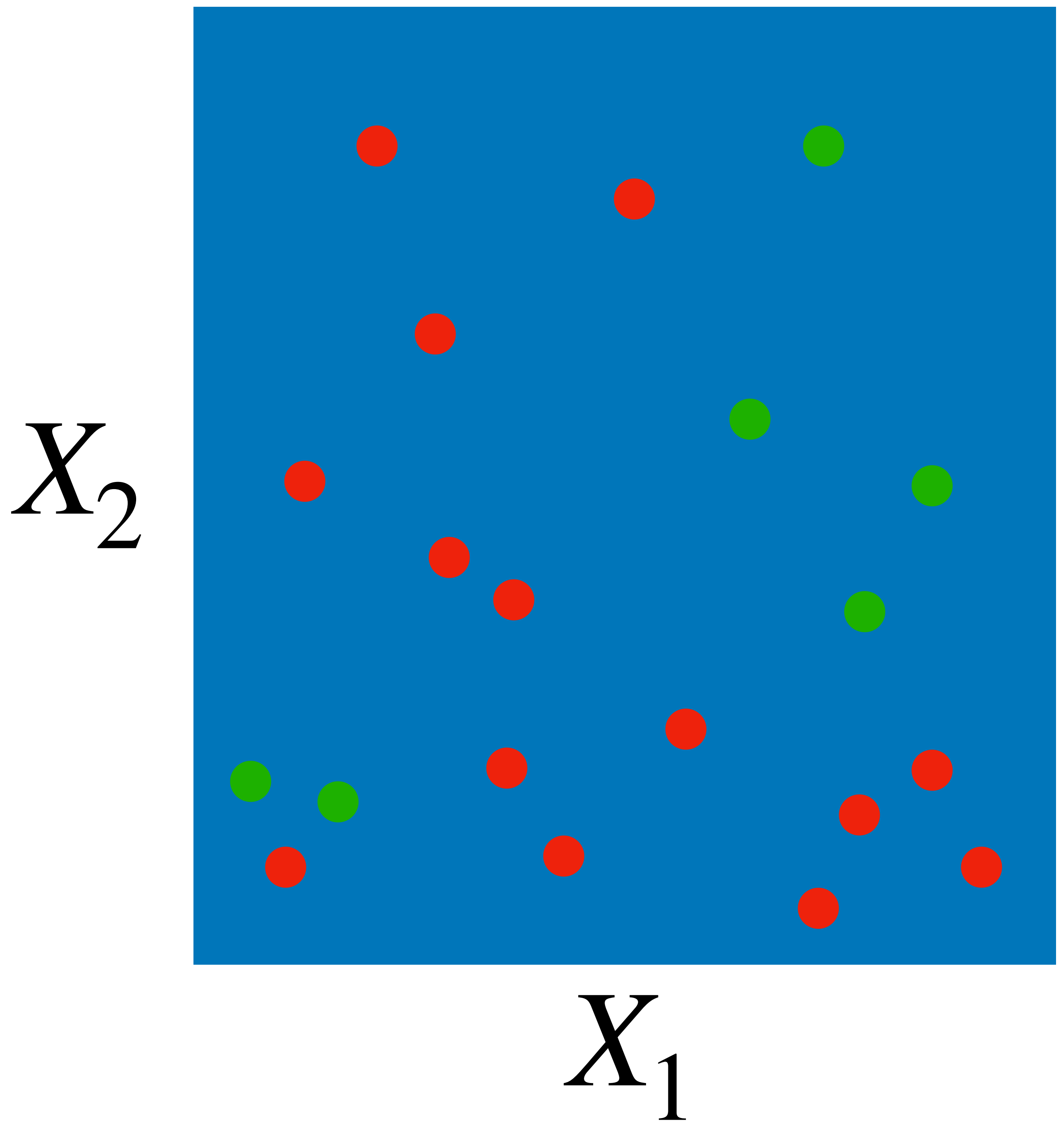
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

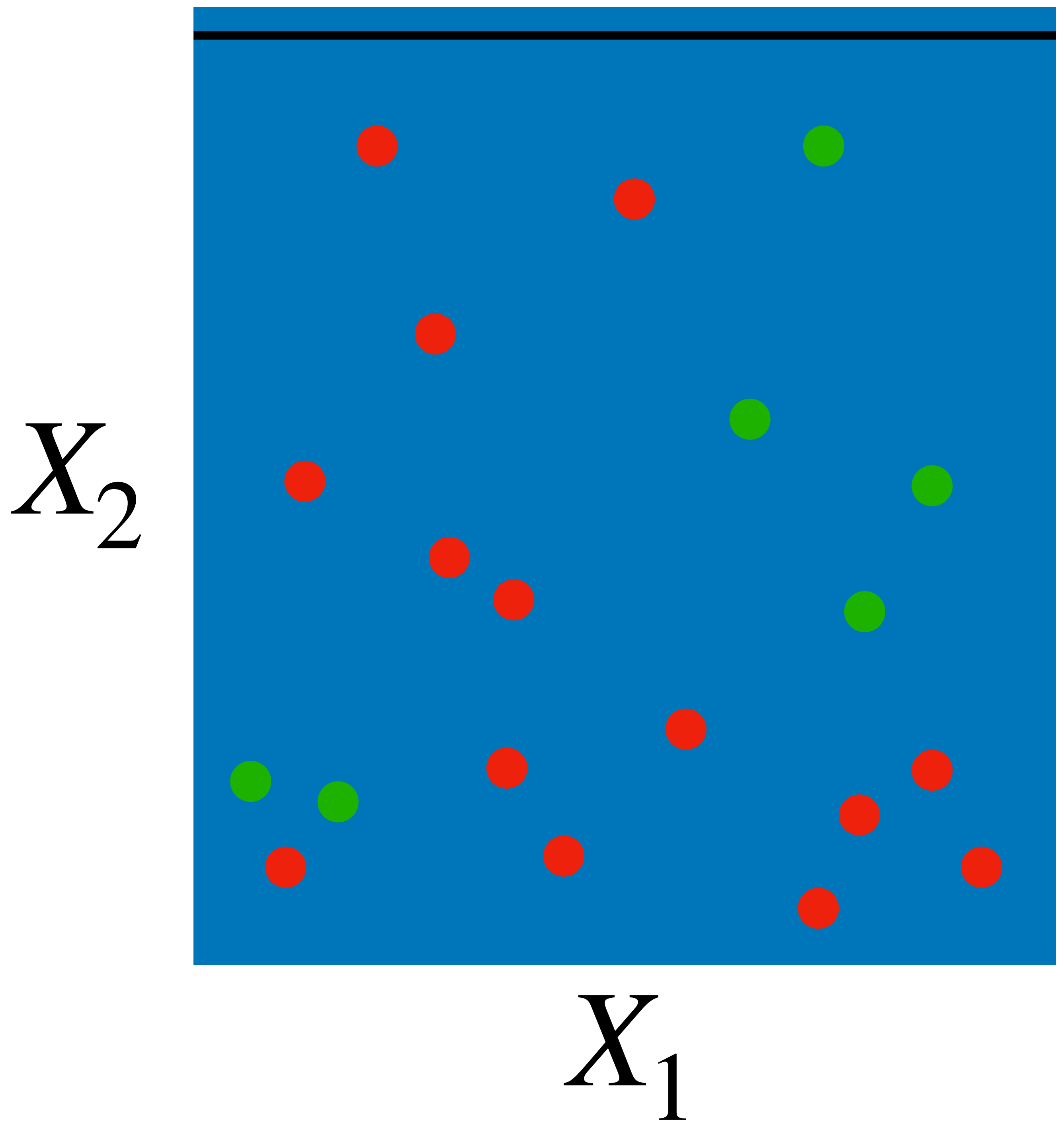
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.
2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

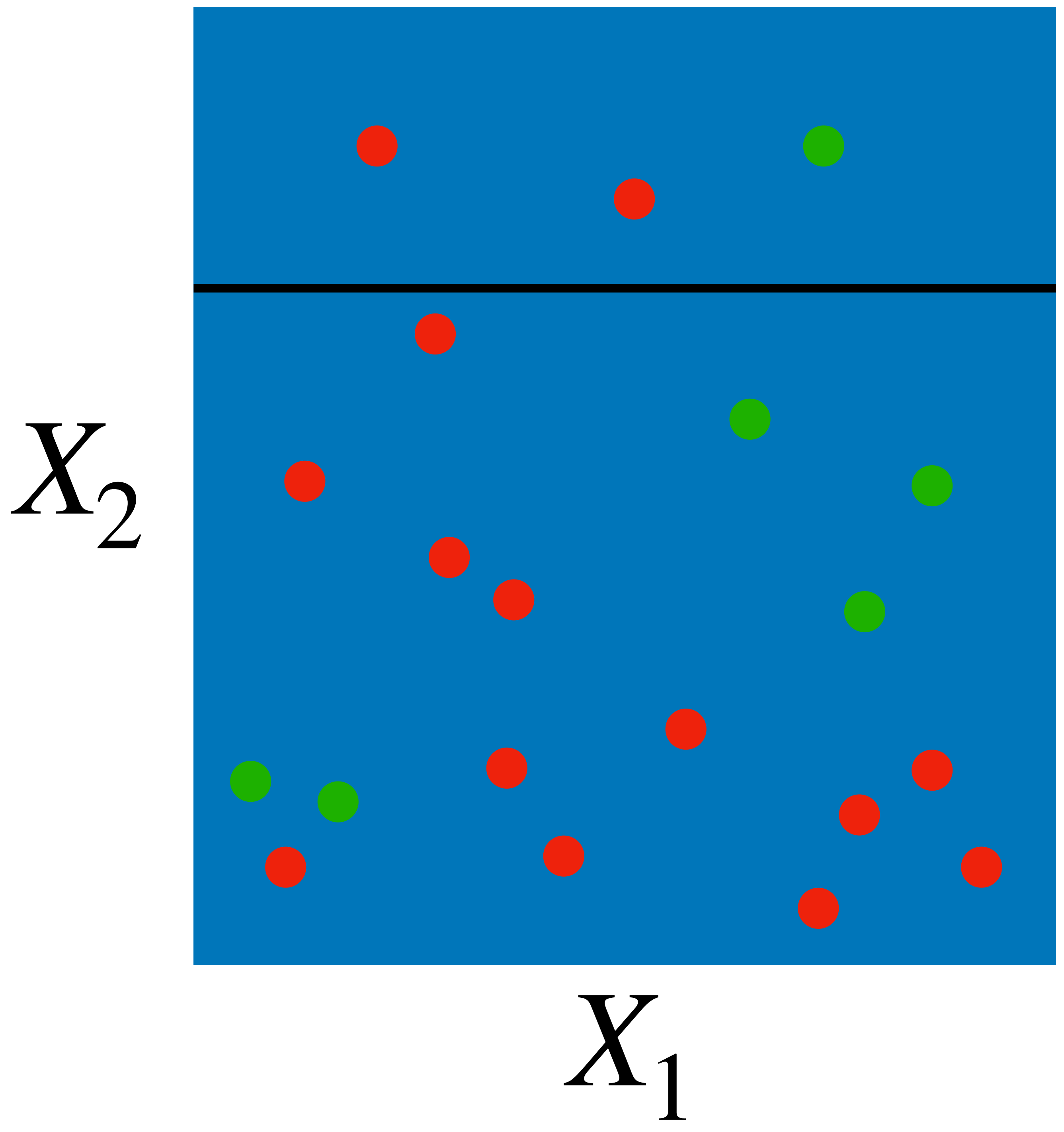
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

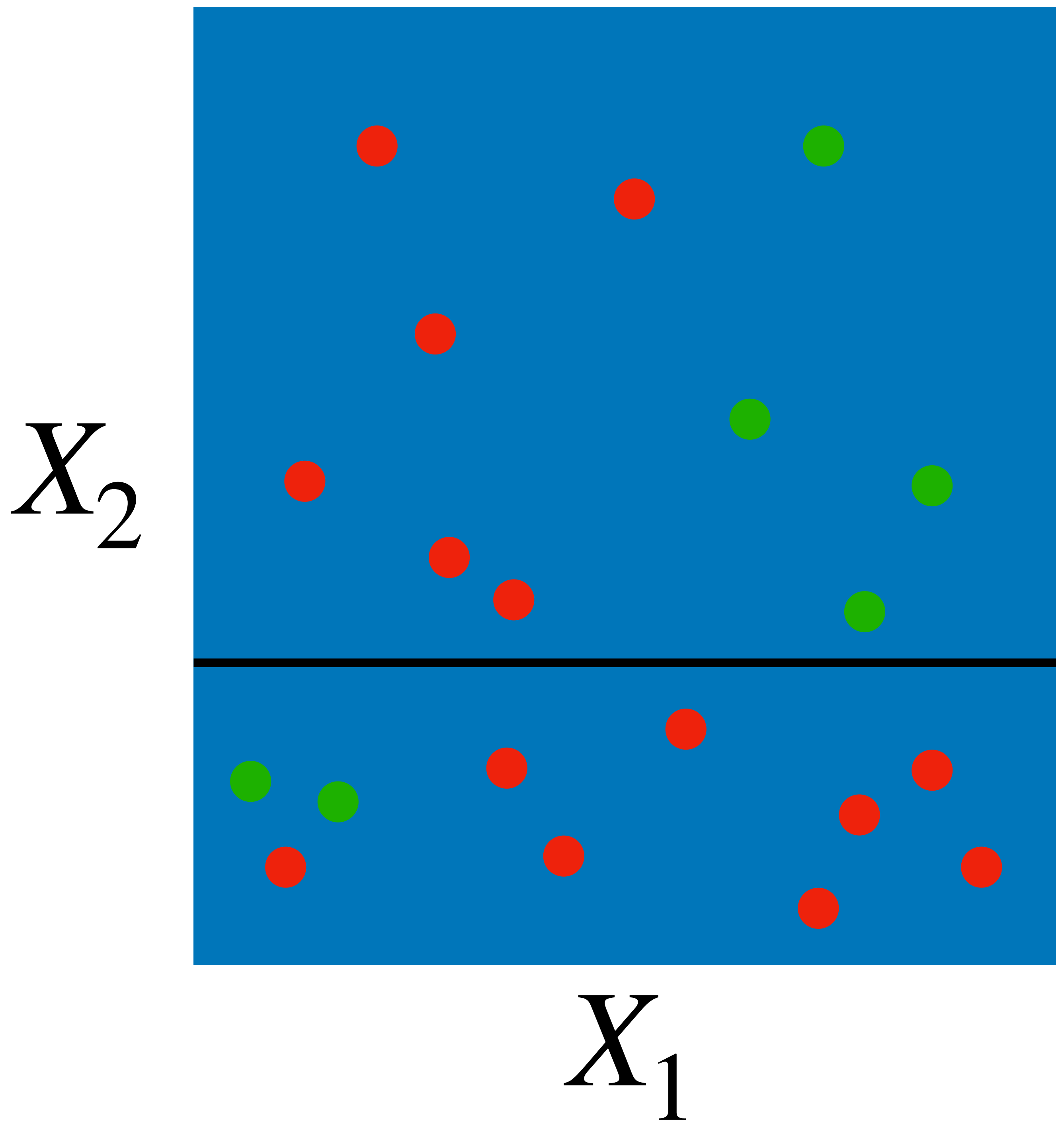
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

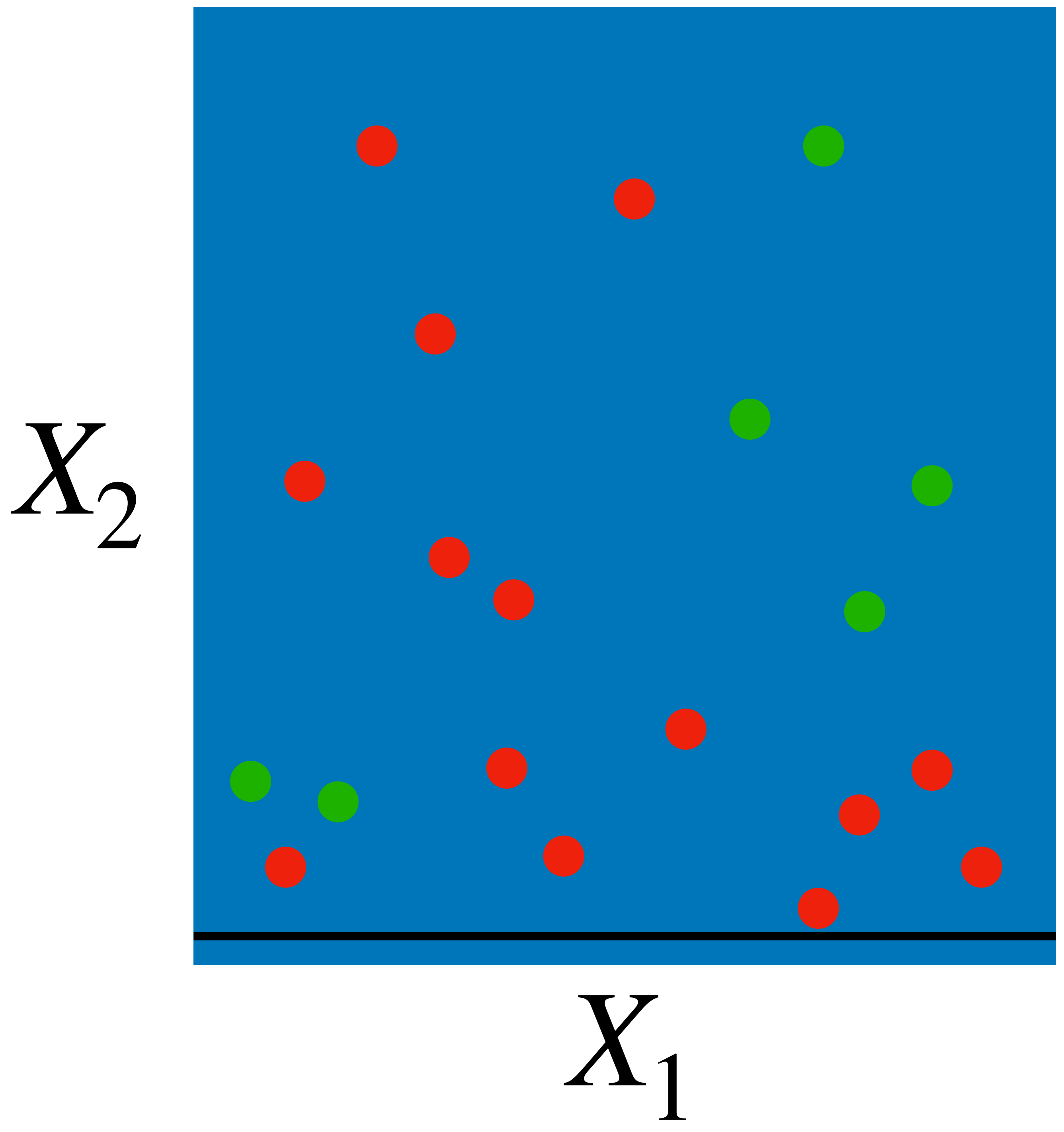
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

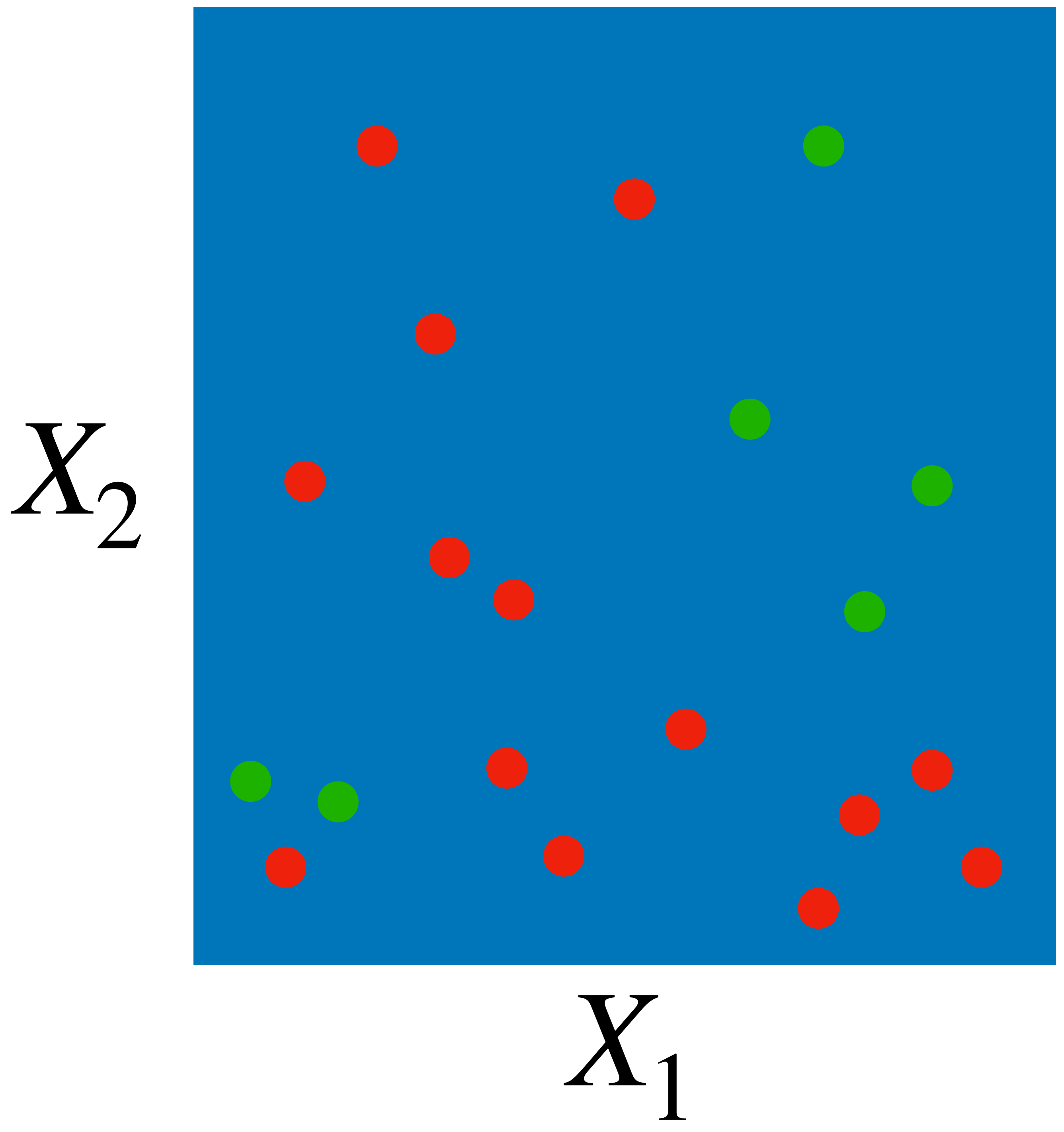
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

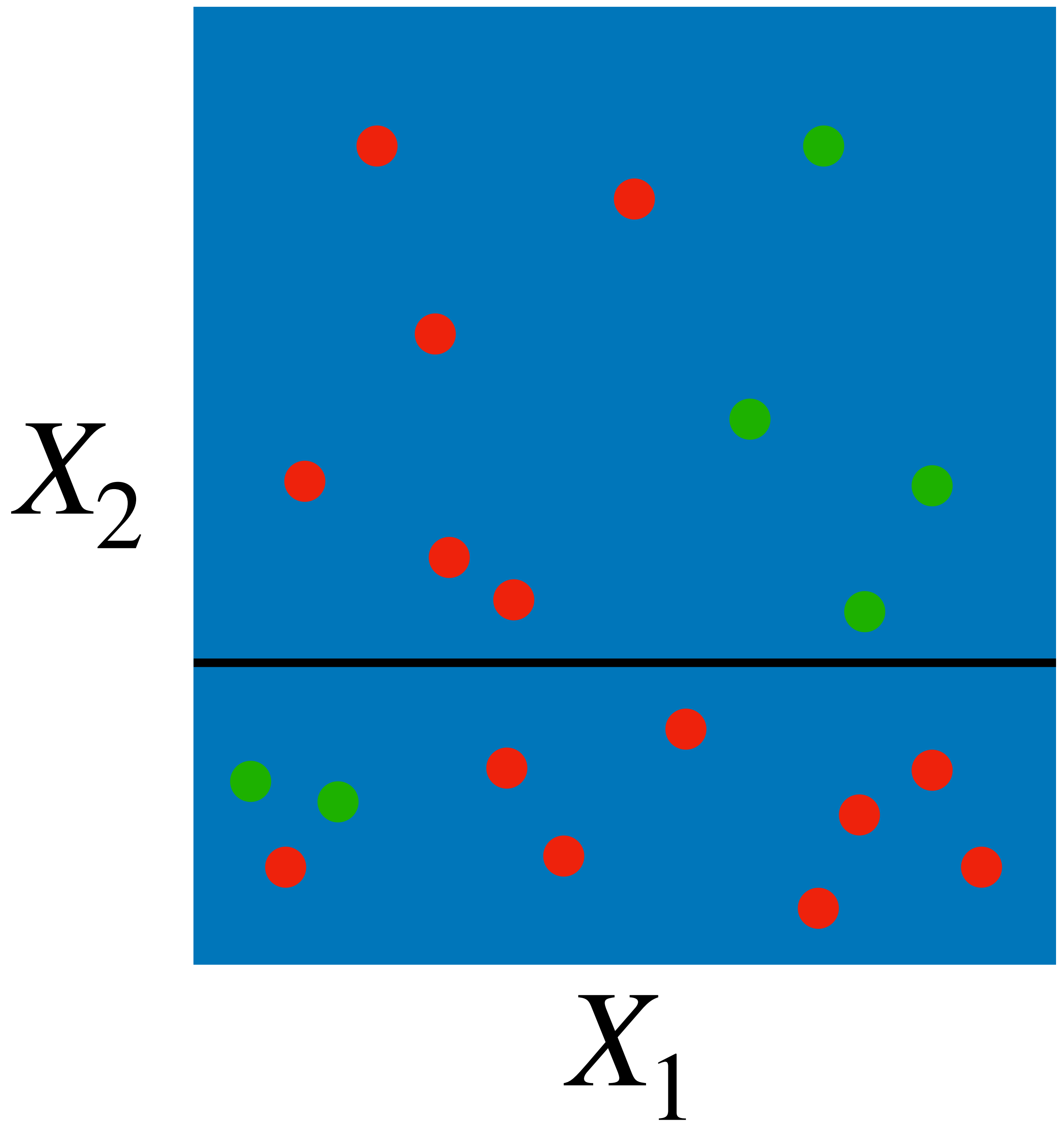
The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

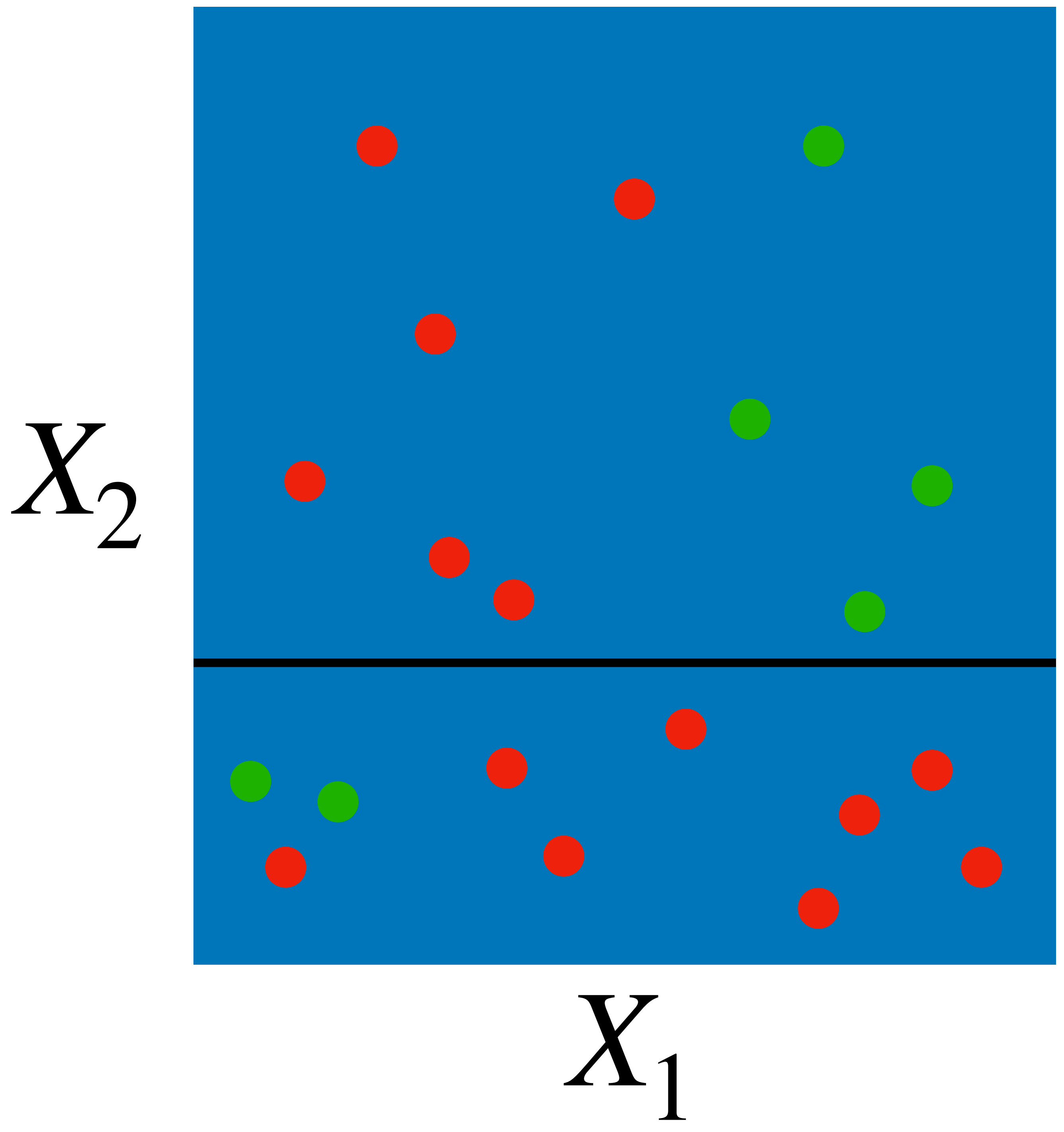
Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.
2. Find split of the whole region that decreases total Gini index the most.
3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1-\hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

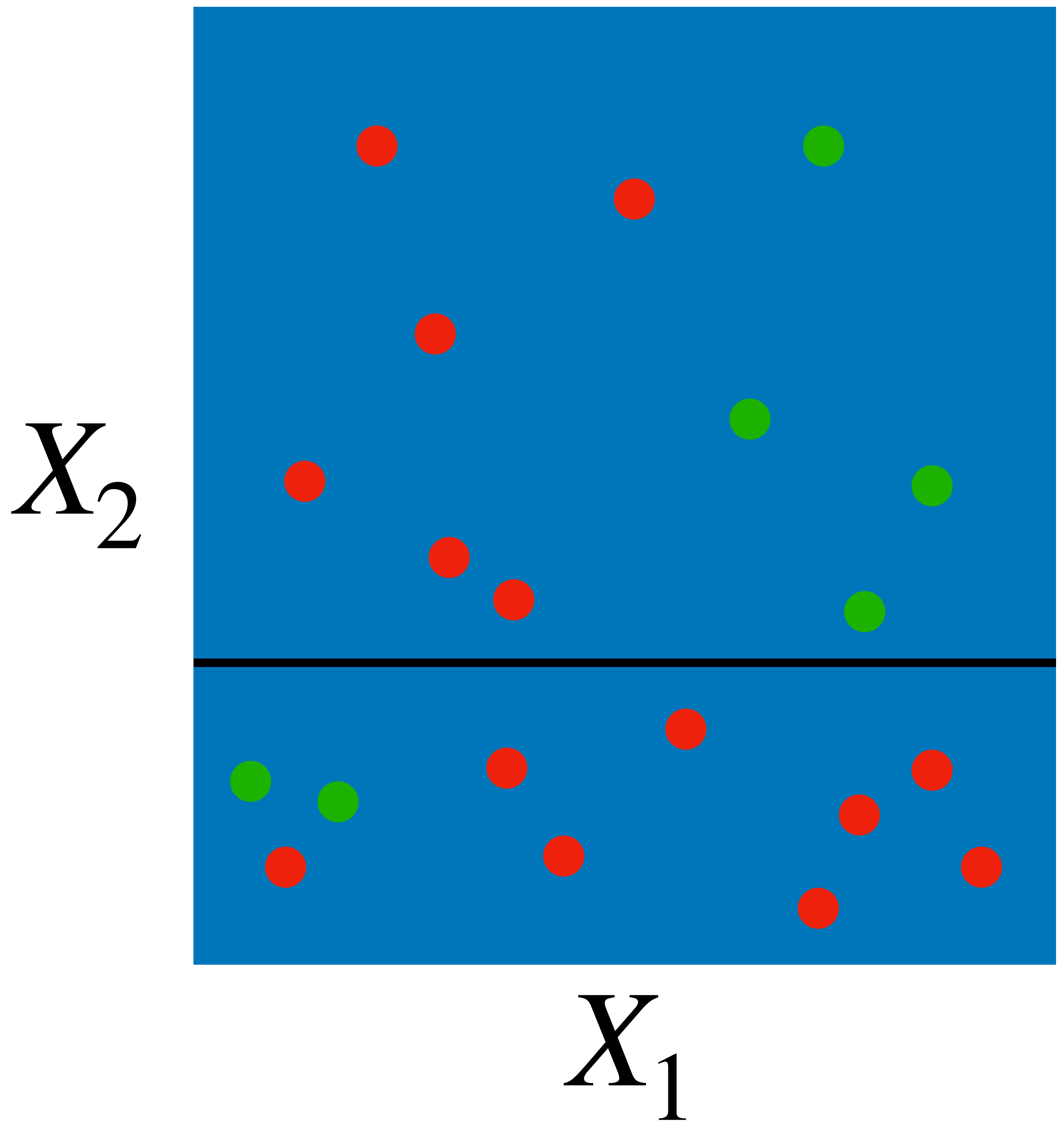
1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1 - \hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1 - \hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1 - \hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

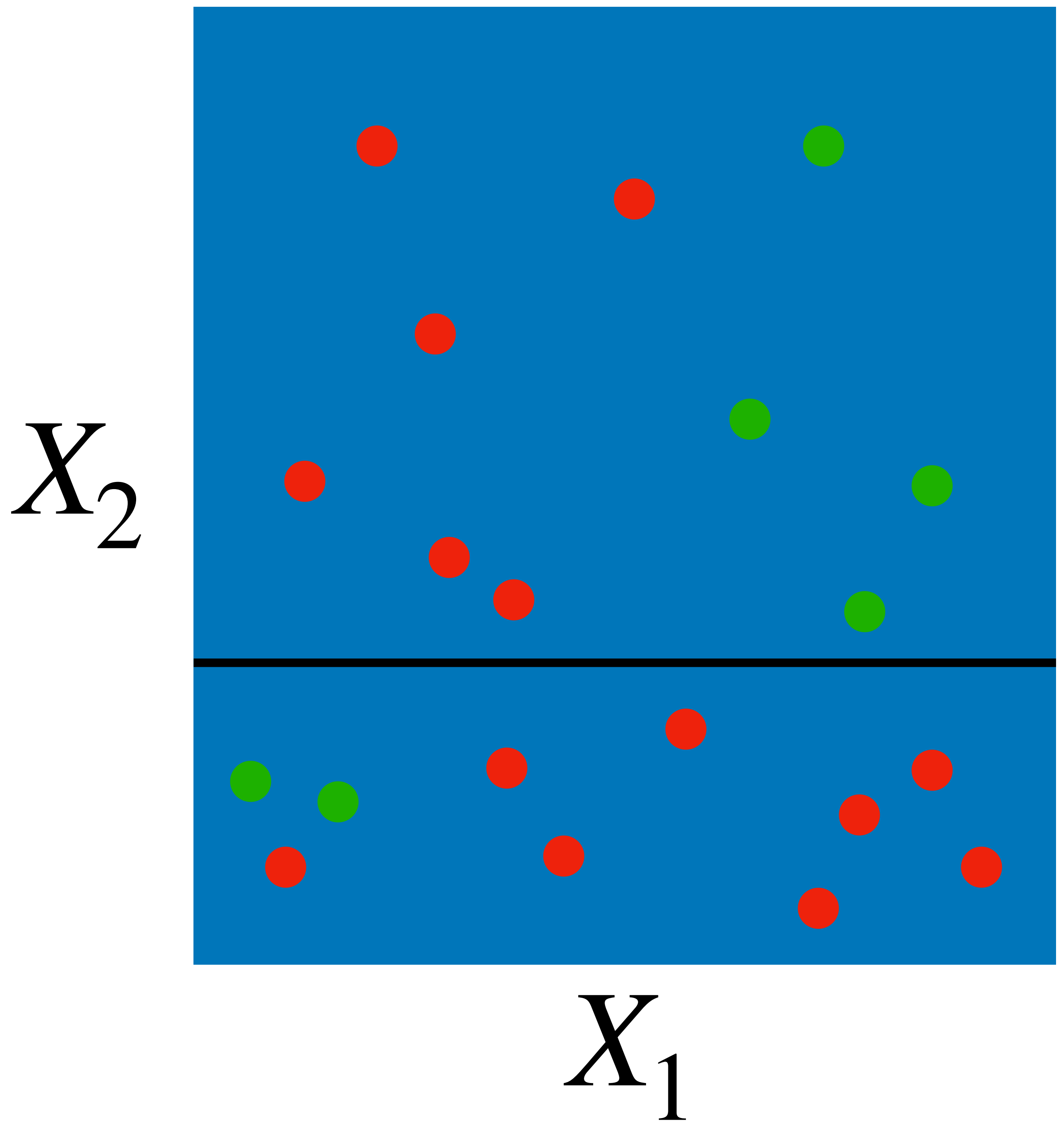
1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1 - \hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1 - \hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1 - \hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

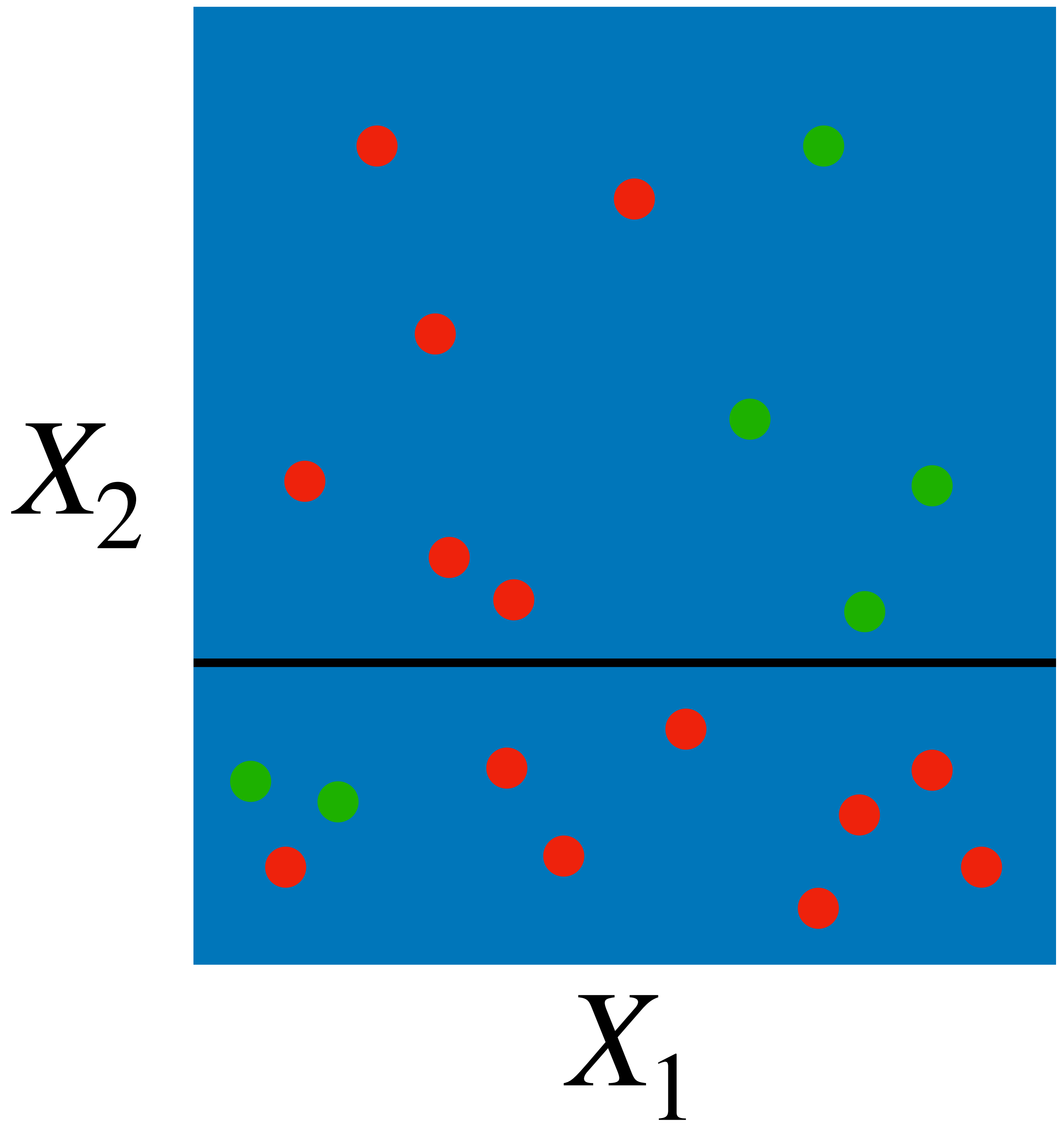
1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1 - \hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1 - \hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1 - \hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

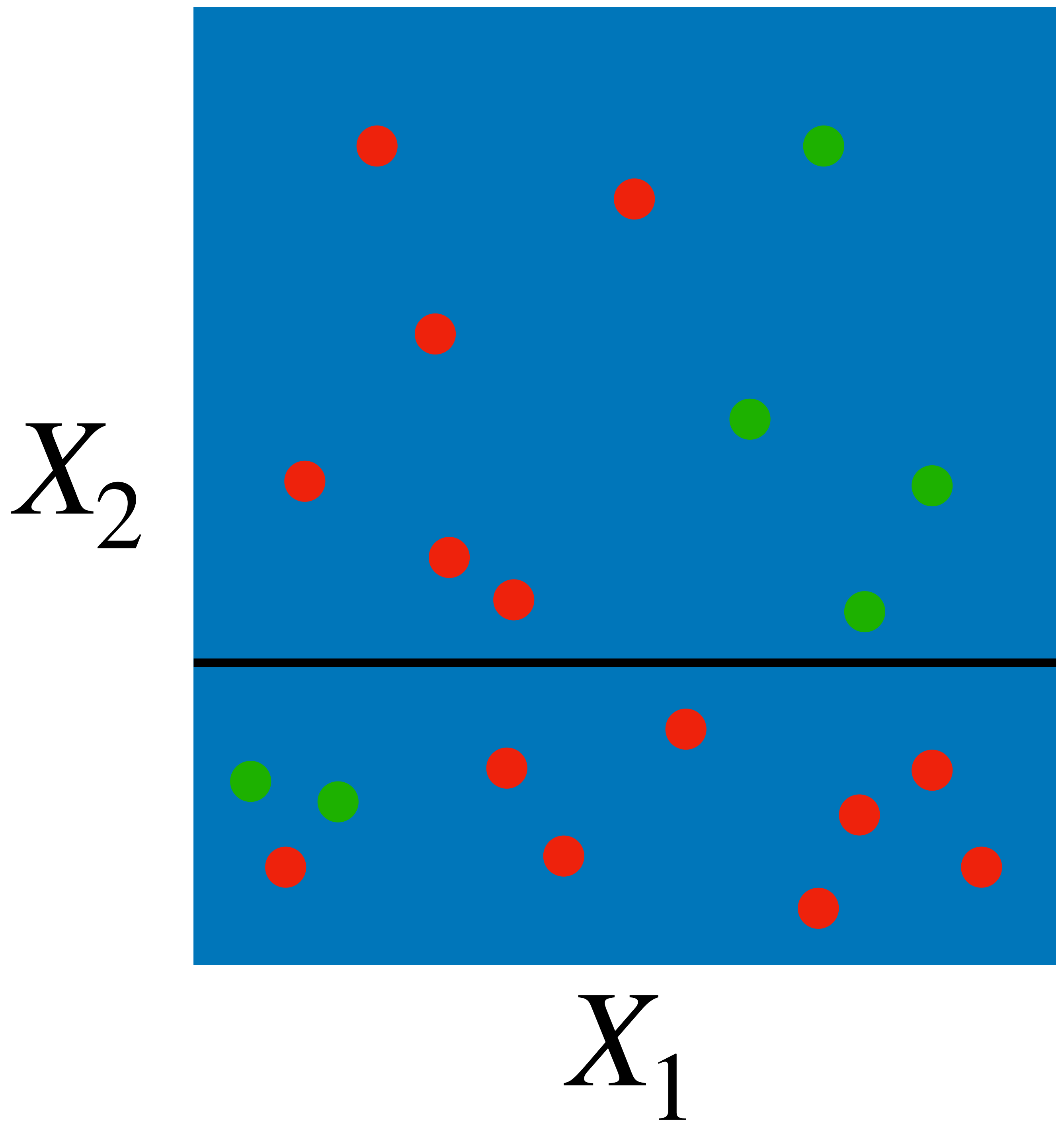
1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1-\hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

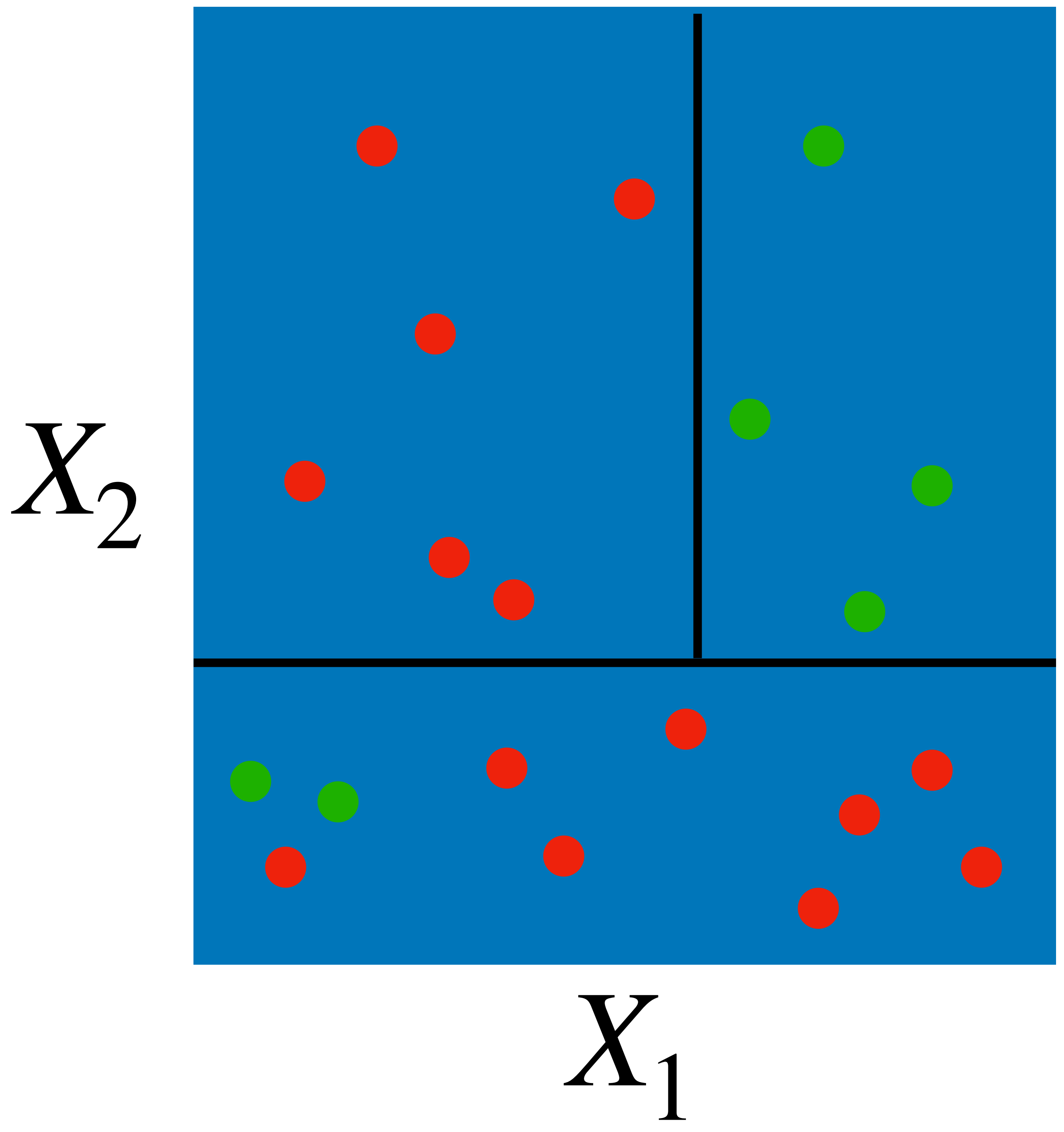
1. Fit constant model to the entire space and calculate (total) Gini index.

2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1-\hat{p}_m)$.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

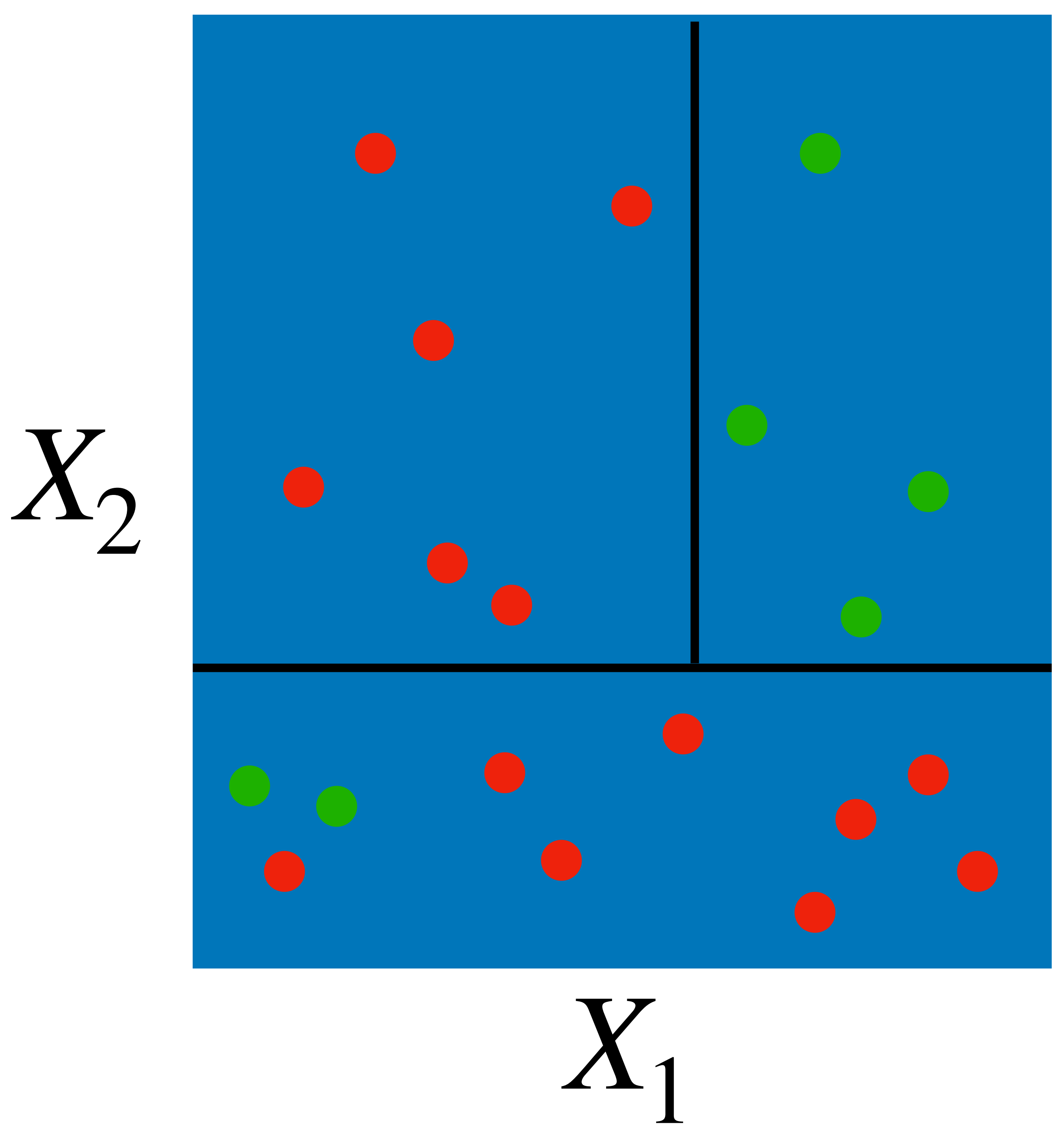
2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1-\hat{p}_m)$.

4. Repeat until there are M regions.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1-\hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1-\hat{p}_M)\right\}$$



Training a classification tree

Finding the rectangles $\widehat{R}_m$

The optimal set of rectangles is computationally intractable to find. In practice, we employ a greedy top-down algorithm based on total Gini index:

1. Fit constant model to the entire space and calculate (total) Gini index.

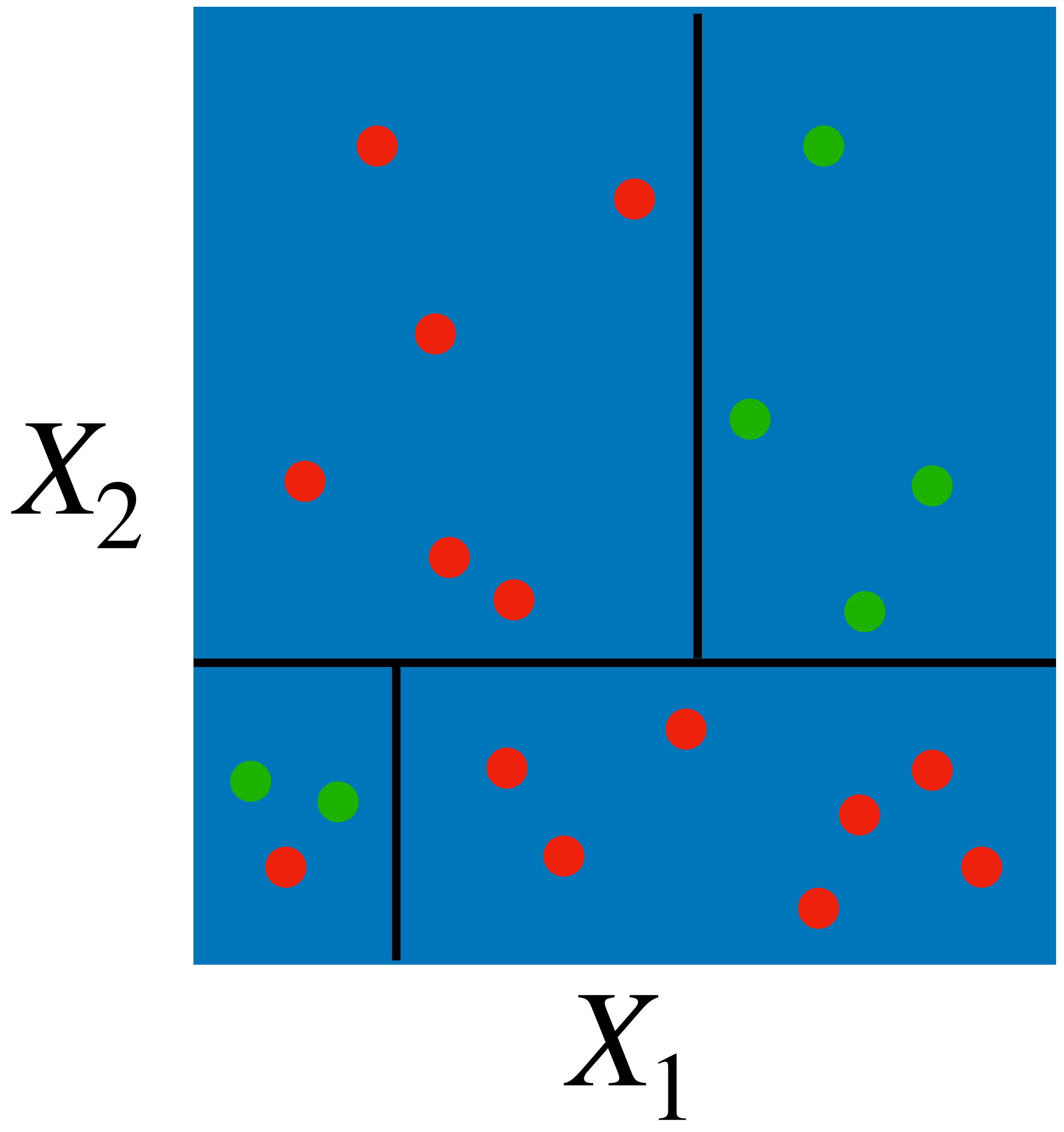
2. Find split of the whole region that decreases total Gini index the most.

3. Find the split for each rectangle that decreases its Gini index the most. Among these, choose largest decrease in $n_m \cdot 2\hat{p}_m(1 - \hat{p}_m)$.

4. Repeat until there are M regions.

Total Gini index =

$$\frac{1}{n}\left\{n_1 \cdot 2\hat{p}_1(1 - \hat{p}_1) + \cdots + n_M \cdot 2\hat{p}_M(1 - \hat{p}_M)\right\}$$

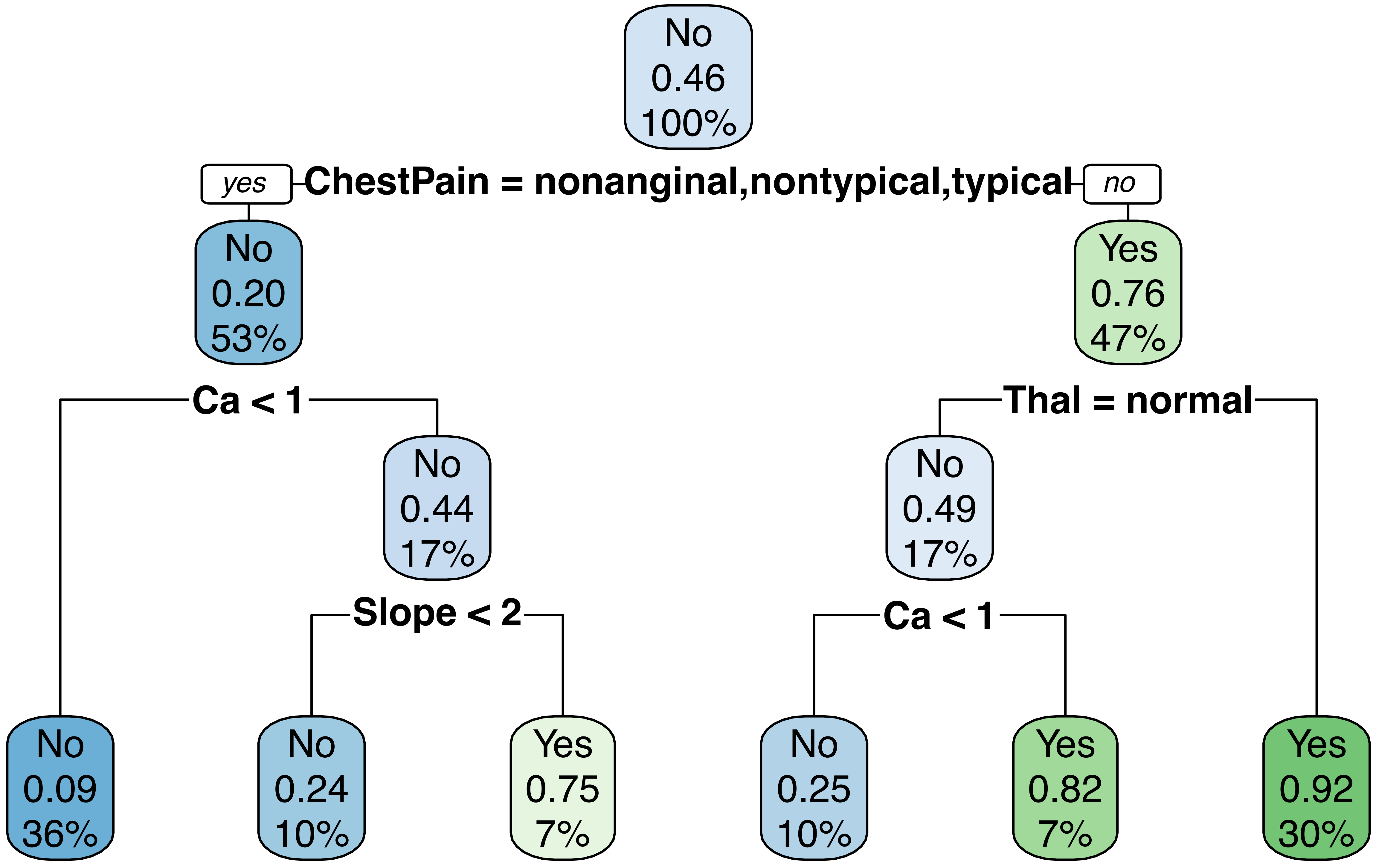


Training a classification tree

Final output

Example: Heart disease data set.

- 303 patients with chest pain
- Binary response HD (heart disease)
- 13 demographic and clinical features

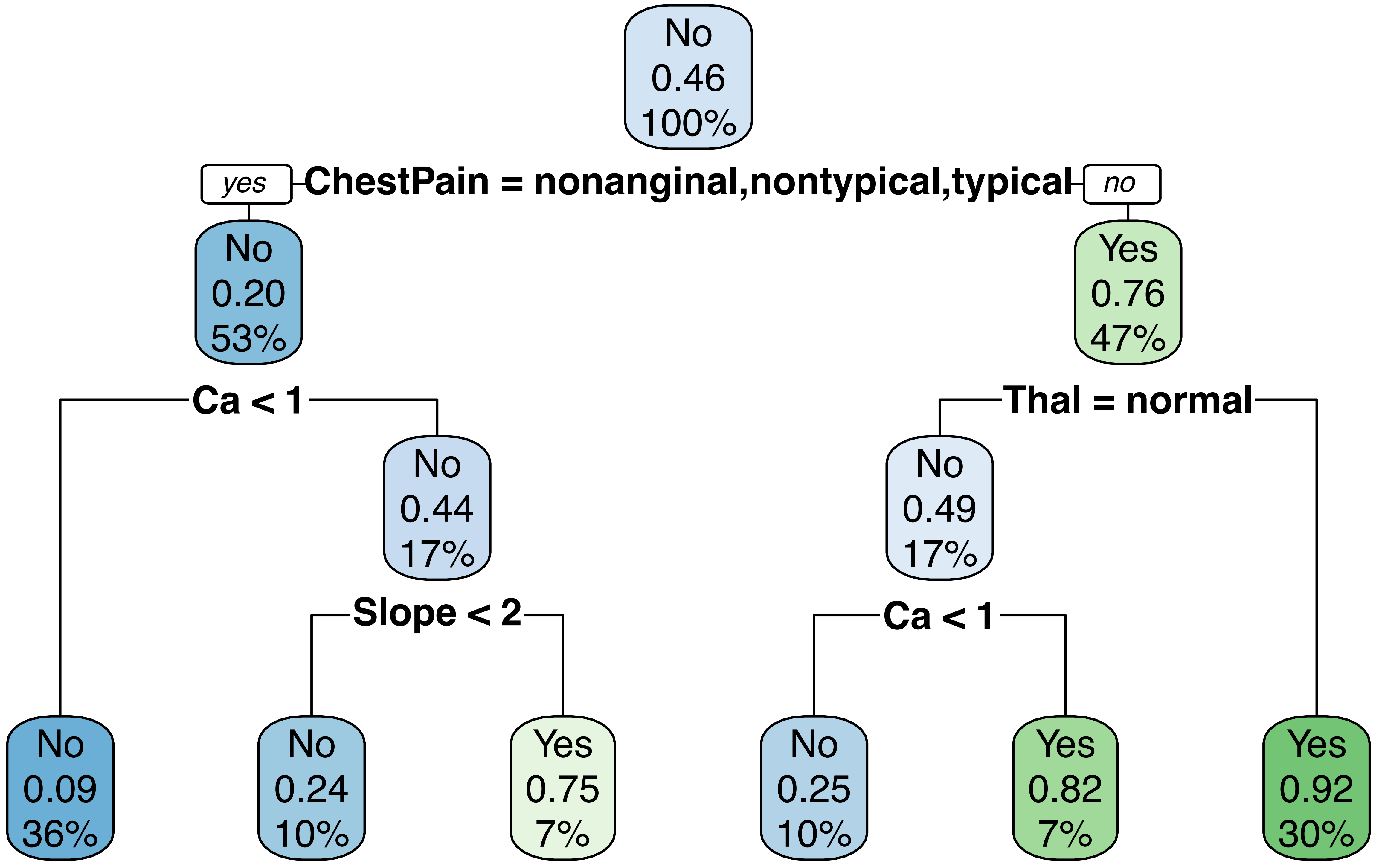


Training a classification tree

Final output

Example: Heart disease data set.

- 303 patients with chest pain
- Binary response HD (heart disease)
- 13 demographic and clinical features



Note: Classification trees extend seamlessly to more than two classes!

Tree-based models versus linear models

Which perform better?

Tree-based models versus linear models

Which perform better?

Neither tree-based nor linear models dominate the other.

Tree-based models versus linear models

Which perform better?

Neither tree-based nor linear models dominate the other.

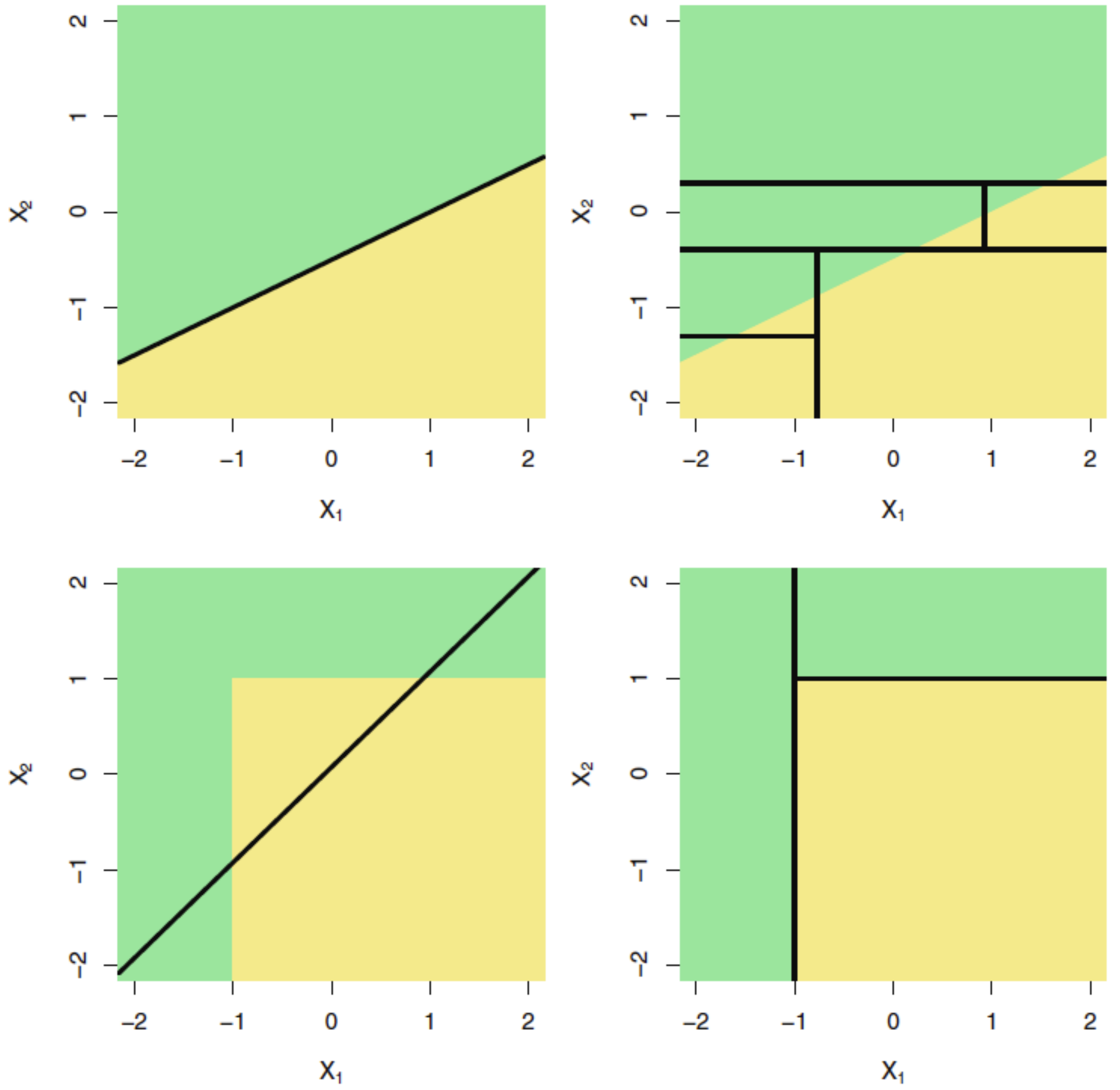
Each prediction method works better when the underlying trend in the data matches its modeling choice.

Tree-based models versus linear models

Which perform better?

Neither tree-based nor linear models dominate the other.

Each prediction method works better when the underlying trend in the data matches its modeling choice.



Classification based on two features (colors indicate the two classes).

Tree-based models versus linear models

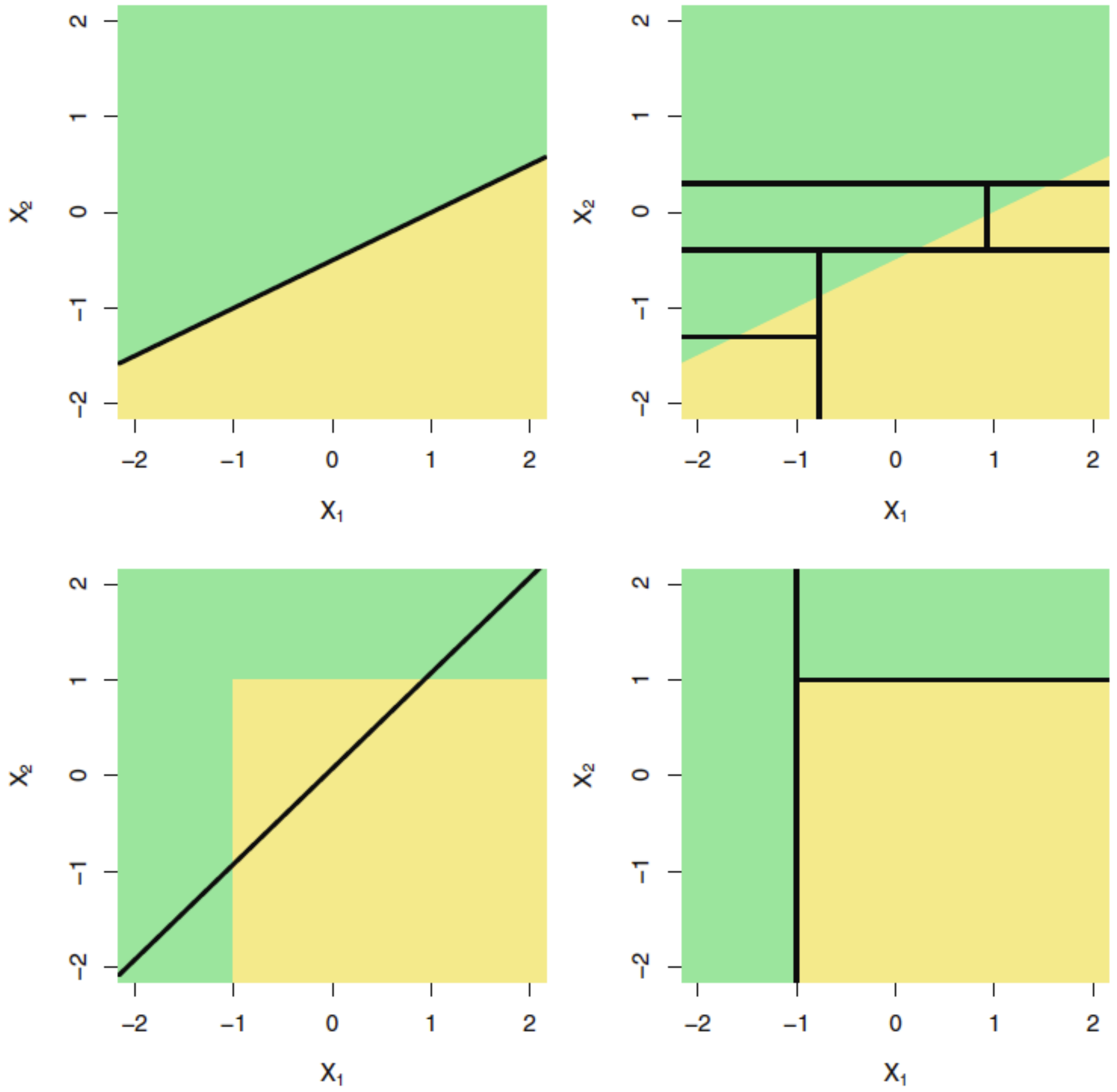
Which perform better?

Neither tree-based nor linear models dominate the other.

Each prediction method works better when the underlying trend in the data matches its modeling choice.

E.g. for classification:

- Linear model → linear decision boundary
- Decision tree → unions of rectangles



Classification based on two features (colors indicate the two classes).

Summary

- Decision trees partition the feature space into axis-aligned nested rectangles, producing a constant prediction for feature vectors in each rectangle.
- Decision trees are built by recursively choosing
  - The optimal rectangle to split
  - The optimal feature to split that rectangle on
  - The optimal split-point for that feature
- Regression and classification trees aim to minimize squared error and misclassification error, respectively.