



Data Analysis Block Course: Statistics and Machine learning

Exercise: Decision Trees by Hand

Prof. Dr. Mikhail Mikhasenko

Assistants: Hendric Jonas, Christoph Tönnis



28 September 2026

Indico schedule

RUB Moodle

Learning goals

After this exercise, you should be able to

- explain what a regression tree predicts inside a leaf,
- explain why the mean is the optimal constant prediction when squared error is used,
- interpret residuals, mean squared error (MSE), and sum of squared errors (SSE),
- use the mean for squared error and the median for absolute error,
- determine all relevant candidate split points for a one-dimensional feature,
- reproduce the main steps that a regression-tree training algorithm performs when choosing a split,
- construct a shallow regression tree by hand,
- explain why a regression tree produces a piecewise-constant prediction,
- explain qualitatively how increasing tree depth affects approximation quality and overfitting,
- extend the same reasoning to a dataset with more than one input feature,
- train trees using absolute error and combine trees fitted to bootstrap samples into a forest,
- distinguish training error from prediction error at unobserved input positions.

The calculations in this sheet are designed to be done by hand. No programming is required.

Part I - Understanding a single leaf

Regression means predicting a numerical value. An **observation** is one data point; its **features** are the input values, such as x , and its **target** y is the value to be predicted. The **training data** are the observations used to choose the model. A prediction is written $\hat{y}$, and the fitted function is written $\hat{f}$.

A **regression tree** divides the input space using a sequence of decisions. A **node** contains the observations that reach a particular stage. The **root node** initially contains all training observations. A **split** divides a parent node into two **child nodes**. An **internal node** makes a split; a **leaf** makes no further split and assigns one constant prediction to every input reaching it.

Throughout this sheet, split rules send $x < t$ to the left child and $x \geq t$ to the right child. The **threshold** t is the boundary used in that decision. We use squared error to train the trees, except when explicitly comparing other losses.

Consider a leaf that contains four target values

observation	target y_i
1	2
2	3
3	4
4	7

The **mean** is the arithmetic average: add the values and divide by their number.

A **residual** $r_i = y_i - \hat{y}_i$ is the signed difference between target and prediction. A **loss** measures prediction error. The **sum of squared errors (SSE)** adds the squared residuals; the **mean squared error (MSE)** divides this sum by the number N of observations. Thus $\text{MSE} = \text{SSE}/N$, with $N > 0$.

Exercise 1.1 - One constant prediction

Assume that the leaf predicts one constant value c for all four observations, i.e. we construct a tree with only a single leaf.

1. Compute the mean of the four target values. This will be the value of the leaf output.
2. Use this mean as the prediction c .
3. Compute the residual

$$r_i = y_i - c$$

for every observation.

4. Compute the sum of squared errors

$$\text{SSE}(c) = \sum_i (y_i - c)^2 \quad (1)$$

5. Compute the mean squared error

$$\text{MSE}(c) = \frac{1}{N} \sum_i (y_i - c)^2 \quad (2)$$

Exercise 1.2 - Is the mean special?

Consider the squared-error loss as a function of an arbitrary constant c :

$$L(c) = \sum_{i=1}^N (y_i - c)^2 \quad (3)$$

1. Differentiate $L(c)$ with respect to c .
2. Set the derivative equal to zero to find a candidate for the minimum.
3. Show that the minimizer is

$$c = \frac{1}{N} \sum_i y_i$$

4. Check that the second derivative is positive for $N > 0$.

Exercise 1.3 - What does MSE actually measure?

The **root mean squared error (RMSE)** is $\sqrt{\text{MSE}}$.

Give short answers:

1. Is the MSE measured in the same units as the target y ?
2. Why do large residuals contribute particularly strongly to the MSE?
3. What would an MSE of zero mean?
4. Calculate the RMSE for Exercise 1.1 and state its units.

Exercise 1.4 - Could we use another criterion?

The **mean absolute error (MAE)** averages the absolute residuals:

$$\text{MAE}(c) = \frac{1}{N} \sum_i |y_i - c| \quad (4)$$

A **median** minimizes this loss. To find it, sort the target values and take the middle value. For an even number of values, use the arithmetic mean of the two middle values.

1. Determine the median of the four targets from Exercise 1.1.
2. Use it as the leaf prediction and calculate the MAE.

Part II - How does a regression tree find its first split?

Consider

$$f(x) = x^2 \quad (5)$$

on the interval

$$0 \leq x \leq 6$$

A **candidate split** is a feature and threshold being considered by the training algorithm. For a numerical feature, use the midpoints between neighboring **distinct**, sorted feature values present in the node. Each candidate must produce two non-empty children.

Exercise 2.1 - Generate the training data by hand

Observe the function at the integer positions

$$x = 0, 1, 2, 3, 4, 5, 6$$

and complete the following table:

x	$y = f(x) = x^2$
0	
1	
2	
3	
4	
5	
6	

Plot these seven training points in a sufficiently large diagram and also sketch the true function $f(x) = x^2$

Exercise 2.2 - Before splitting: the root leaf

Initially, the tree contains only one leaf. Therefore all seven observations receive the same prediction (i.e. the mean value as you have seen in Part 1).

1. Compute the mean target value in the root leaf.
2. Draw this constant prediction in the same graph as $f(x)$.
3. Compute the SSE of the root leaf.

This is the error before the tree has made any split.

Exercise 2.3 - Which split points does the algorithm need to test?

The feature values are already sorted:

0, 1, 2, 3, 4, 5, 6

A split has the form

$$x < t$$

A threshold only changes the division of the training data when it moves from one side of an observed x value to the other.

1. List all candidate midpoints for this dataset.
2. Explain why testing both $t = 0.6$ and $t = 0.8$ would be unnecessary.

Exercise 2.4 - Follow one candidate split step by step

For each candidate, compute a separate mean prediction for each child. Choose the candidate with the smallest **total child SSE**, $SSE_L + SSE_R$. This is equivalent to minimizing the child MSEs weighted by their numbers of observations, not their unweighted average.

The **error reduction** is the parent SSE minus the total child SSE.

Tree depth is the largest number of splits along a path from the root to a leaf. A single root leaf has depth 0; splitting it once produces depth 1. A **piecewise-constant prediction** has a constant value inside each leaf region and may jump at its boundaries.

Take the candidate threshold $t = 2.5$, with split rule $x < 2.5$.

1. Assign observations to the two child nodes.

Write down the x and y values that go to the left child:

$x < 2.5$	y

Write down the x and y values that go to the right child:

$x \geq 2.5$	y
--------------	-----

2. Determine the prediction in each child.

Compute

$$\hat{y}_L = \text{mean}(y_{\text{in left child}})$$

and

$$\hat{y}_R = \text{mean}(y_{\text{in right child}})$$

3. Measure the error after the split.

Compute

$$\text{SSE}_L = \sum_{i \in L} (y_i - \hat{y}_L)^2 \quad (6)$$

$$\text{SSE}_R = \sum_{i \in R} (y_i - \hat{y}_R)^2 \quad (7)$$

and finally

$$\text{SSE}_{\text{split}} = \text{SSE}_L + \text{SSE}_R \quad (8)$$

Compare this error with the SSE of the unsplit root leaf.

Did the split improve the model?

Exercise 2.5 - Let the tree choose the best split

The training algorithm repeats exactly the calculation above for every candidate threshold.

Complete the table.

threshold t	mean left $\hat{y}_L$	mean right $\hat{y}_R$	SSE_L	SSE_R	total SSE
0.5					
1.5					
2.5					
3.5					
4.5					
5.5					

The decision rule of the tree is:

"Choose the split with the smallest total child error"

Equivalently, choose the split with the largest error reduction

$$\Delta\text{SSE} = \text{SSE}_{\text{parent}} - (\text{SSE}_L + \text{SSE}_R)$$

1. Which threshold is selected?
2. Write down the two leaf predictions.
3. Draw the resulting depth-1 tree.
4. Draw the resulting prediction in the graph from Exercise 2.1. Explain why it is piecewise constant.

Exercise 2.6 - Split the left child

Apply the same split-selection procedure within each child node. Allow one additional split in each child, producing a depth-2 tree.

The **training error** is the prediction error on the observations used to build the tree. **Noise** is random variation in the measured targets. **Overfitting** occurs when a model adapts to sample-specific details or noise in a way that harms its predictions on new data.

Take the left child produced by the best split.

1. List the observations contained in this child.
2. Determine all candidate thresholds that can still split these observations into two non-empty groups.
3. For every candidate, calculate the total child SSE.
4. Select the best split.

Exercise 2.7 - Split the right child

1. List the observations in the right child and its candidate thresholds.
2. If only two observations remain, how many candidate midpoints are possible?
3. Calculate the total child SSE and determine the best split.

Exercise 2.8 - Draw the depth-2 tree

1. Draw the complete depth-2 tree. Label every internal node with its split condition and every leaf with its constant prediction.
2. Draw the corresponding step function $\hat{f}(x)$ in the same coordinate system as $f(x) = x^2$.

Exercise 2.9 - Compare the approximations

Use the errors already calculated to complete the table.

Model	Training SSE
One root leaf	
Depth-1 tree	
Depth-2 tree	

Which of these three models has the smallest training error?

Part III - More than one input feature

A regression tree can choose not only where to split, but also which feature to split. Consider the two-dimensional function

$$f(x_1, x_2) = x_1^2 + 4x_2 \quad (9)$$

We sample

$$x_1 \in \{0, 1, 2, 3\} \quad x_2 \in \{0, 1\}$$

The **feature space** consists of all possible combinations of input values. An **axis-aligned split** tests one feature at a time. In the (x_1, x_2) plane, draw x_1 horizontally and x_2 vertically; such splits have boundaries parallel to a coordinate axis. In two dimensions, the resulting regions are rectangles, possibly unbounded.

Exercise 3.1 - Understand the function before building a tree

First fix $x_2 = 0$.

Then

$$f(x_1, 0) = x_1^2 \quad (10)$$

Now fix $x_2 = 1$.

Then

$$f(x_1, 1) = x_1^2 + 4 \quad (11)$$

1. Evaluate both functions at $x_1 = 0, 1, 2, 3$.
2. Complete the table.

x_1	x_2	$y = f(x_1, x_2)$
0	0	
1	0	
2	0	
3	0	
0	1	
1	1	
2	1	
3	1	

3. In separate coordinate systems, plot $f(x_1, 0)$ and $f(x_1, 1)$ as functions of x_1 .

Exercise 3.2 - Which feature is chosen first?

The tree may now test conditions involving either feature. For x_1 and x_2 , these are 4 thresholds in total, 3 for conditions involving x_1 and one for x_2 , thus, at the root node, the algorithm must compare four candidate splits.

Complete the following table by repeating the same steps as in Exercises 2.4-2.5:

1. assign observations to the left and right child,
2. calculate the mean target in each child,
3. calculate the SSE in each child,
4. add the two errors.

feature	threshold	mean left	mean right	total SSE
x_1	0.5			
x_1	1.5			
x_1	2.5			
x_2	0.5			

Which feature and threshold are selected?

Exercise 3.3 - Grow one more level

Now allow each child node to split once more.

For each child:

1. determine which observations are present,
2. determine which candidate splits are still possible,
3. calculate their total SSE,
4. choose the best split.

Draw the resulting depth-2 tree.

Then again sketch the prediction as a function of x_1

- for $x_2 = 0$,
- for $x_2 = 1$.

Part IV - A forest by hand

You will first train a regression tree using absolute error. You will then train three trees on bootstrap samples, average their predictions to form a forest, and compare the predictions with those of the single tree.

Exercise 4.1 - Training data

Consider

$$f(x) = \sin(x) + \frac{x}{\pi} \quad 0 \leq x \leq 2\pi \quad (12)$$

Each input position occurs exactly once in the training dataset.

x	$f(x)$
0	
$\pi/2$	
π	
$3\pi/2$	
2π	

1. Complete the table.
2. Plot the five training points and sketch the function.

Exercise 4.2 - Train a tree using absolute error

The **median** is the middle value of a sorted list. For an odd number of values, take the single middle value. For an even number, use the arithmetic mean of the two middle values.

For example, the median of 1, 3, 8 is 3, and the median of 1, 3, 5, 8 is $(3 + 5)/2 = 4$.

A leaf prediction c has mean absolute error

$$\text{MAE}(c) = \frac{1}{N} \sum_{i=1}^N |y_i - c| \quad (13)$$

A median minimizes this error. Determine the leaf median by sorting the **target values**.

For each threshold, send $x < t$ to the left child and $x \geq t$ to the right child. The child sets are given below. Dashes mark unused cells. Fill in the function values and enter each child's median and MAE in its $f(x_i)$ row.

Threshold $t = \pi/4$

Child		1	2	3	4	Median	MAE
L	x_i	0	-	-	-	-	-
	$f(x_i)$		-	-	-		
R	x_i	$\pi/2$	π	$3\pi/2$	2π	-	-
	$f(x_i)$						

Threshold $t = 3\pi/4$

Child		1	2	3		Median	MAE
L	x_i	0	$\pi/2$	-		-	-
	$f(x_i)$			-			
R	x_i	π	$3\pi/2$	2π		-	-
	$f(x_i)$						

Threshold $t = 5\pi/4$

Child		1	2	3	Median	MAE
L	x_i	0	$\pi/2$	π	-	-
	$f(x_i)$					
R	x_i	$3\pi/2$	2π	-	-	-
	$f(x_i)$			-		

Threshold $t = 7\pi/4$

Child		1	2	3	4	Median	MAE
L	x_i	0	$\pi/2$	π	$3\pi/2$	-	-
	$f(x_i)$						
R	x_i	2π	-	-	-	-	-
	$f(x_i)$		-	-	-		

Compare the thresholds

1. Complete the four tables above.
2. Calculate the total absolute error for each threshold:

$$S = N_L \text{MAE}_L + N_R \text{MAE}_R$$

where N_L and N_R are the numbers of observations in the children.

3. Enter the results below.

Threshold	Sum S
$\pi/4$	
$3\pi/4$	
$5\pi/4$	
$7\pi/4$	

4. Select the threshold with the smallest S . If thresholds give the same error, choose the smaller threshold.
5. Draw the depth-1 tree and label its leaves with their median predictions.
6. Calculate the tree's training MAE as $S/5$.
7. Calculate the median and MAE of the unsplit dataset. Compare the errors before and after splitting.

Exercise 4.3 - Train three trees

A **bootstrap sample** is obtained by drawing observations with replacement from the training dataset. After each draw, all original observations remain available for the next draw. An observation can therefore occur several times or be absent from a sample.

Use five draws per tree. Three possible samples are specified below; each column sums to five.

x	Sample A	Sample B	Sample C
0	2	1	2
$\pi/2$	1	0	0
π	1	1	2
$3\pi/2$	0	1	0
2π	1	2	1

Train one depth-1 tree per sample using the prepared tables below.

1. Fill in the function values and calculate the median and MAE for each child. Count each repeated observation separately when sorting targets and calculating errors.
2. Complete the summary table for each sample using $S = N_L \text{MAE}_L + N_R \text{MAE}_R$.
3. Select the best split using the rule from Exercise 4.2.
4. Draw each tree and write its prediction as a piecewise-constant function.

The candidate thresholds are the midpoints between neighboring distinct input values present in each sample.

Sample A

Threshold $t = \pi/4$

Child		1	2	3	Median	MAE
L	x_i	0	0	-	-	-
	$f(x_i)$			-		
R	x_i	$\pi/2$	π	2π	-	-
	$f(x_i)$					

Threshold $t = 3\pi/4$

Child		1	2	3	Median	MAE
L	x_i	0	0	$\pi/2$	-	-
	$f(x_i)$					
R	x_i	π	2π	-	-	-
	$f(x_i)$			-		

Threshold $t = 3\pi/2$

Child		1	2	3	4	Median	MAE
L	x_i	0	0	$\pi/2$	π	-	-
	$f(x_i)$						
R	x_i	2π	-	-	-	-	-
	$f(x_i)$		-	-	-		

Threshold	Sum S
$\pi/4$	
$3\pi/4$	
$3\pi/2$	

Sample B

Threshold $t = \pi/2$

Child		1	2	3	4	Median	MAE
L	x_i	0	-	-	-	-	-
	$f(x_i)$		-	-	-		
R	x_i	π	$3\pi/2$	2π	2π	-	-
	$f(x_i)$						

Threshold $t = 5\pi/4$

Child		1	2	3		Median	MAE
L	x_i	0	π	-		-	-
	$f(x_i)$			-			
R	x_i	$3\pi/2$	2π	2π		-	-
	$f(x_i)$						

Threshold $t = 7\pi/4$

Child		1	2	3		Median	MAE
L	x_i	0	π	$3\pi/2$		-	-
	$f(x_i)$						
R	x_i	2π	2π	-		-	-
	$f(x_i)$			-			

Threshold	Sum S
-----------	---------

$\pi/2$

$5\pi/4$

$7\pi/4$

Sample C

Threshold $t = \pi/2$

Child		1	2	3		Median	MAE
L	x_i	0	0	-		-	-
	$f(x_i)$			-			
R	x_i	π	π	2π		-	-
	$f(x_i)$						

Threshold $t = 3\pi/2$

Child		1	2	3	4	Median	MAE
L	x_i	0	0	π	π	-	-
	$f(x_i)$						
R	x_i	2π	-	-	-	-	-
	$f(x_i)$		-	-	-		
Threshold		Sum S					
$\pi/2$							
$3\pi/2$							

Exercise 4.4 - Combine the predictions

A **forest** combines predictions from several trees. Use their arithmetic mean:

$$\hat{f}_{\text{forest}}(x) = \frac{\hat{f}_A(x) + \hat{f}_B(x) + \hat{f}_C(x)}{3}$$

The leaf predictions are medians; the forest prediction is the mean of the three tree predictions.

1. Complete the table at the five original input positions.

x	Target $f(x)$	Tree A	Tree B	Tree C	Forest	Absolute forest error
0						
$\pi/2$						
π						
$3\pi/2$						
2π						

2. Write the forest prediction as a piecewise-constant function. Include every threshold used by any tree.
3. Sketch the forest prediction.
4. Calculate the forest's MAE on the original five observations, counting each once. Compare it with the tree from Exercise 4.2.

In a random forest with several input features, a random subset of features can also be considered at each split. Here there is only one feature, so the variation between trees comes from the bootstrap samples.

Short conceptual questions

1. What value does a regression-tree leaf predict when squared error is used?
2. Why is that value the mean?
3. What quantity is minimized when choosing a split?
4. Where do the candidate thresholds come from for a continuous feature?
5. Why is the prediction of a regression tree piecewise constant?
6. What is the main effect of increasing tree depth?
7. Why can a deep tree overfit noisy data?
8. In a multidimensional input space, what additional choice must the tree make at every split?
9. What shape do the regions in the two-dimensional example have when only axis-aligned splits are used?
10. Why can bootstrap samples produce different trees from the same original dataset, and how do repeated observations affect the calculation?
11. Does averaging the predictions of several regression trees make the forest prediction continuous? Explain.
12. Does a smaller training error guarantee better predictions between the training points? Give a short reason.

Key concepts

Concept	Meaning in this exercise
Regression	Predicting a numerical target value.
Observation / feature / target	One data point / an input value / the value to be predicted.
Training data	Observations used to choose splits and leaf predictions.
Regression tree	A sequence of splits dividing the input space into regions with constant predictions.
Node / root node	A stage containing observations / the initial node containing all training observations.
Parent / child node	A node before a split / one of the nodes produced by that split.
Internal node / leaf	A node that splits / a final node assigning a constant prediction.
Split / threshold	A division such as $x < t$ versus $x \geq t$ / its boundary t .
Candidate split	A feature-threshold choice tested during training; both children must be non-empty.
Residual	Signed prediction error $r_i = y_i - \hat{y}_i$.
Loss	A numerical measure of prediction error.
SSE	Sum of squared residuals.
MSE / RMSE	Mean squared residual / its square root, in target units.
MAE	Mean absolute residual.
Mean / median	Arithmetic average / middle of the sorted values, conventionally averaging the middle pair for even sample size.
Error reduction	Parent SSE minus the sum of child SSEs.
Tree depth	Maximum number of splits on a root-to-leaf path.
Piecewise-constant prediction	A prediction constant inside each leaf region.
Training error	Prediction error on the observations used to build the model.
Noise / overfitting	Random target variation / fitting sample-specific details in a way that harms prediction on new data.
Feature space	The space of possible combinations of input features.
Axis-aligned split	A split testing only one feature.
Bootstrap sample	A sample drawn with replacement; observations may repeat or be absent. Here each sample has five draws.
Forest	A collection of trees whose predictions are combined; here by their arithmetic mean.
Total absolute split error	$S = N_L \text{MAE}_L + N_R \text{MAE}_R$, counting each observation in its child.