Quiz Week 3: Tree-based methods

Willow Liu
29/01/2025

Q1: Classification and regression trees (CART) have hyper-parameters. Which of the following statements are correct?

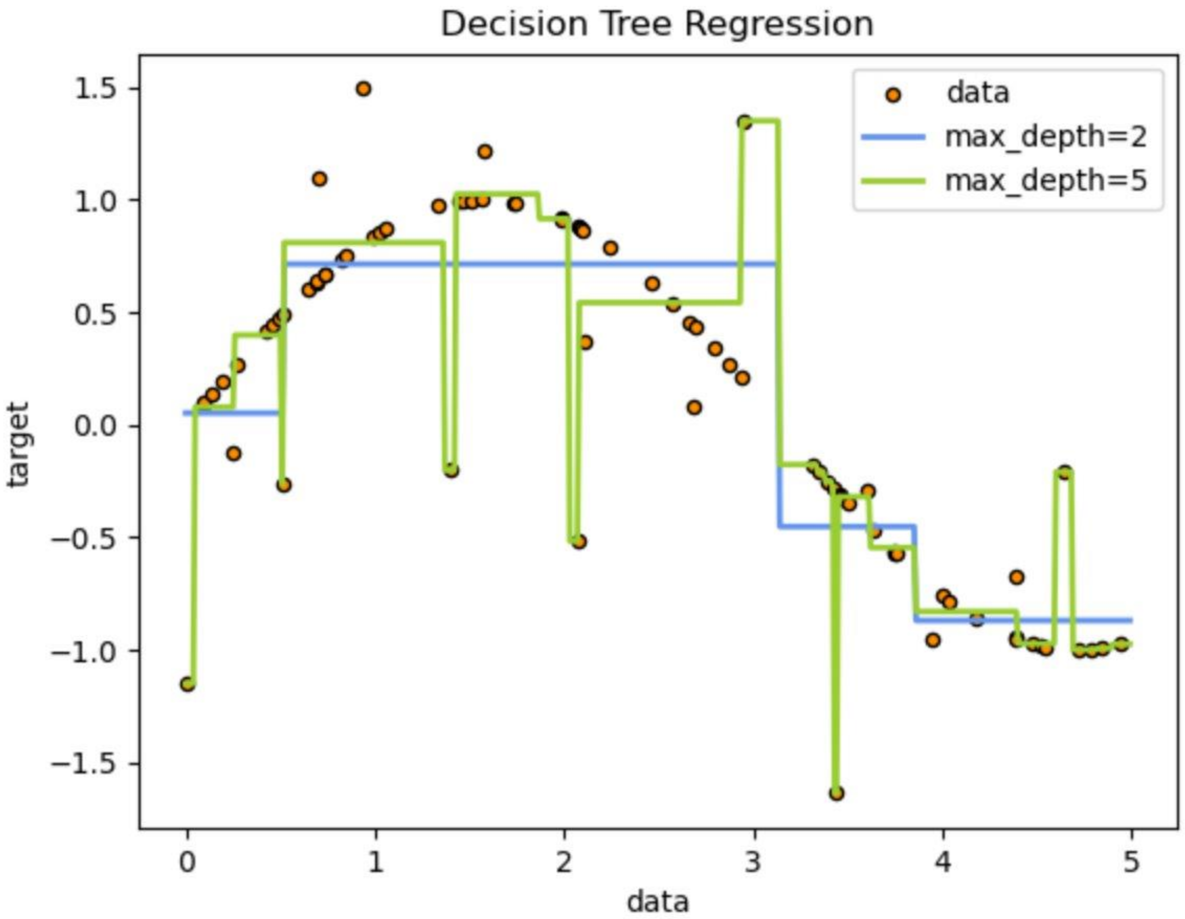
a. CART's hyper-parameters represent a trade-off between performance and overfitting and are user-defined, though can be tuned by cross-validation

b. CART's hyper-parameters include the minimal depth of the tree and the maximal number of records on a node

c. CART's hyper-parameters include the maximal depth of the tree and the minimal number of records on a node

d. The values of the hyper-parameters are inferred from the data via the learning process (training)

a. CART's hyper-parameters represent a trade-off between performance and overfitting and are user-defined, though can be tuned by cross-validation



With greater max_depth, we can model more regions of the plane and increase the model’s complexity

c. CART's hyper-parameters include the maximal depth of the tree and the minimal number of records on a node

The hyperparameters of a decision tree are:

max_depth: The maximum depth of the tree.

min_samples_per_leaf: The minimum number of samples required to be at a leaf node.

min_samples_split: The minimum number of samples required to split an internal node.

max_leaf_nodes: Limits the total number of leaf nodes in the tree.

min_impurity_decrease: A node will be split if this split induces a decrease of the impurity greater than or equal to this value.

https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Q2: Which of the following statements are correct about classification and regression trees (CART)?

a. One advantage of CART is smoothness: small perturbations in the input data do not dramatically change the response

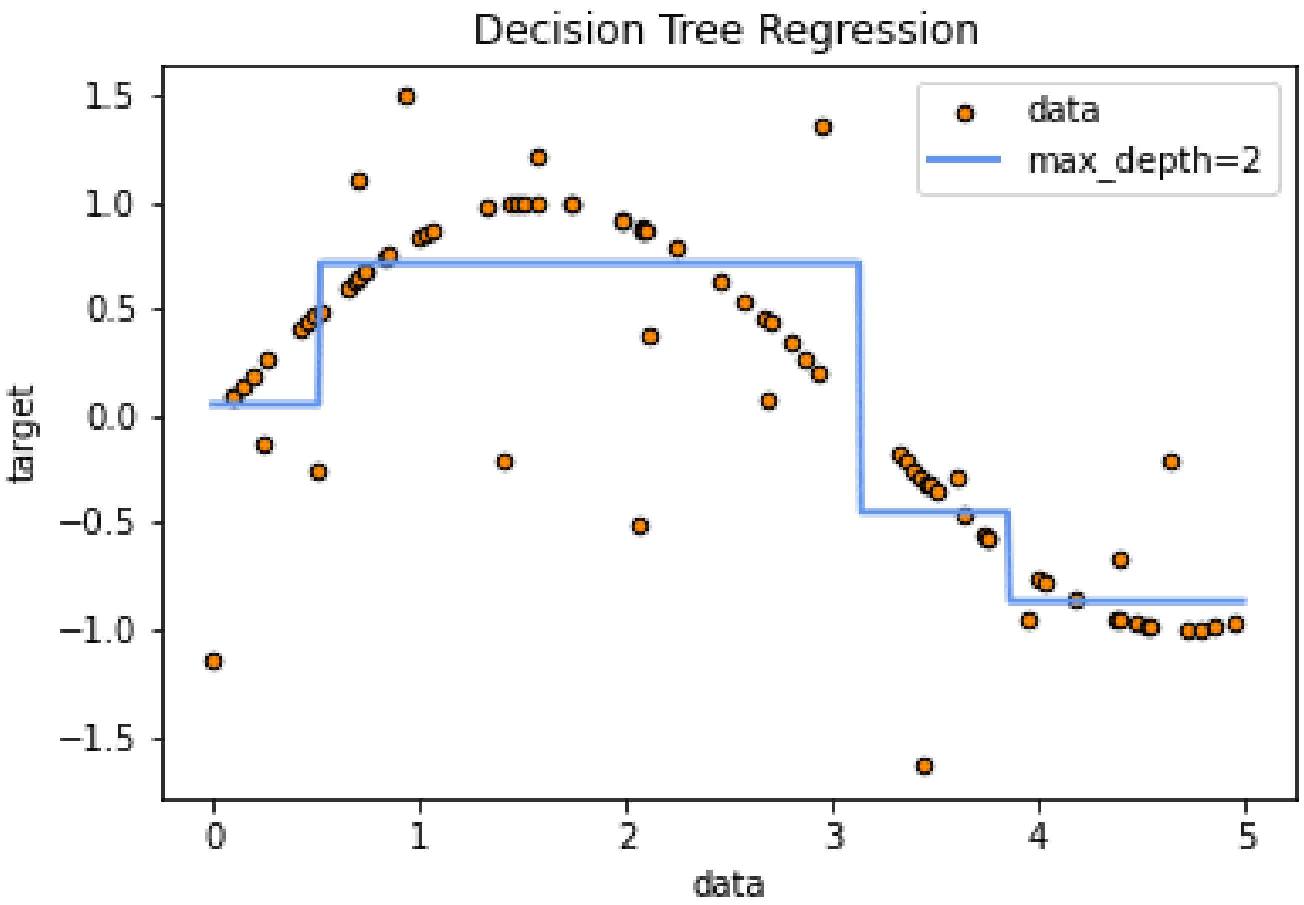
b. One advantage of CART is interpretability: it is easy to understand which features learnt generated the predictions

c. One advantage of CART is flexibility: no assumptions of data distribution and no transformations needed

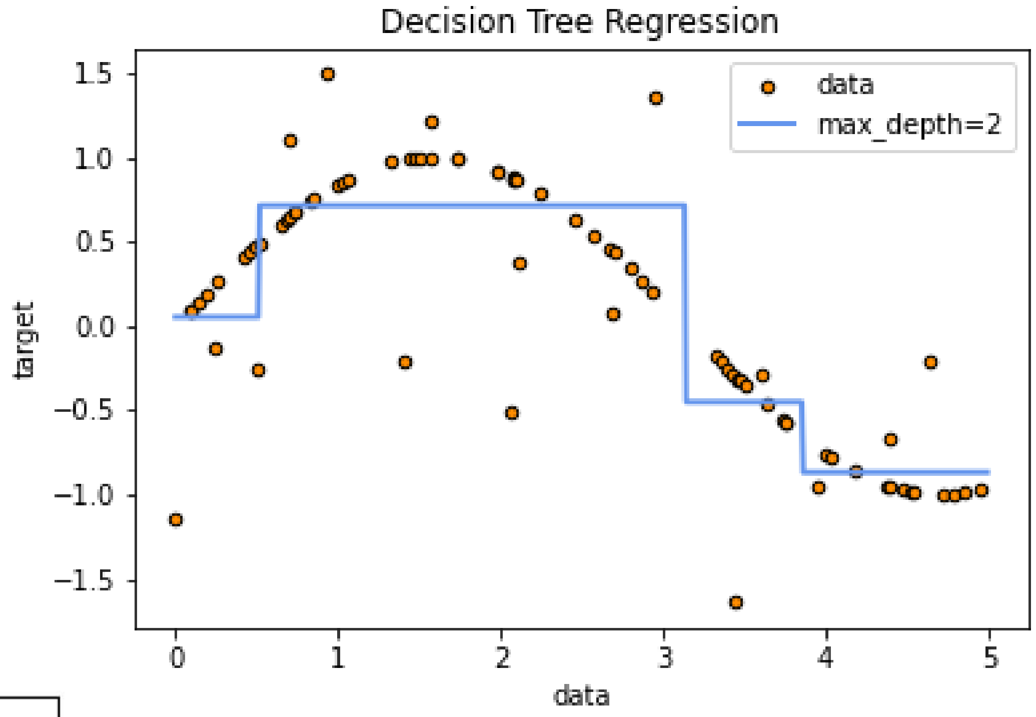
d. One disadvantage of CART is overfitting: they do not easily generalise to new unseen data

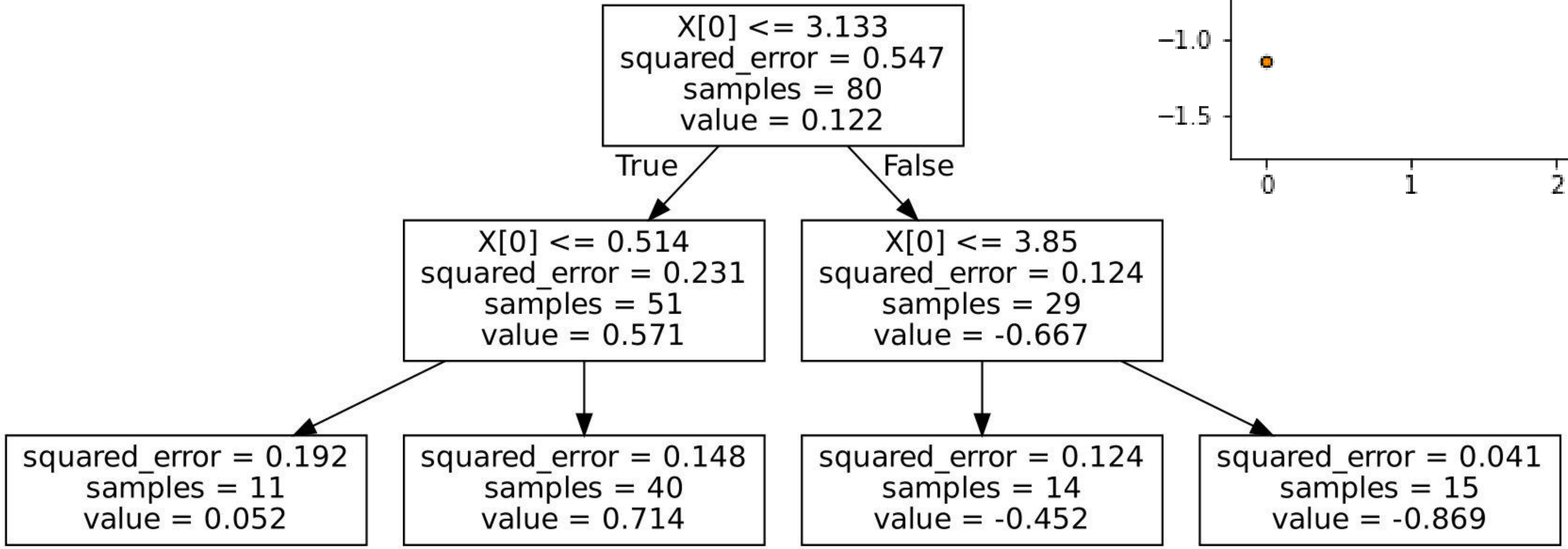
c. One *advantage* of CART is *flexibility*: no assumptions of data distribution and no transformations needed -> True

here are fitting a non-linear function using CART



How can we visualize decision tree?


Decision Tree Regression


X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

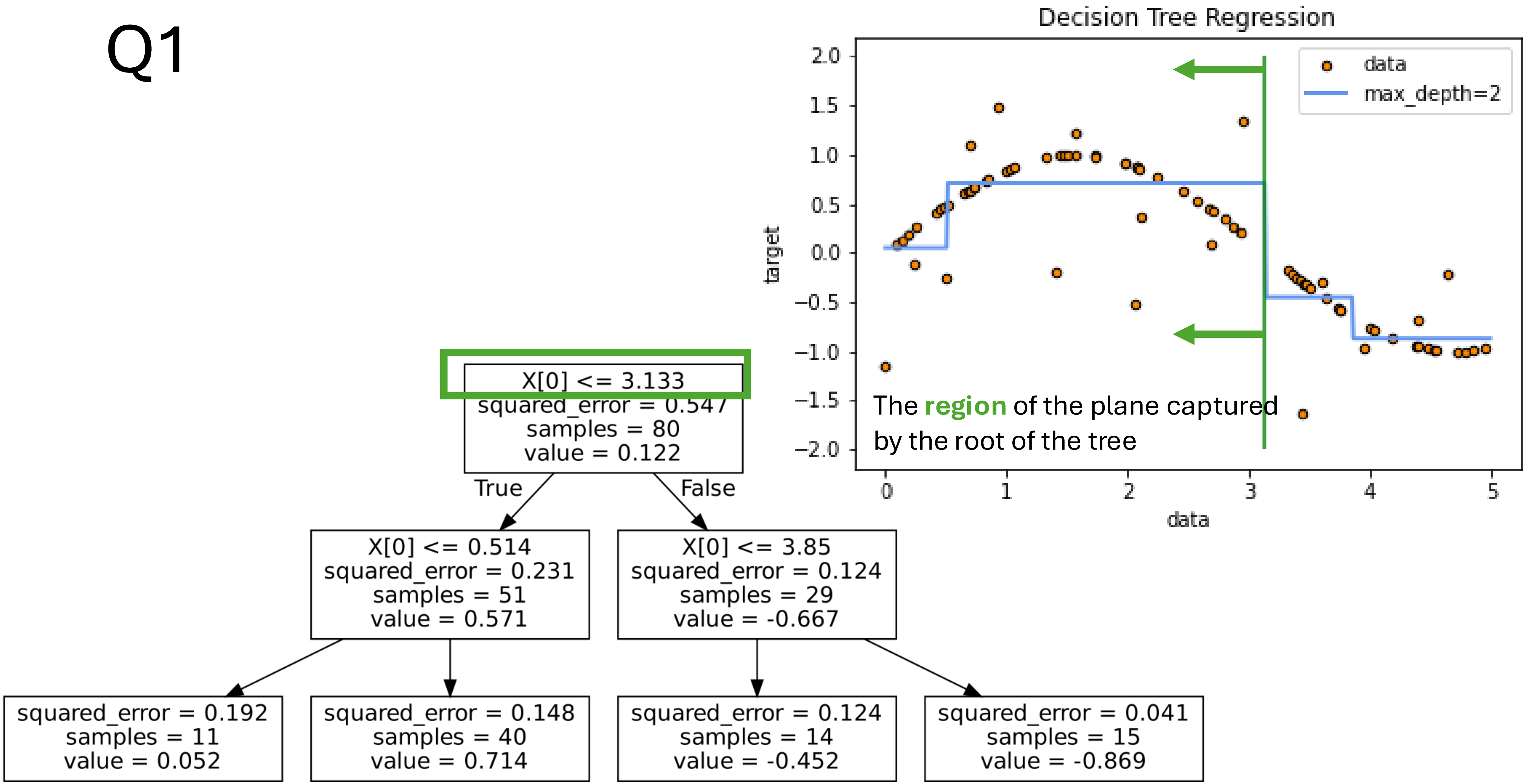
squared_error = 0.192
samples = 11
value = 0.052

squared_error = 0.148
samples = 40
value = 0.714

squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q1



X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

squared_error = 0.148
samples = 40
value = 0.714

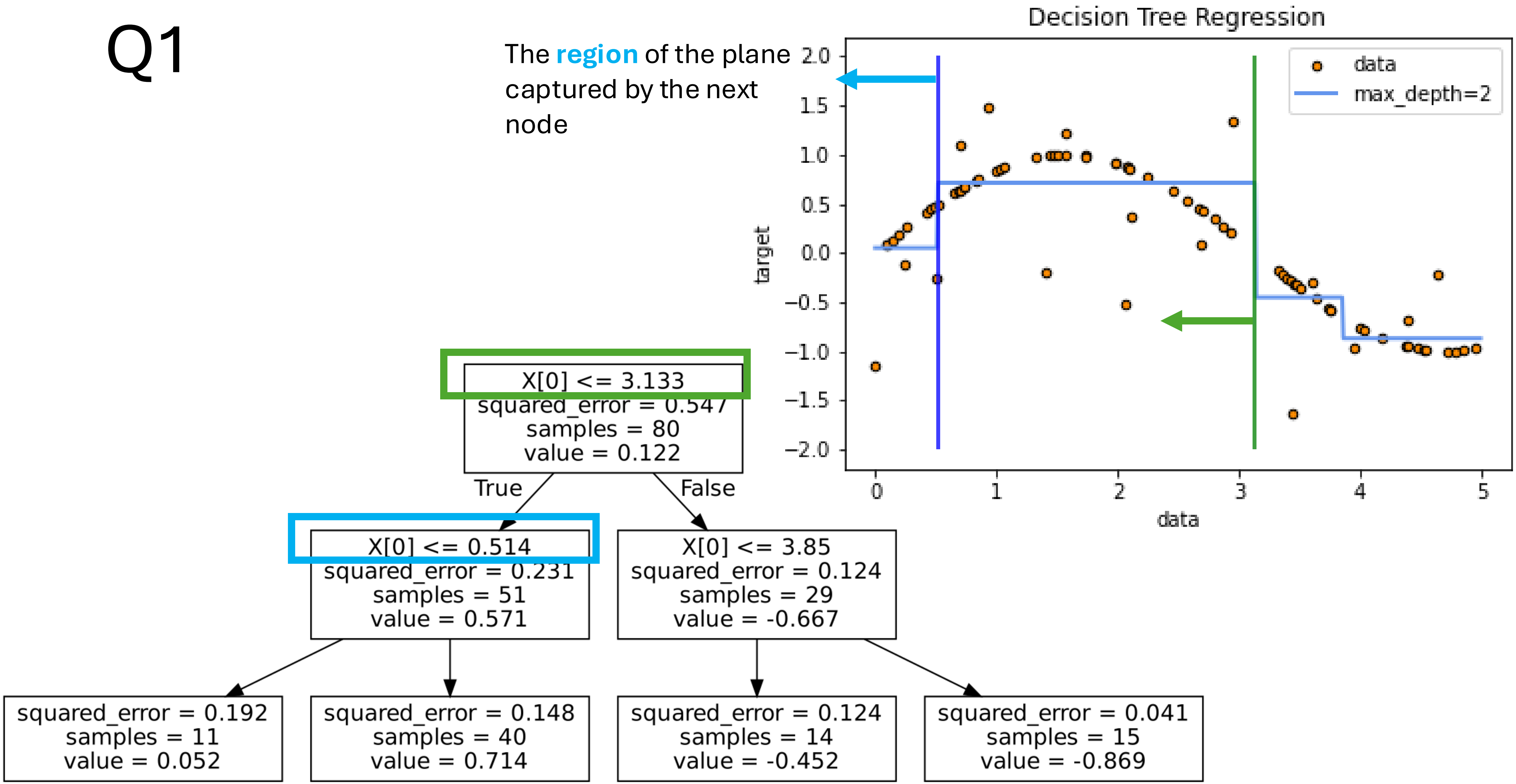
squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Decision Tree Regression

The **region** of the plane captured by the root of the tree

Q1



The region of the plane captured by the next node

X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

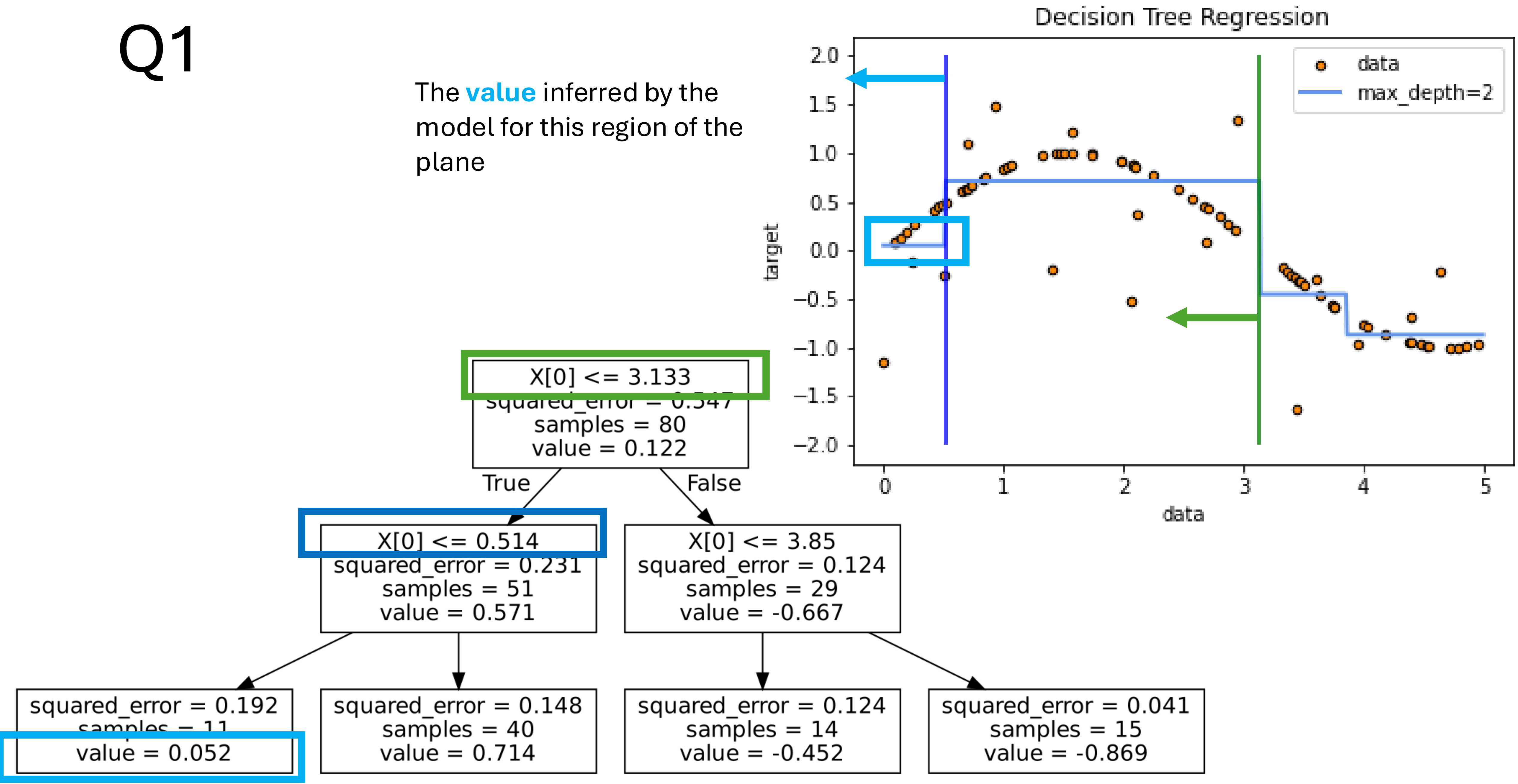
squared_error = 0.148
samples = 40
value = 0.714

squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q1

The **value** inferred by the model for this region of the plane



X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

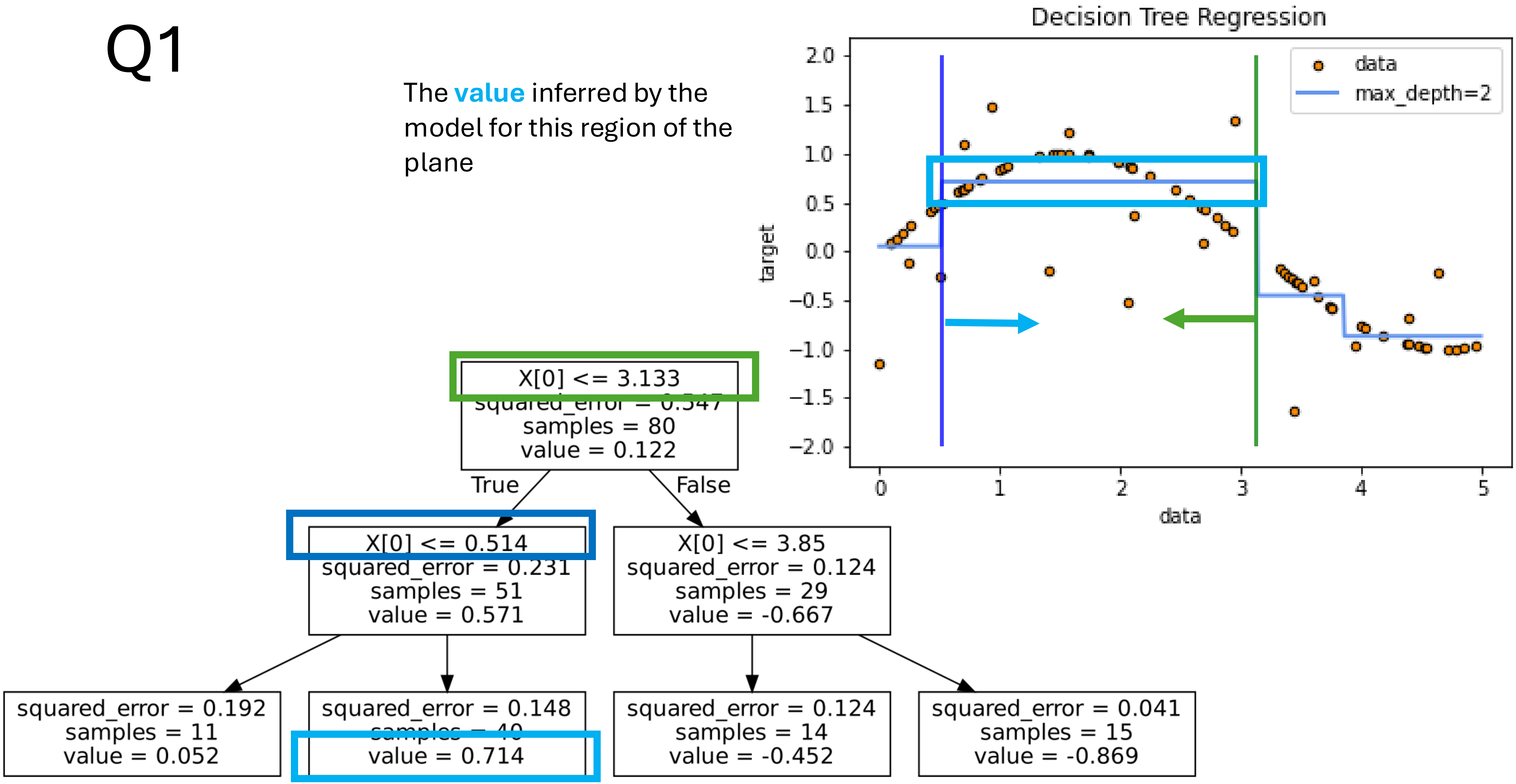
squared_error = 0.192
samples = 11
value = 0.052

squared_error = 0.148
samples = 40
value = 0.714

squared_error = 0.124
samples = 14
value = -0.452

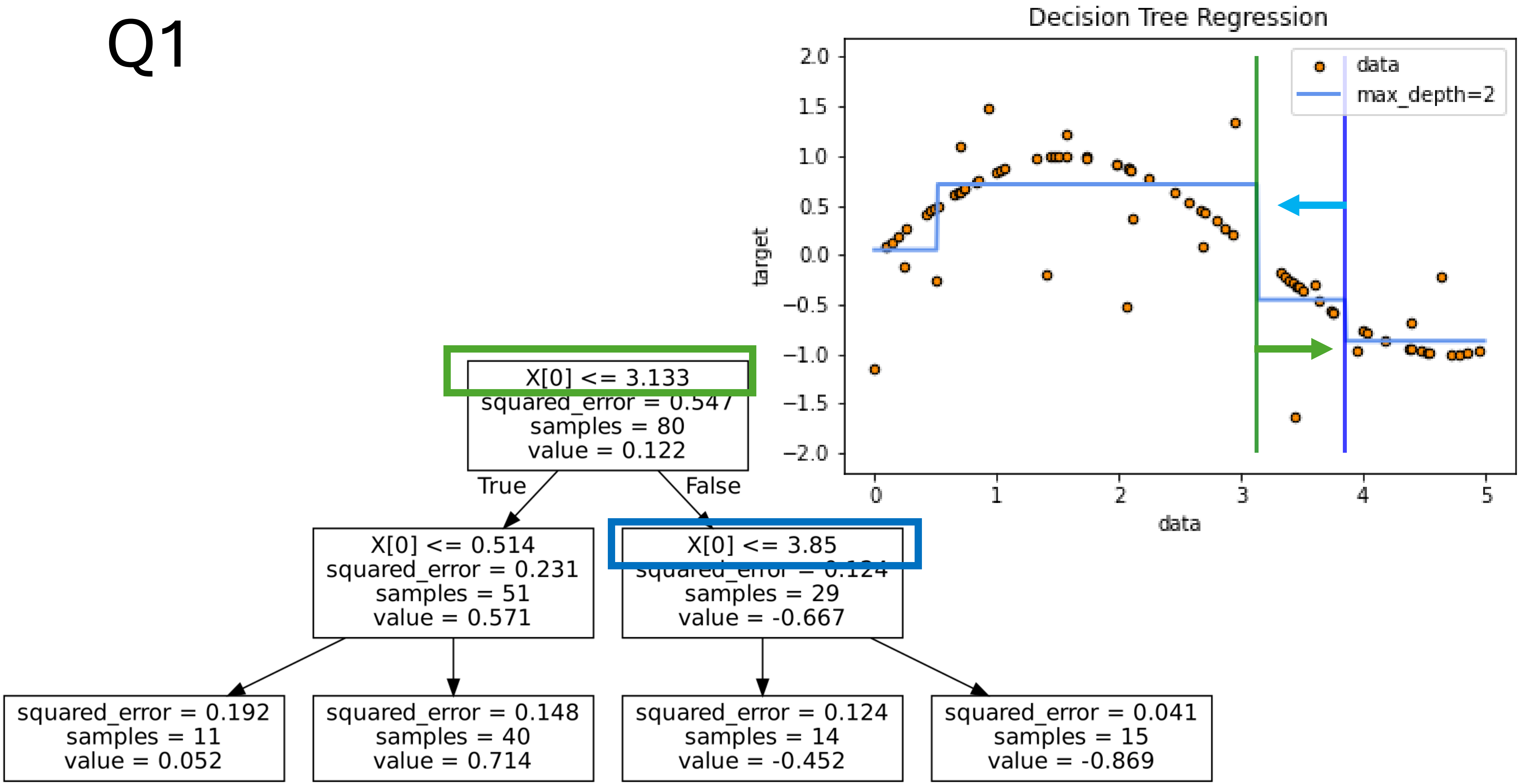
squared_error = 0.041
samples = 15
value = -0.869

Q1



The value inferred by the model for this region of the plane

Q1



X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

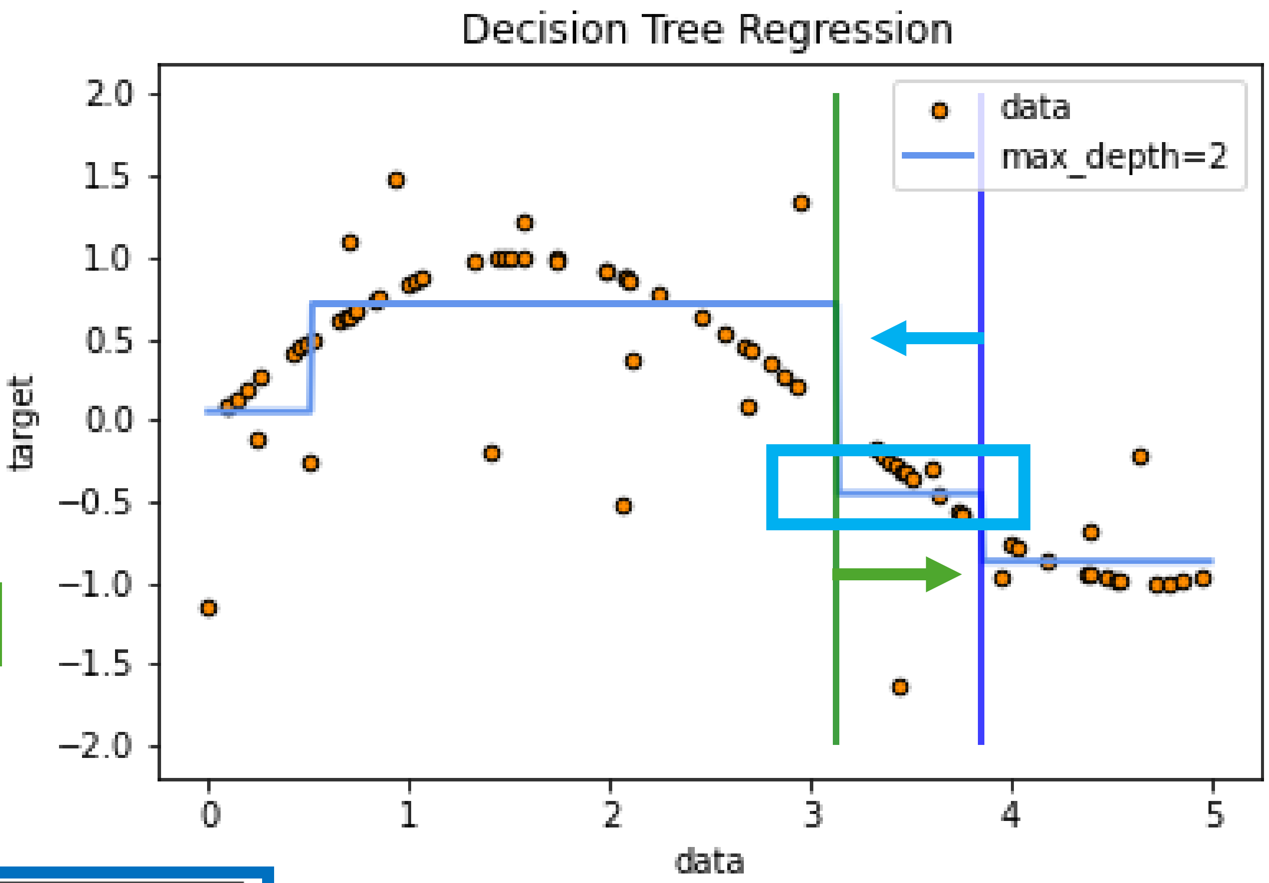
squared_error = 0.148
samples = 40
value = 0.714

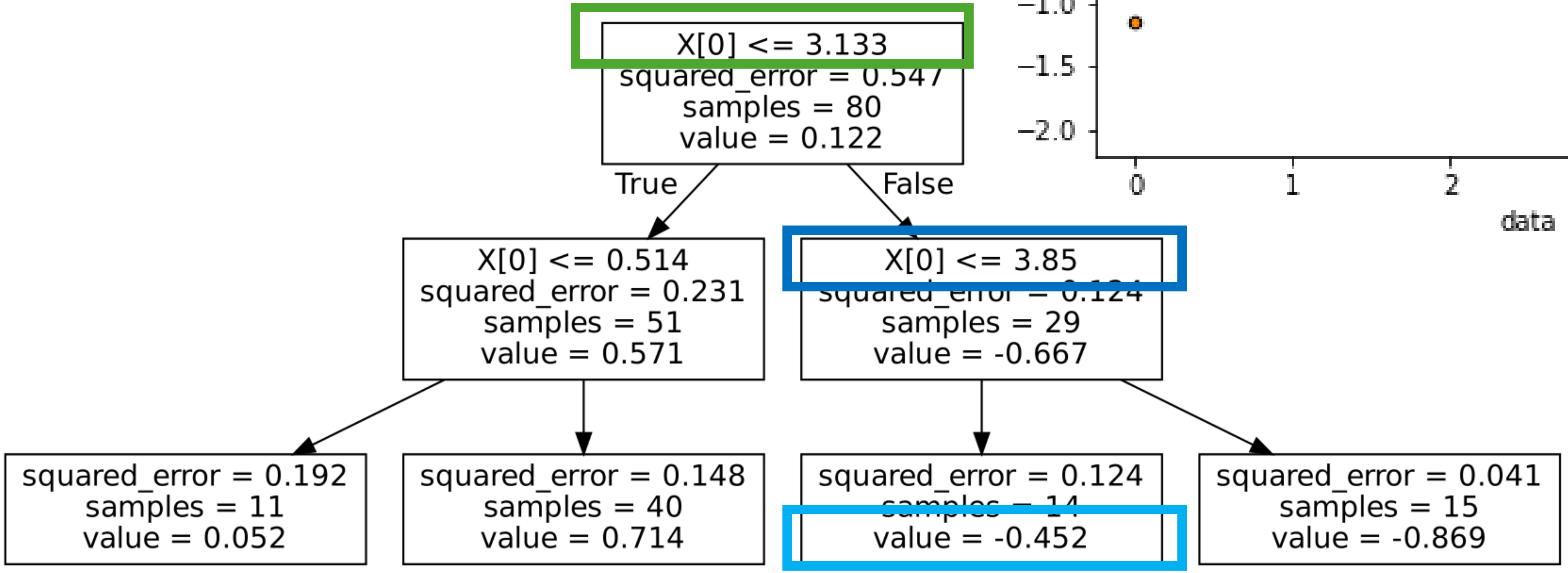
squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q1

The **value** inferred by the model for this region of the plane





X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

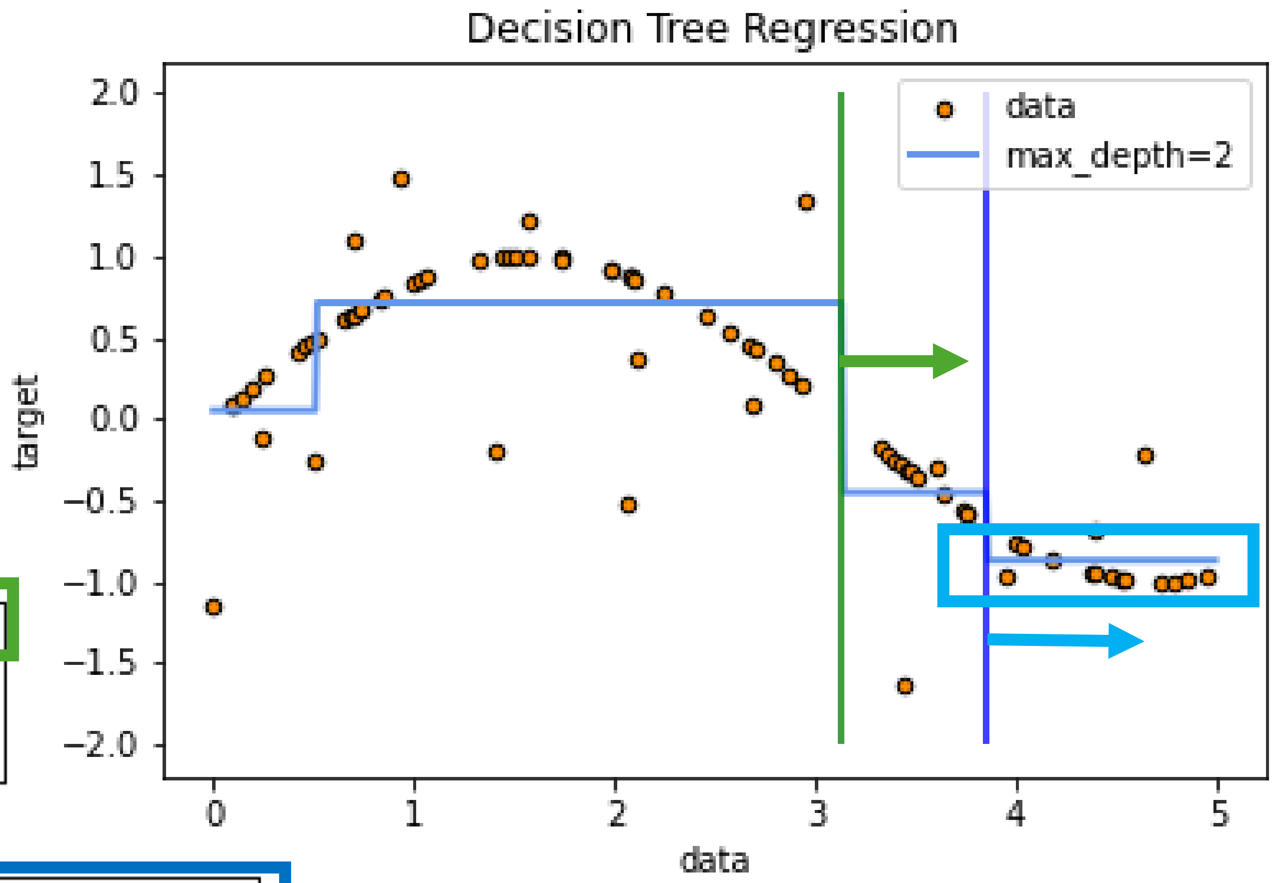
squared_error = 0.148
samples = 40
value = 0.714

squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q1

The **value** inferred by the model for this region of the plane



X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

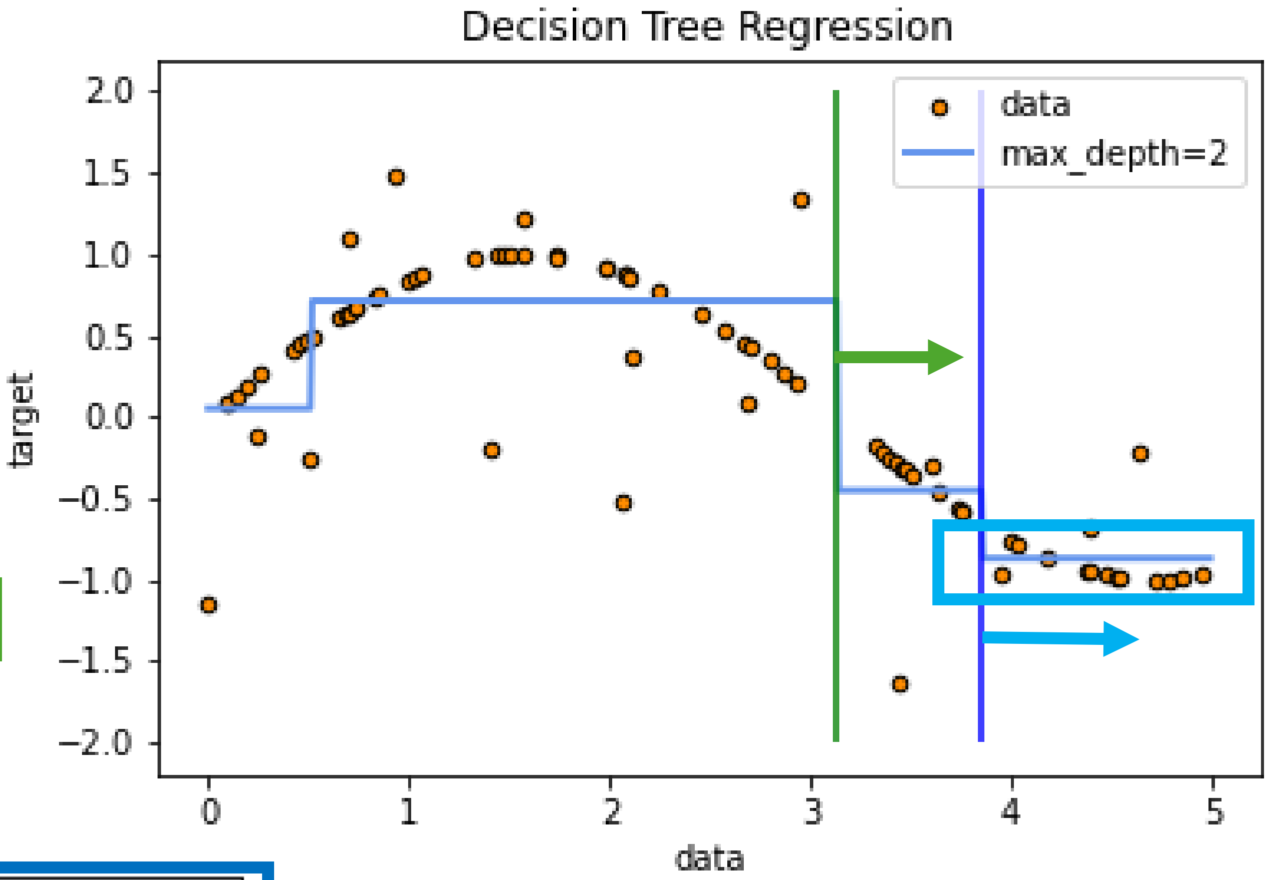
squared_error = 0.148
samples = 40
value = 0.714

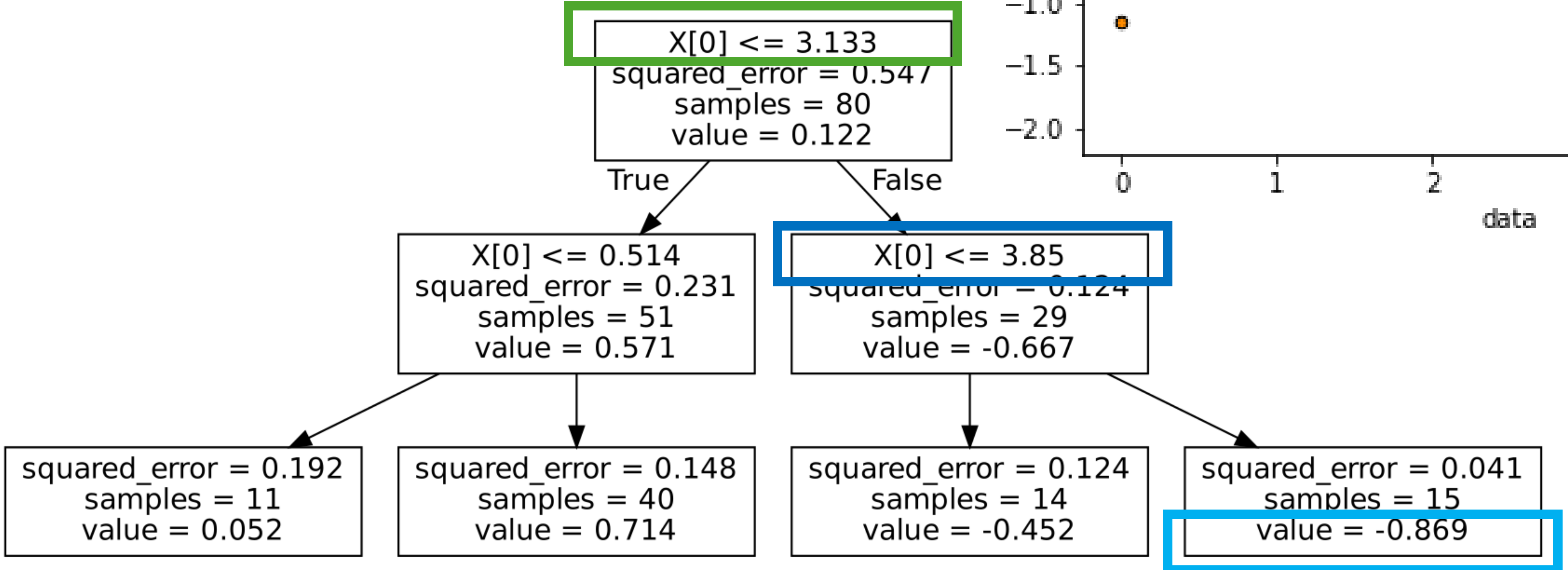
squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q1

2. One *advantage* of CART is *interpretability*: it is easy to understand which features learnt generated the predictions -> True


Decision Tree Regression



X[0] <= 3.133
squared_error = 0.547
samples = 80
value = 0.122

True

False

X[0] <= 0.514
squared_error = 0.231
samples = 51
value = 0.571

X[0] <= 3.85
squared_error = 0.124
samples = 29
value = -0.667

squared_error = 0.192
samples = 11
value = 0.052

squared_error = 0.148
samples = 40
value = 0.714

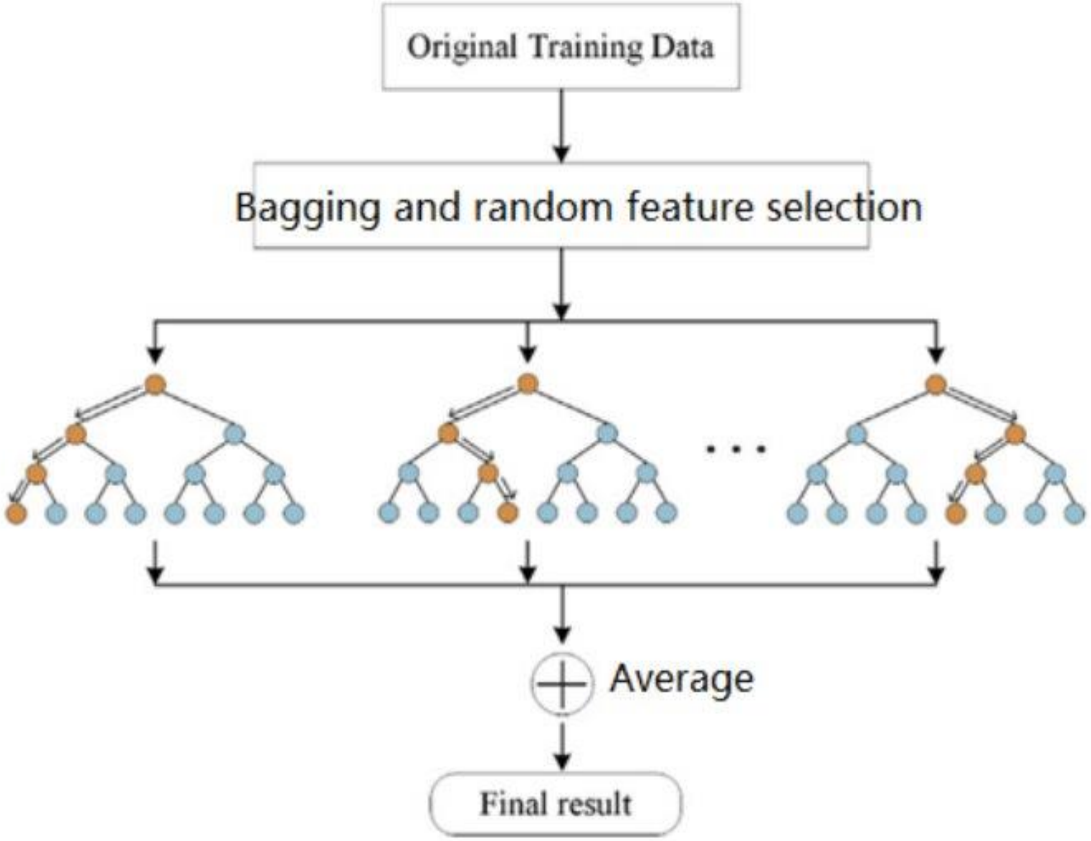
squared_error = 0.124
samples = 14
value = -0.452

squared_error = 0.041
samples = 15
value = -0.869

Q3: CART are usually used as the base predictors of random forest (RF). Which of the following are correct?

a. RF constructs an ens

✓ Random Forest bu nd in parallel, which makes it computationally effic



b. RF repeatedly samp producing different trees

✓ Random Forest use vith replacement) for data points and selects a **ra** ig to diverse trees that reduce overfitting.

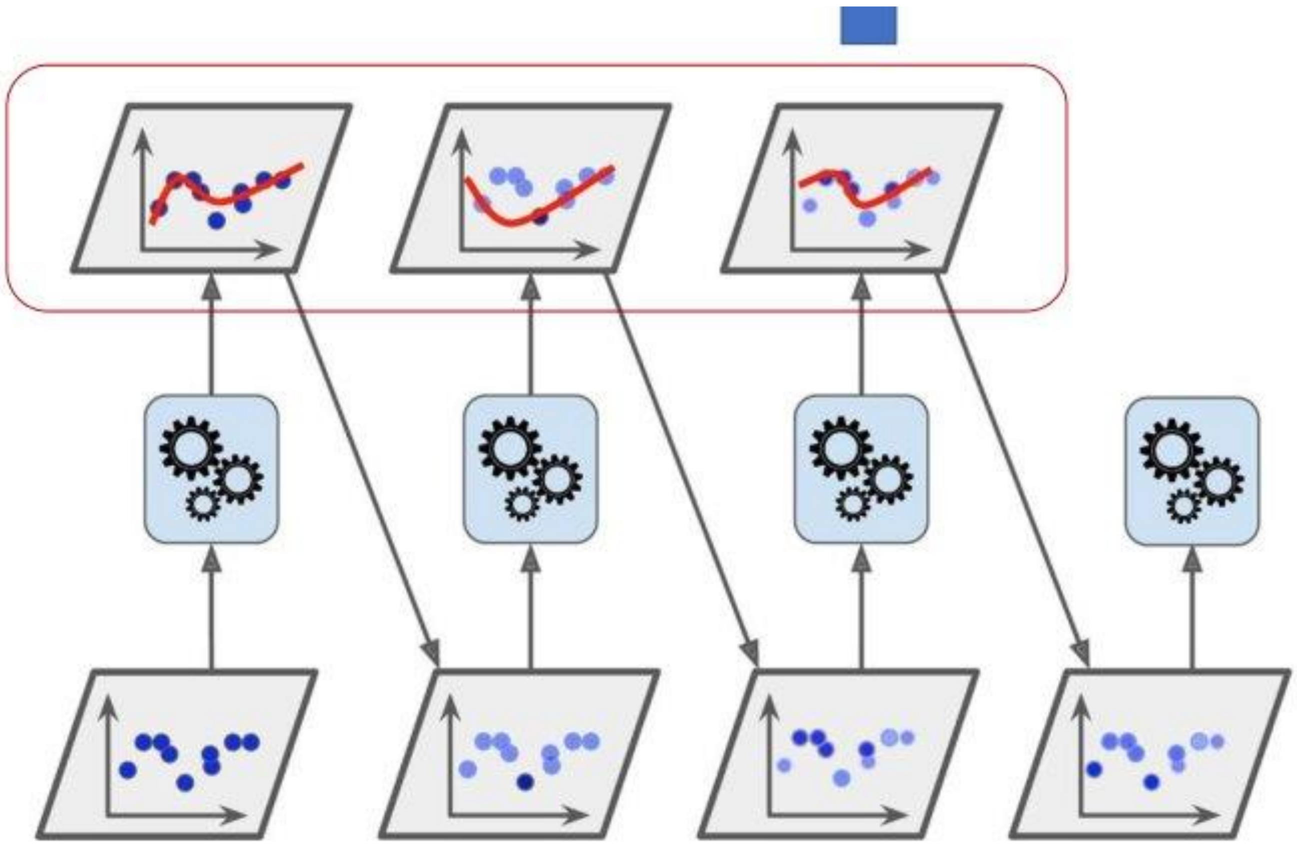
c. RF can be used for b

✓ Random Forest is e.g., predicting categories like spam or not spam) s value like house price).

Q4: CART are usually used as the base predictors of gradient-boosted decision trees (GBDT). Which of the following are correct?

a. GBDT constructs an ensemble of trees in parallel
b. During GBDT's fitting, a new CART predictor is trained using the residual from the last CART as the weight, considering the largest residuals
c. GBDT has been used successfully in many data science competitions
d. XGBoost is one efficient and scalable implementation of GBDT

- GBDT build trees **sequentially**, not in parallel. Each tree is trained to minimize the residual errors of the previous trees, which is why boosting is inherently sequential.



Q5: We have a dataset at a decision tree node with the following class distribution:

- Class A: 40 samples
- Class B: 30 samples
- Class C: 30 samples

Calculate the Gini Impurity for this node.

a. 0.64

b. **0.66**

c. 0.44

d. 0.36

Gini Impurity:

$$I_G(p) = \sum_{i=1}^{J} p_i(1 - p_i)$$

Gini Impurity is a measure used to evaluate the quality of a split in decision trees, particularly in the CART (Classification and Regression Trees)

A Gini Impurity of 0.66 indicates that the node is quite impure, with no dominant class.

	Class A	**Class B**	**Class C**
Samples	40	30	30
Probability	0.4	0.3	0.3
Gini Impurity	$0.4(1-0.4) + 0.3 * (1-0.3) + 0.3 * (1-0.3)$ =0.66		