Introduction to machine learning in Hydrology

Lazaro J. Perez & Marc Berghouse

contact: lazaro.perez@dri.edu

Outline

- Decision Trees
- Decision Trees in Matlab
- Classification and Regression Learner App

Decision trees

Definition

A decision tree is a non-parametric supervised learning algorithm that is utilized for both classification and regression tasks

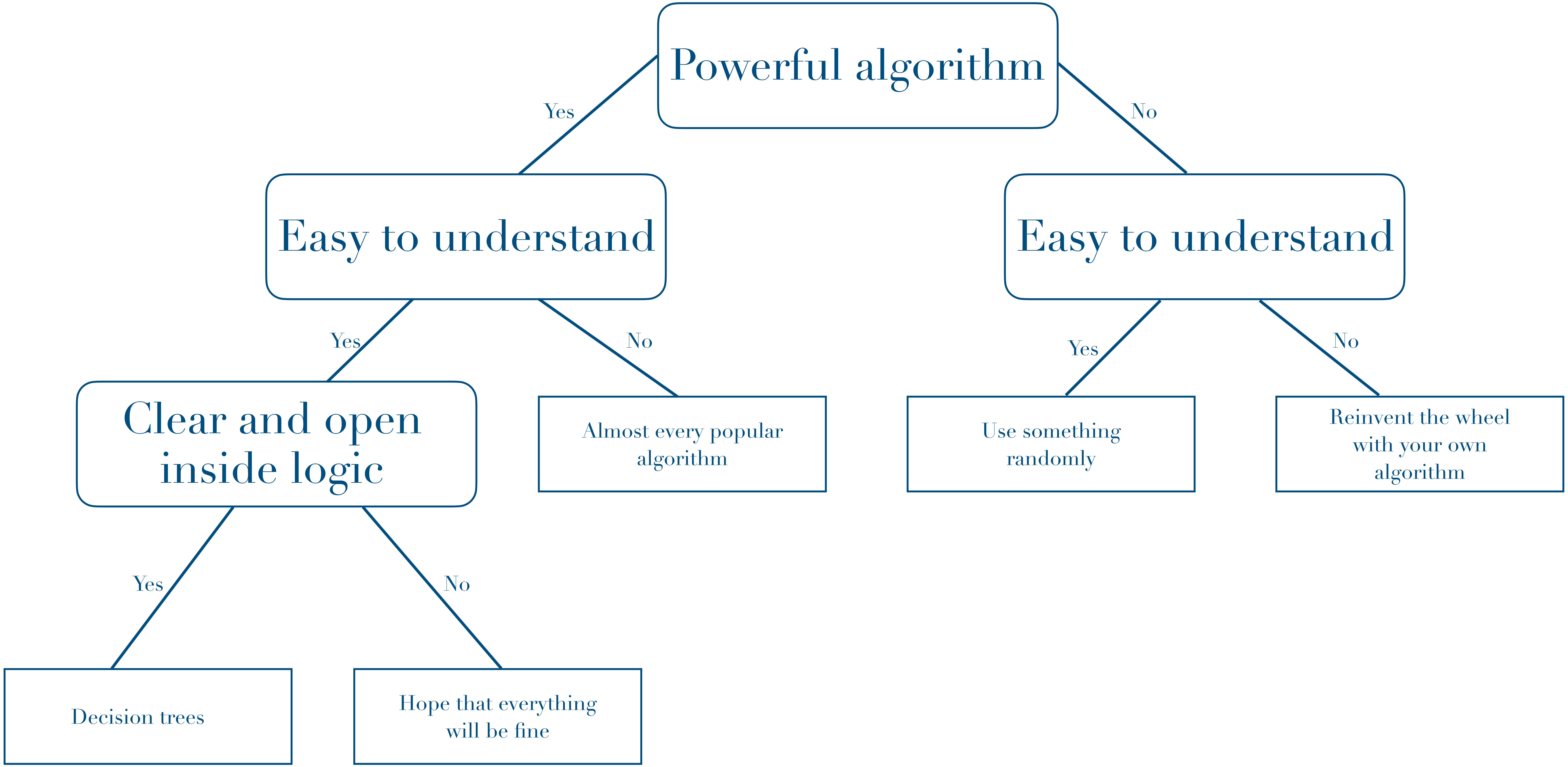
It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes

General consensus

It’s a binary tree that recursively splits the dataset until we’re left with pure leaf nodes

A decision tree analysis is a divide-and-conquer approach to classification and regression

Decision trees: Graphical representation

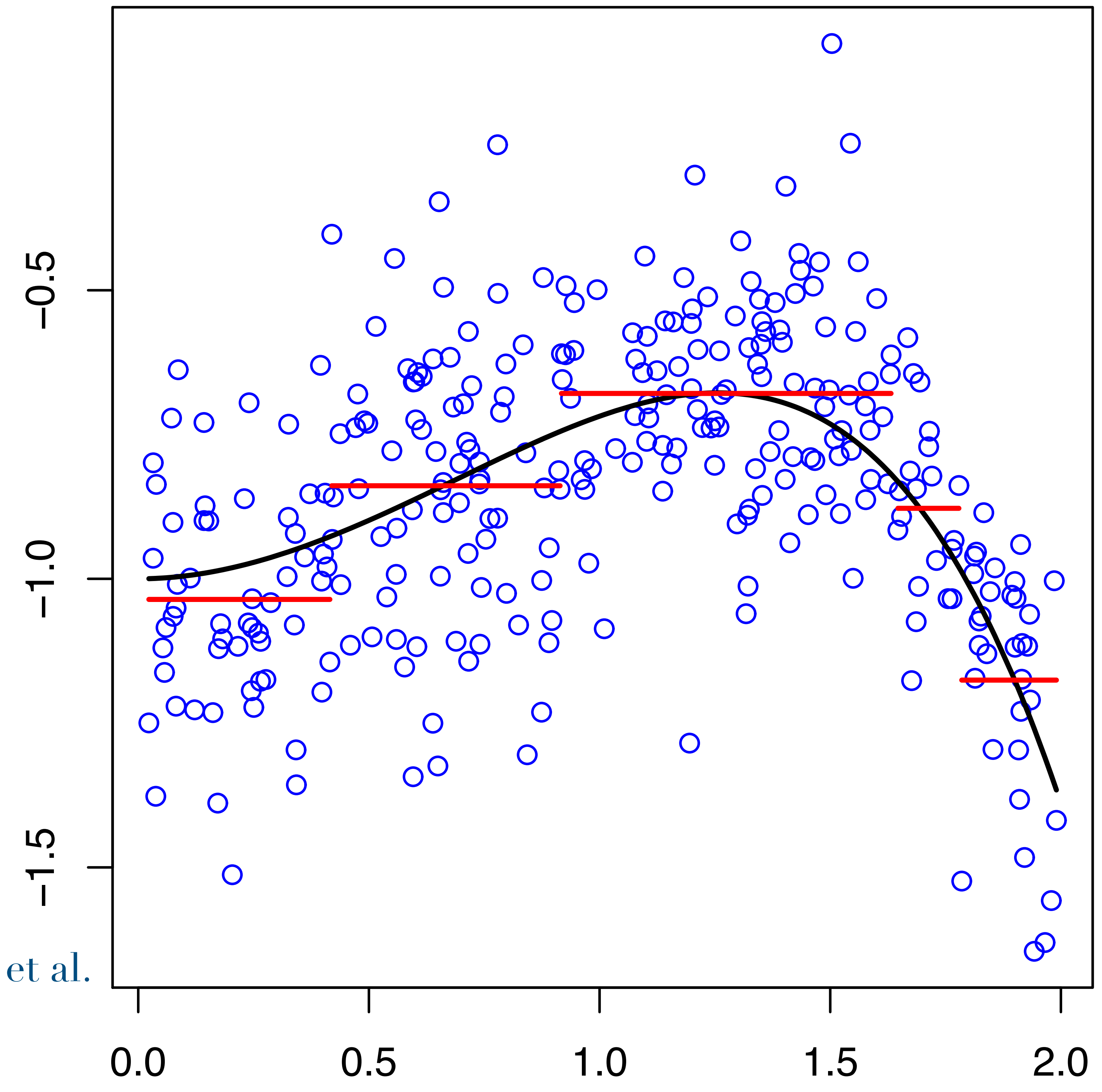

Powerful algorithm
Yes
No
Easy to understand
Easy to understand
Yes
No
Yes
No
Clear and open inside logic
Almost every popular algorithm
Use something randomly
Reinvent the wheel with your own algorithm
Yes
No
Decision trees
Hope that everything will be fine

Decision trees

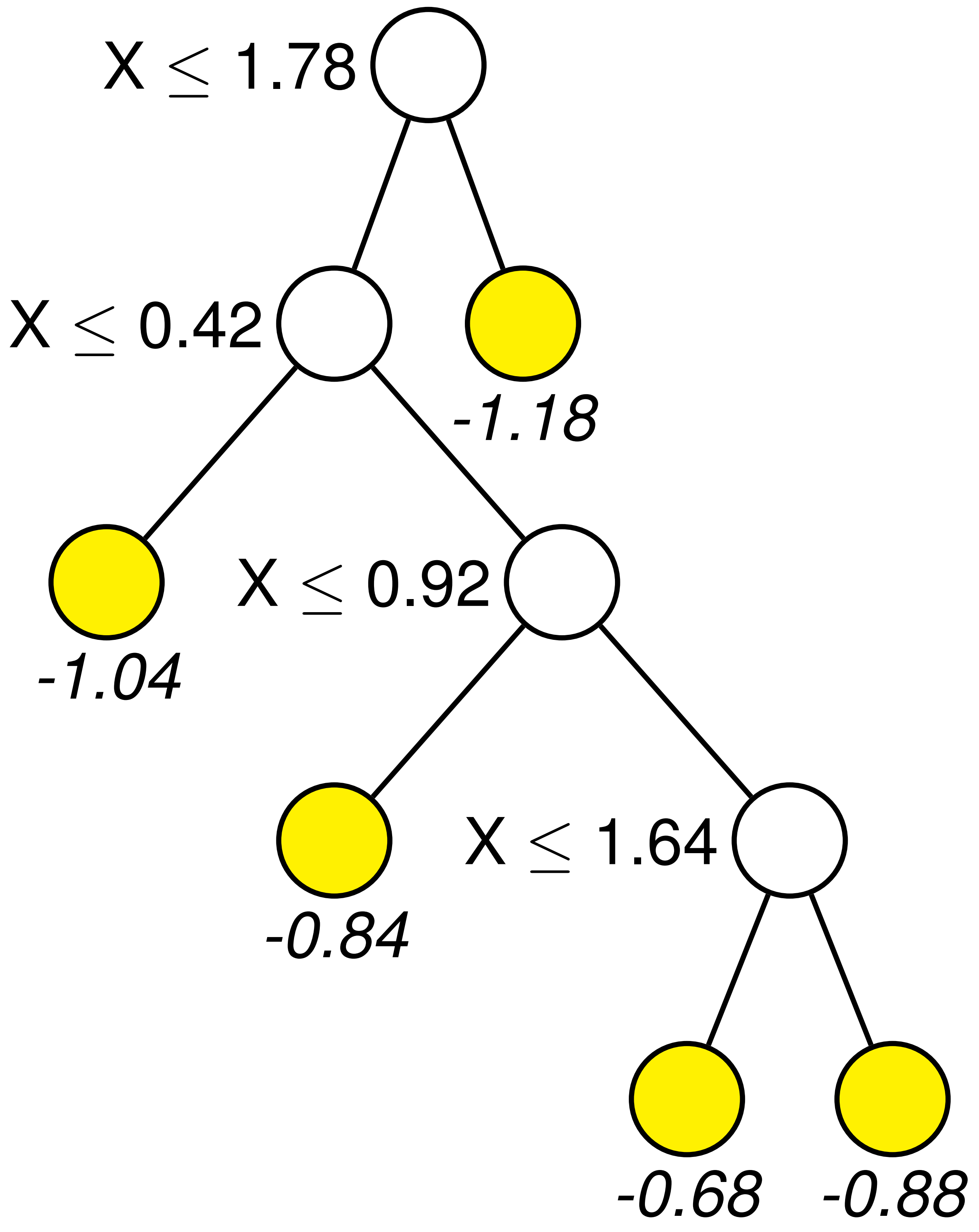
Development

First generation

Automatic interaction detection (AID): Morgan and Sonquist, 1963



Loh et al.

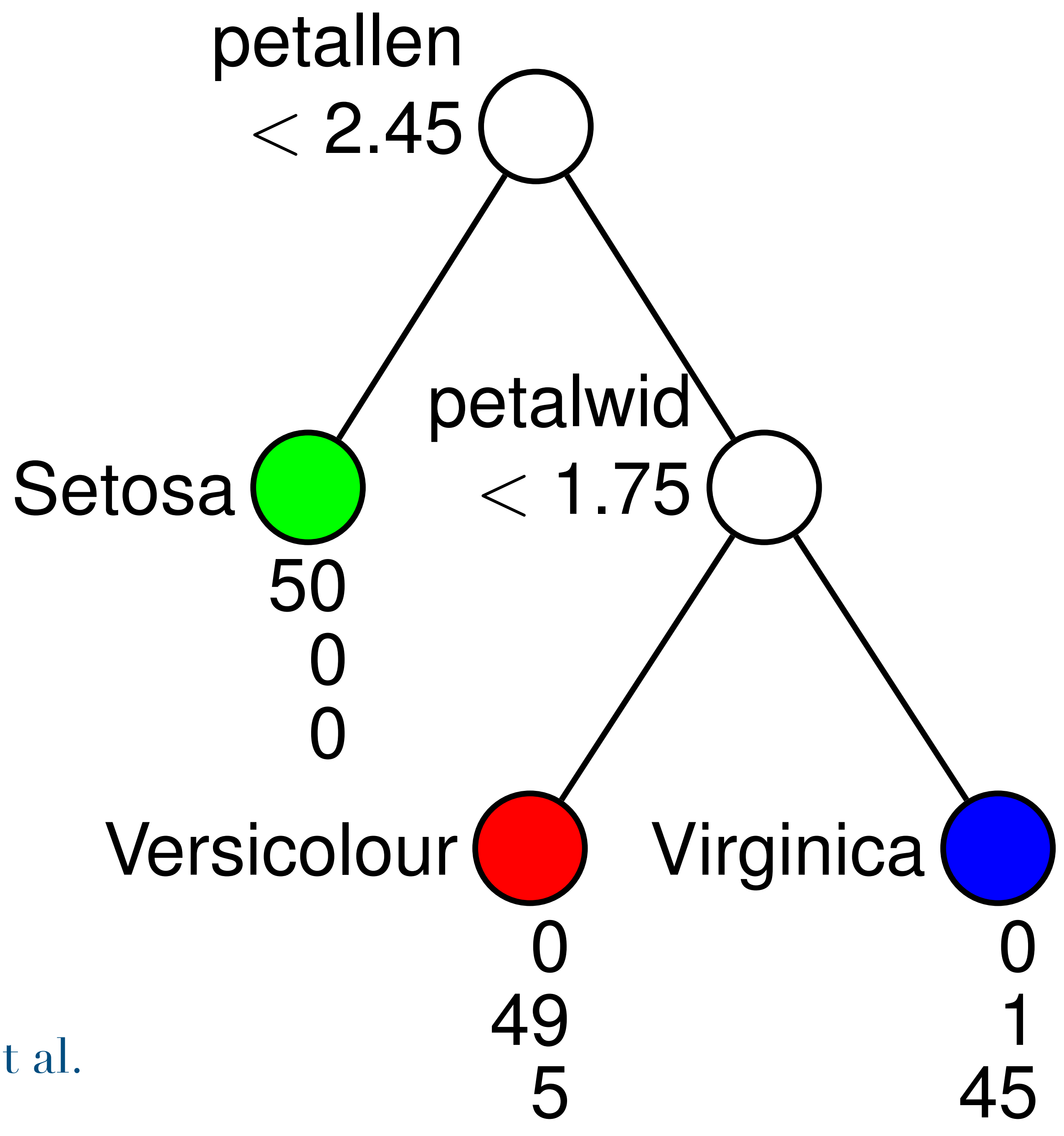


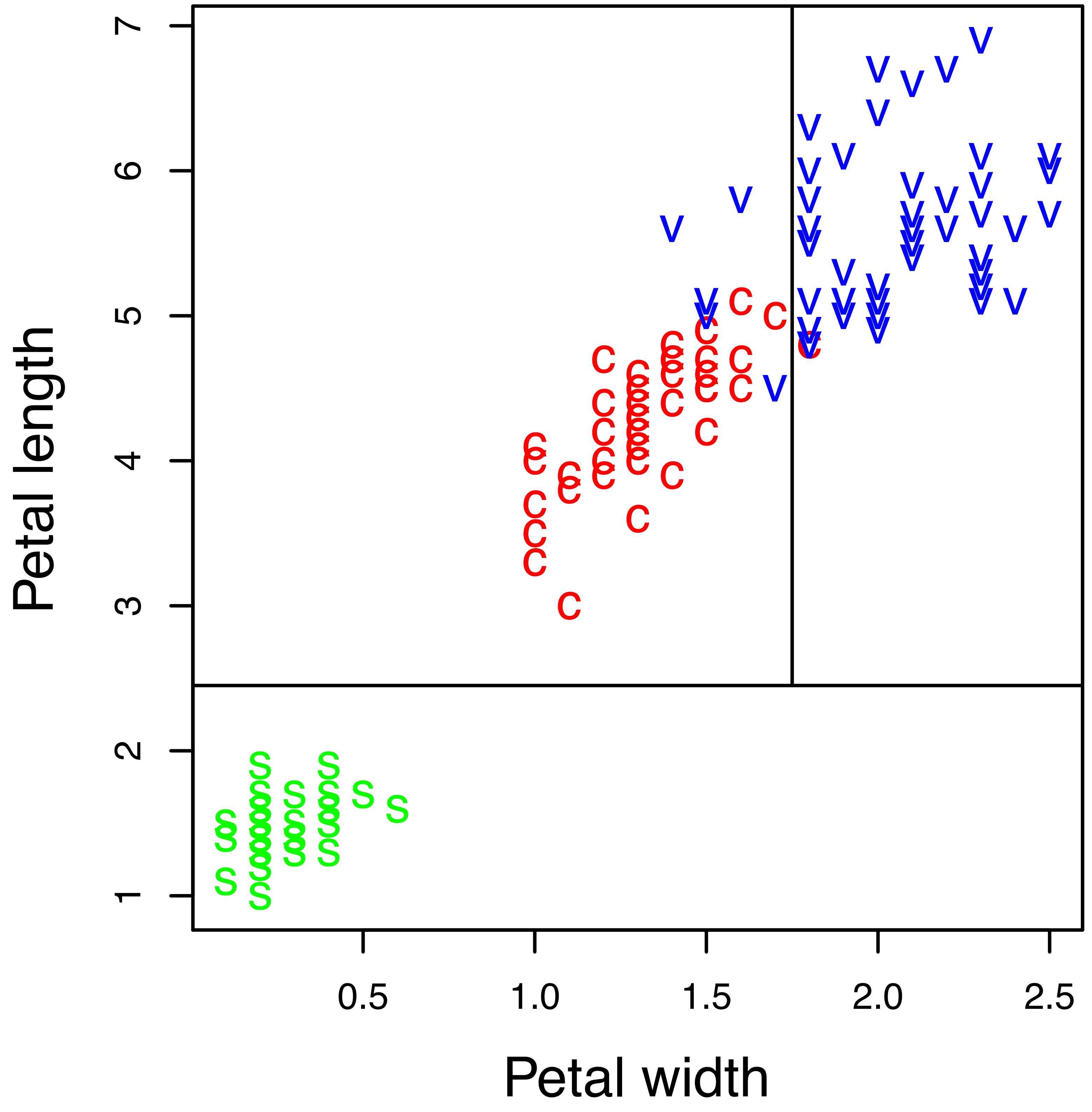
Decision trees

Development

First generation

Theta automatic interaction detection (THAID): Messenger and Mandell, 1972





Loh et al.

Decision trees

Development

Second generation

Classification and regression tree (CART): Breiman et al., 1984

- Adds cross-validation
- Adds pruning

Third generation

Quick unbiased efficient statistical tree (QUEST): Loh and Shih, 1997

- Merge classes to get binary splits

Fourth generation

Random forest: Breiman et al., 2001

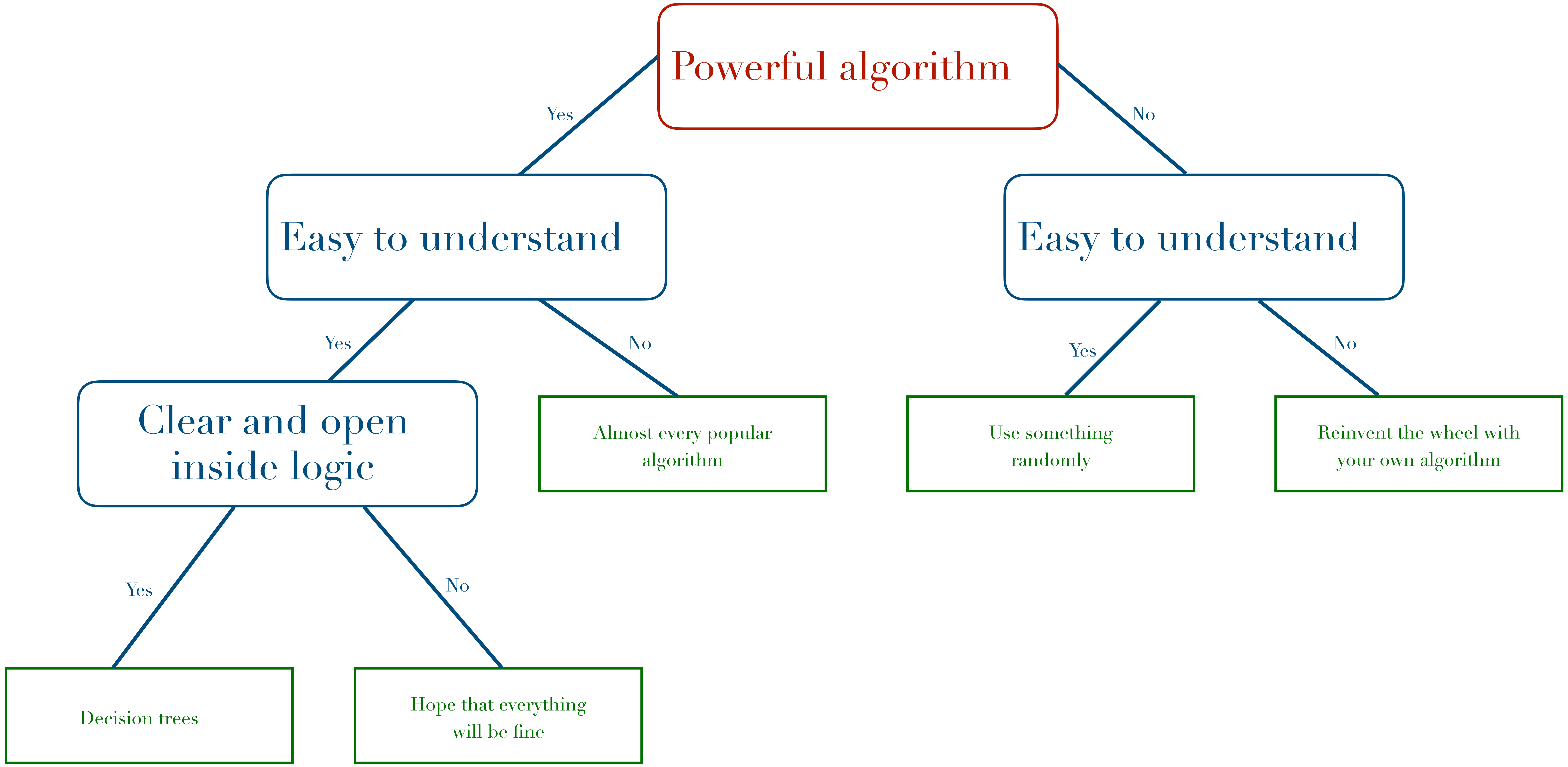
- Final predicted value is average of values from the trees

Decision trees: Key concepts

Root: Top node

Node: Each object in a tree

Leaf node: Final node


Powerful algorithm
Yes
No
Easy to understand
Easy to understand
Yes
No
Yes
No
Clear and open inside logic
Almost every popular algorithm
Use something randomly
Reinvent the wheel with your own algorithm
Yes
No
Decision trees
Hope that everything will be fine

Decision trees: Splitting criterion

Entropy

$$S = -\sum_{k \in K} p(k) \log_2 p(k)$$

Entropy values fall between 0 and 1

Pure leaf nodes (only one class) has S=0

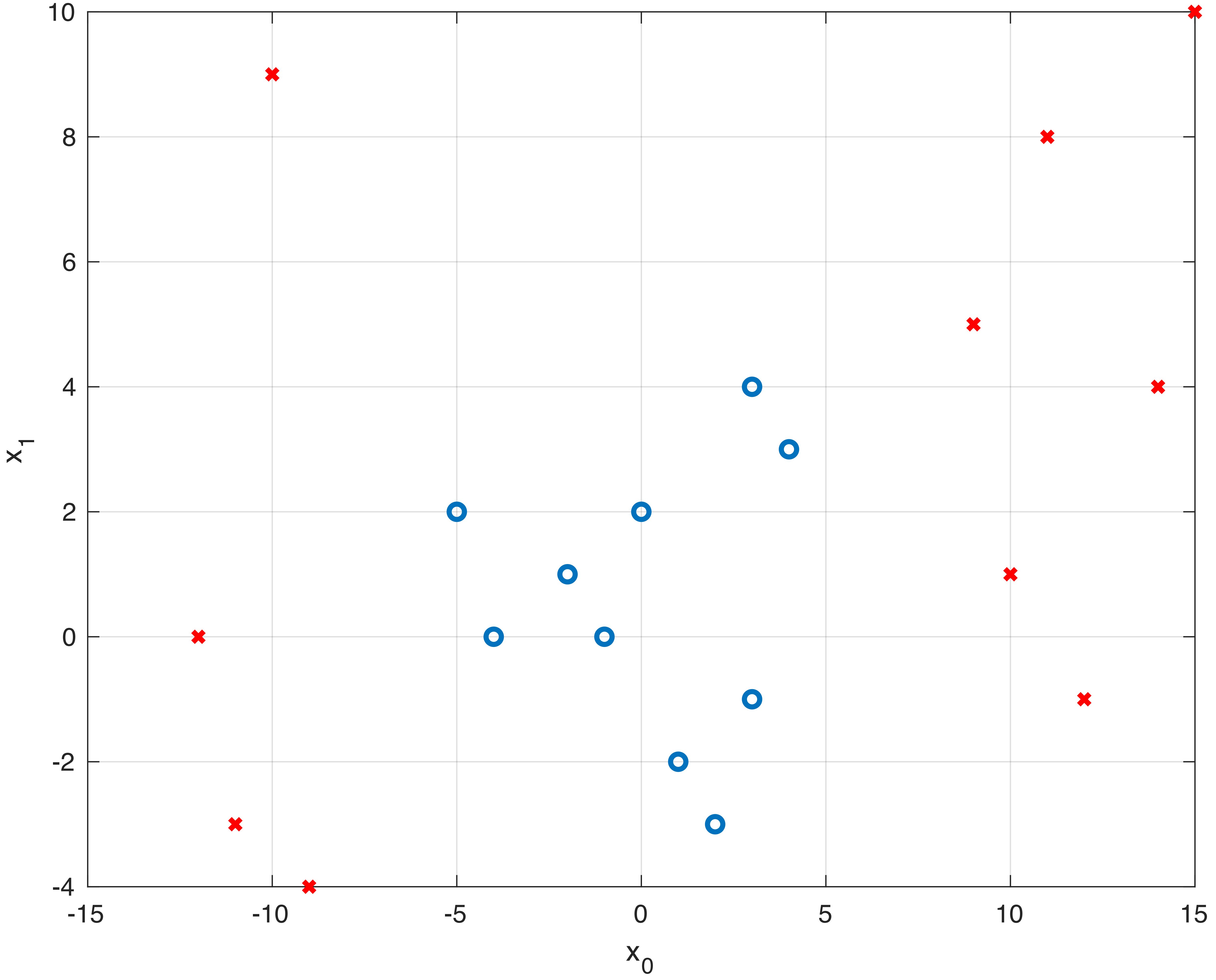
Information gain

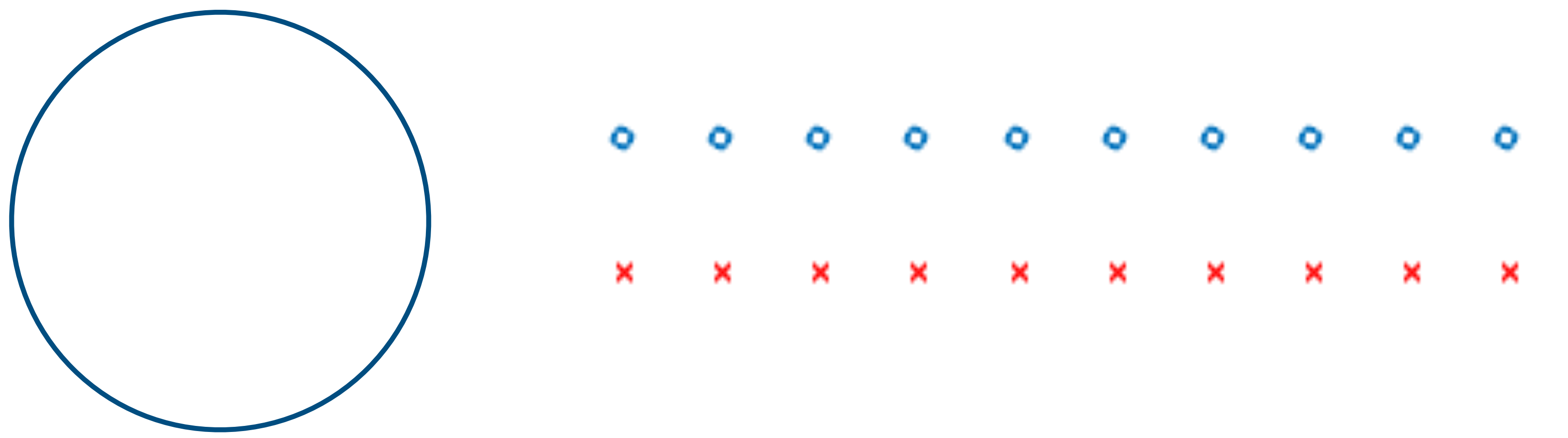
$$\Theta = S(p) - \sum w_i S(c_i)$$

Information contained in a state

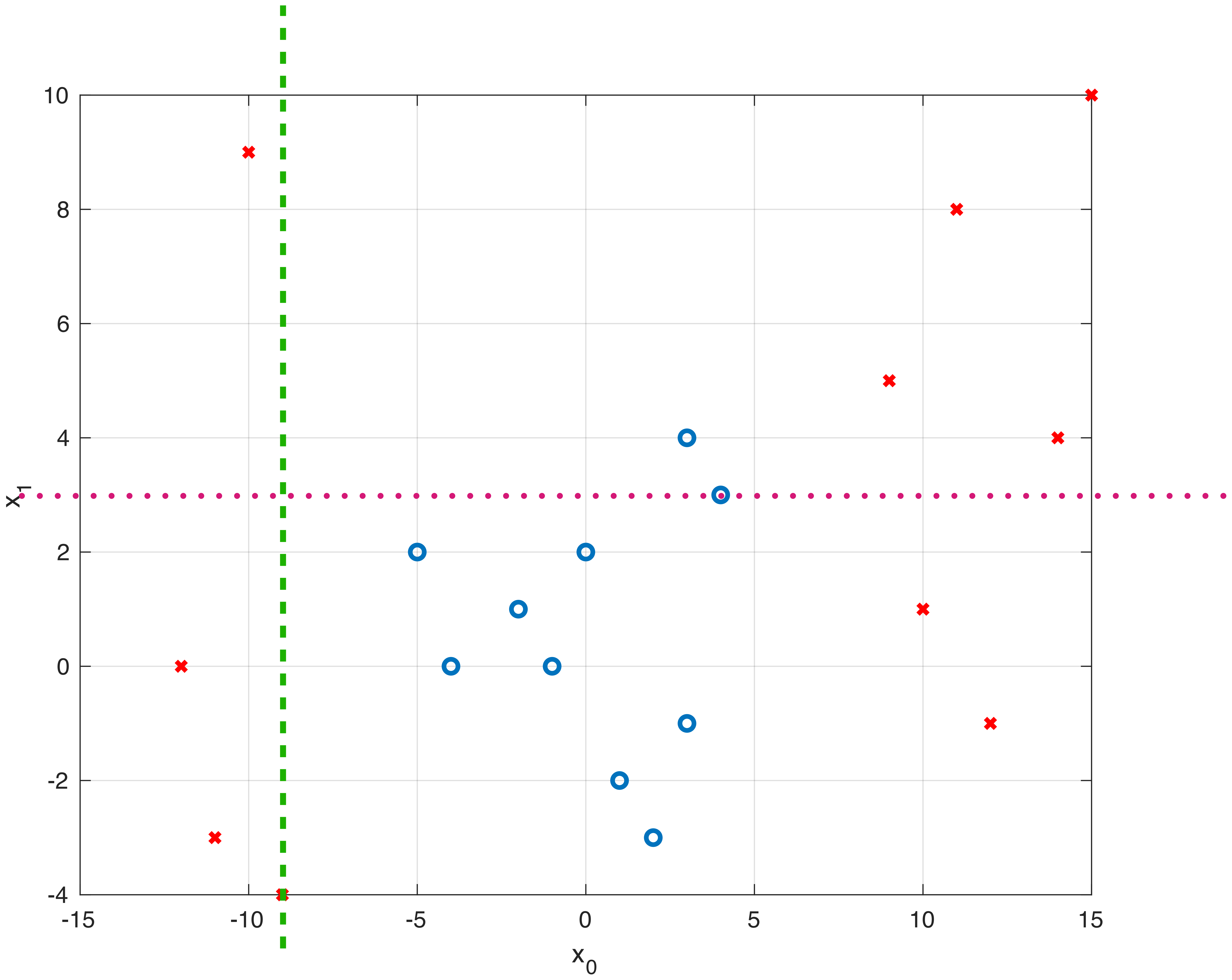
The attribute with the highest information gain will produce the best split

Decision trees: Example



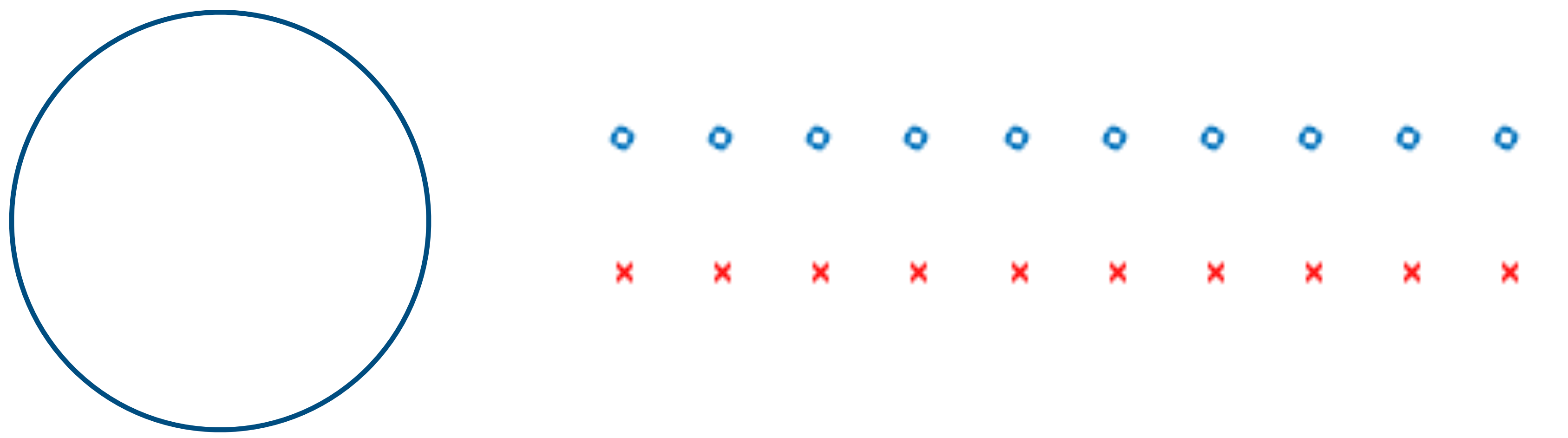


Decision trees: Example

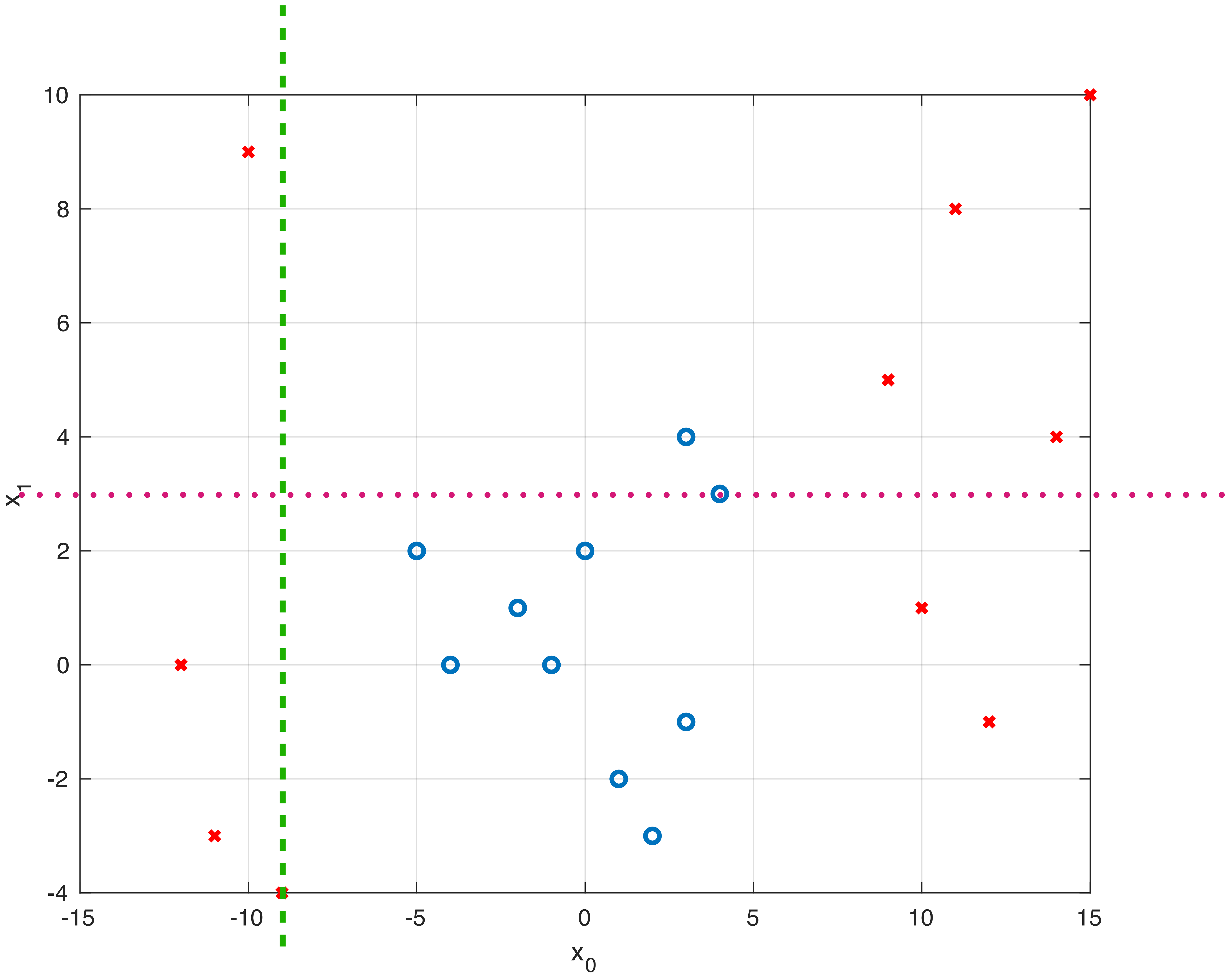


$x_1 \leq 3$

$x_0 \leq -9$

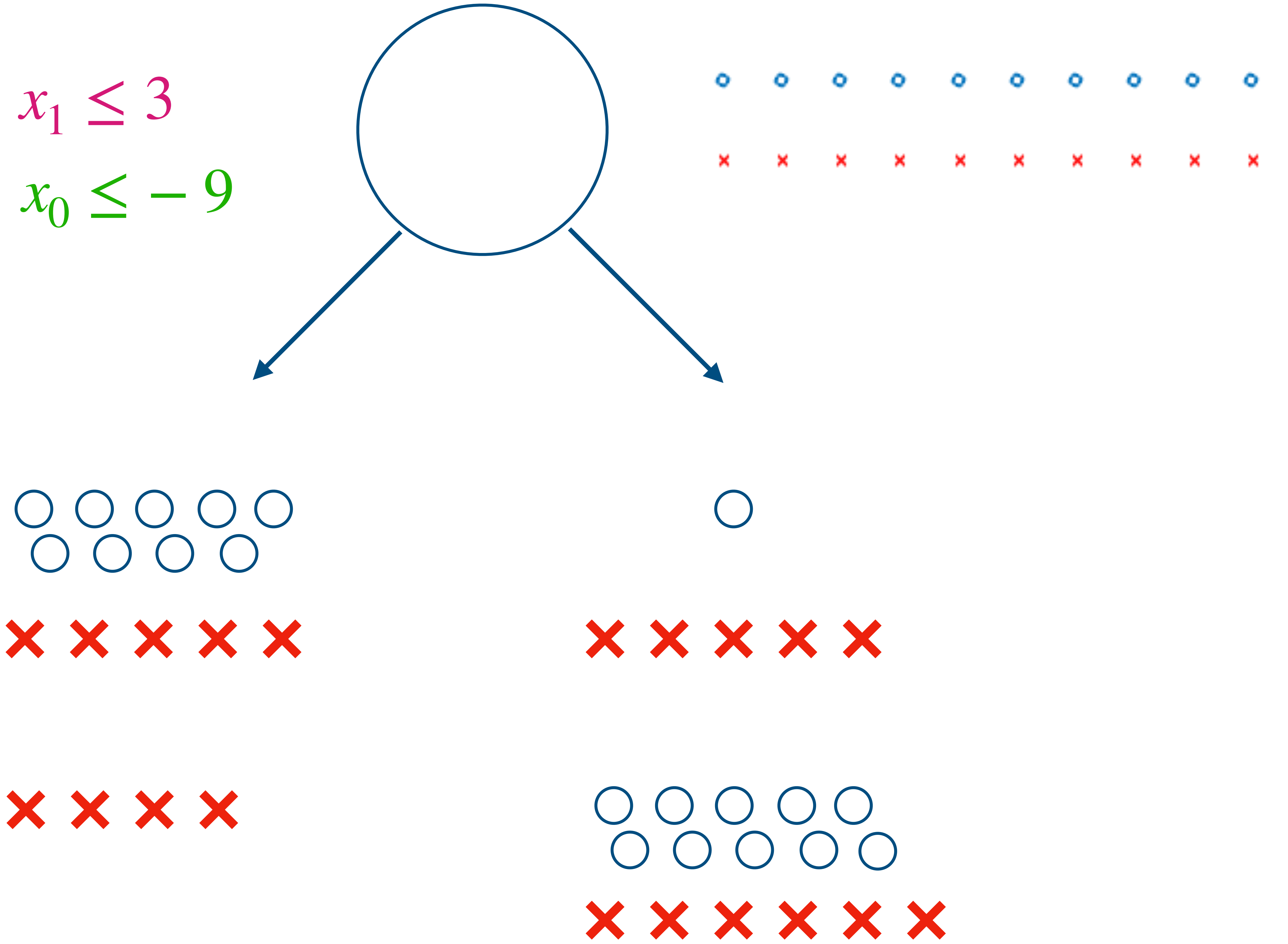


Decision trees: Example

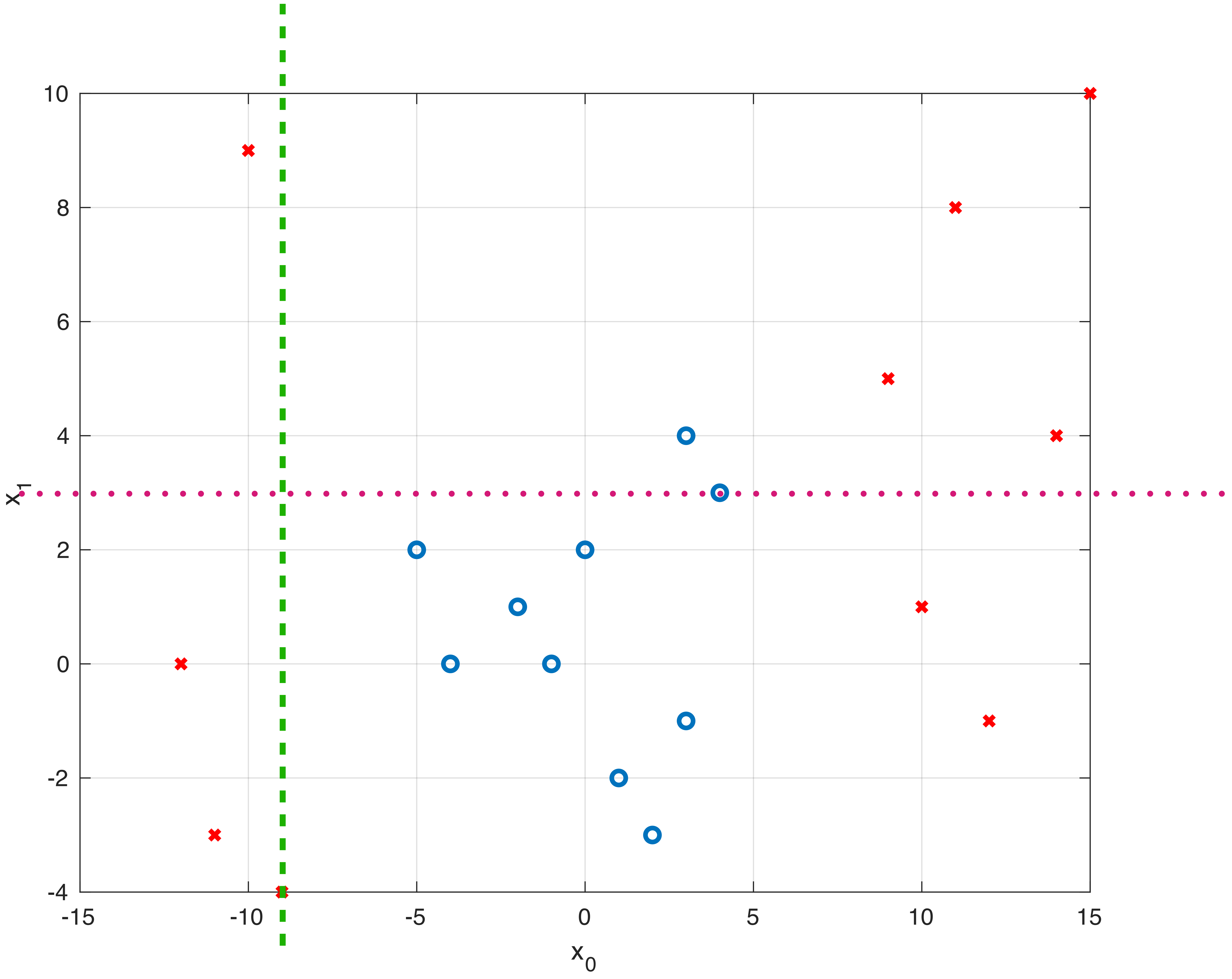


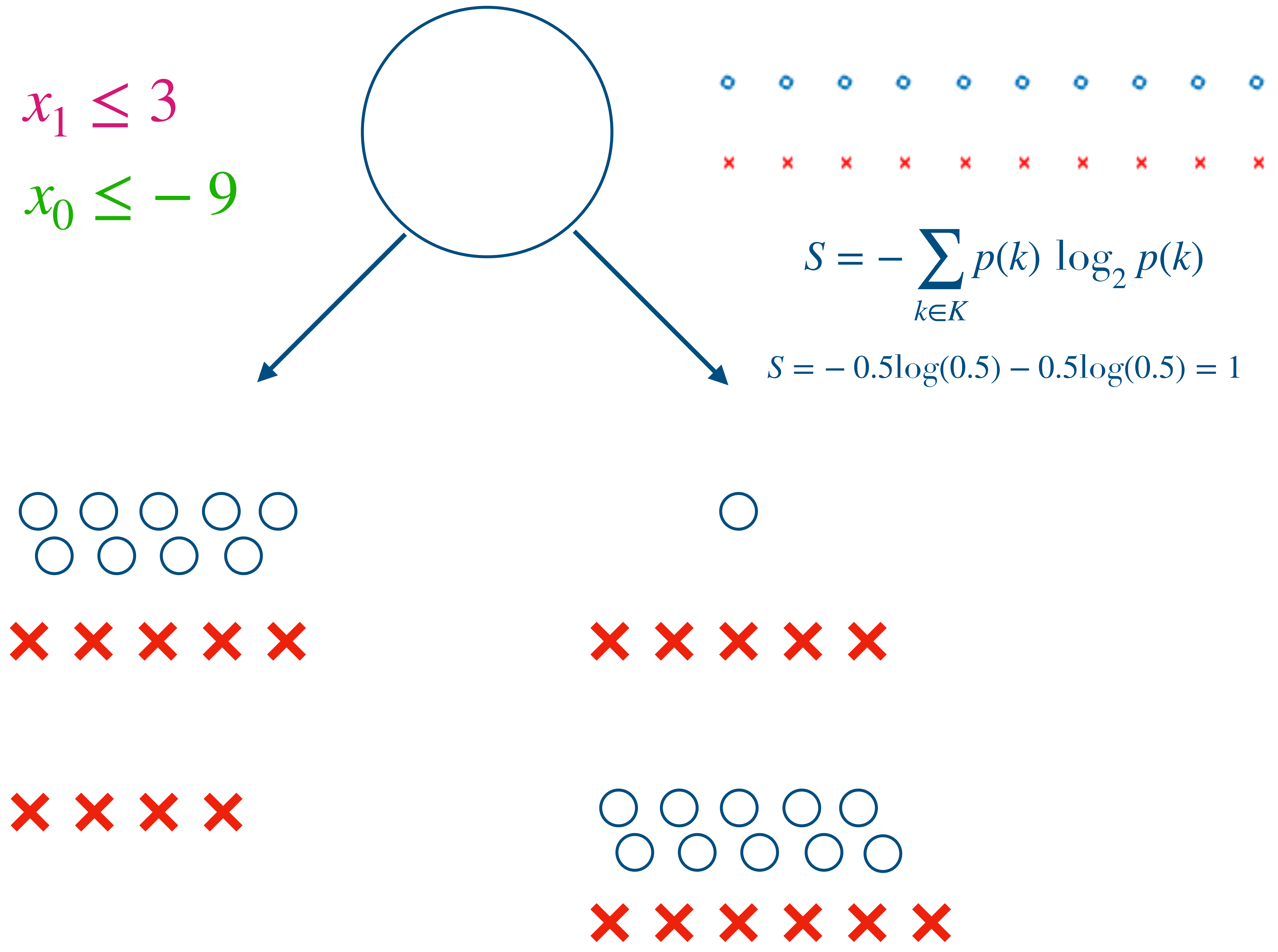
$x_1 \leq 3$

$x_0 \leq -9$



Decision trees: Example





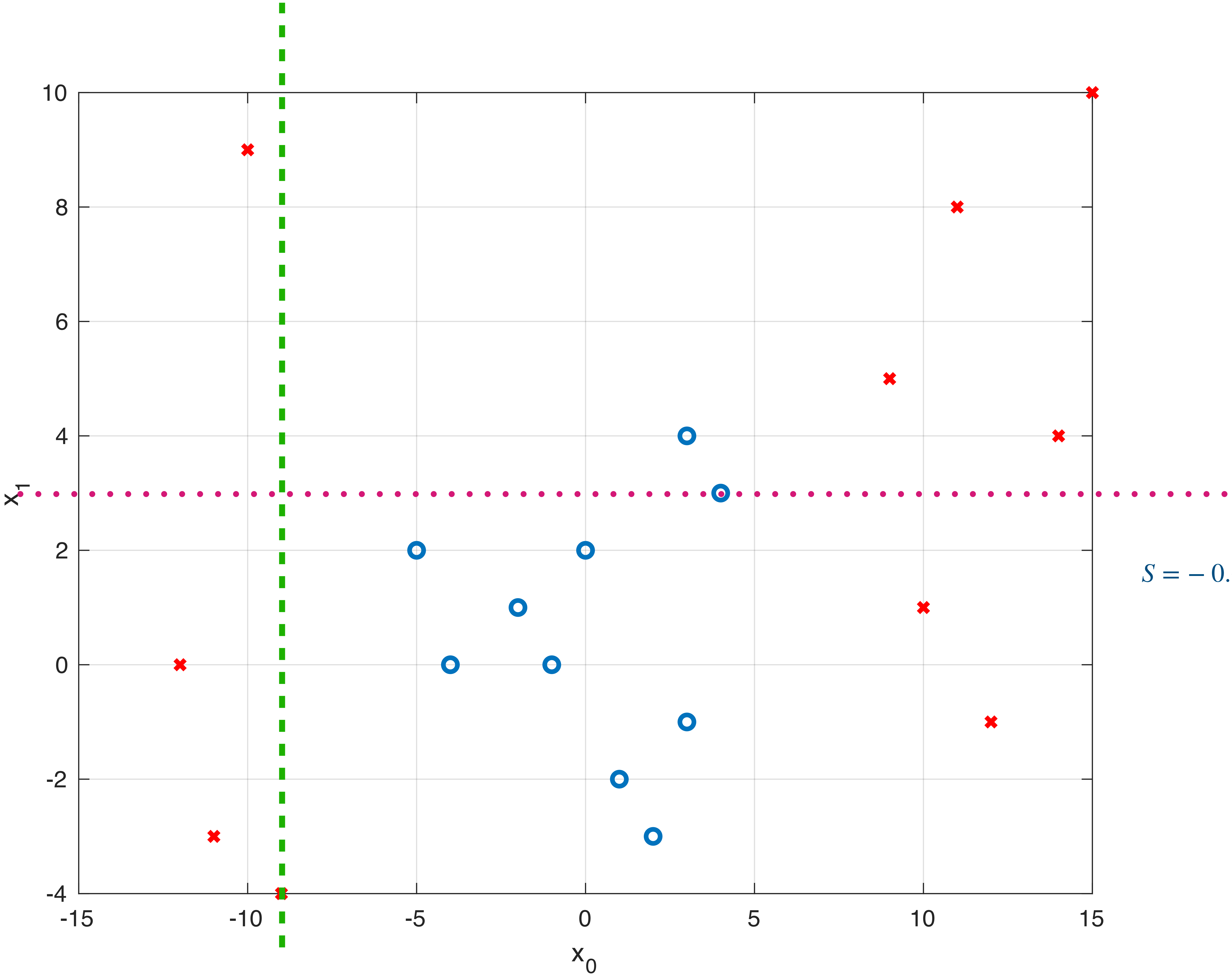
$x_1 \leq 3$

$x_0 \leq -9$

$$S = -\sum_{k \in K} p(k)\ \log_2 p(k)$$

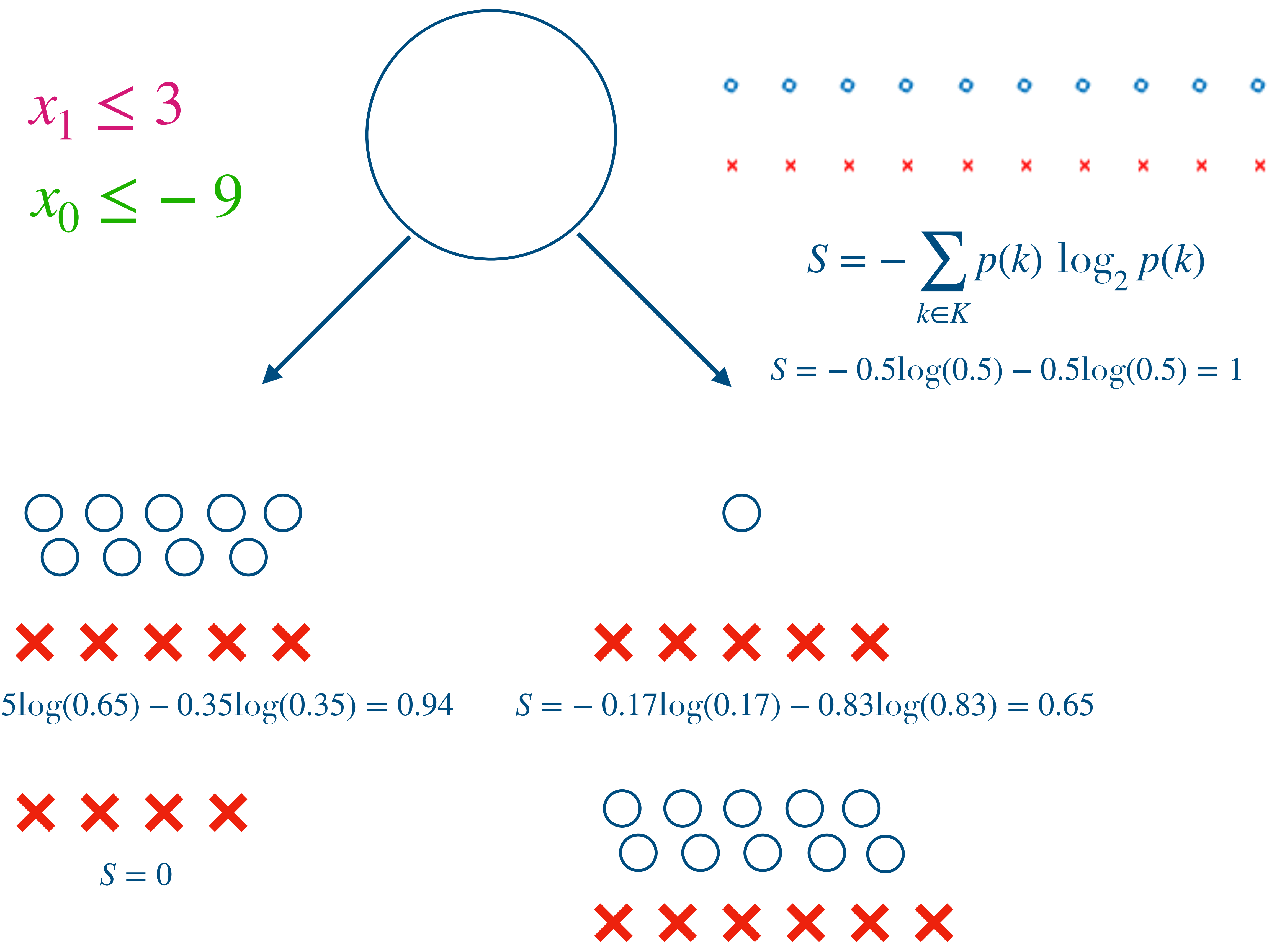
$$S = -0.5\log(0.5) - 0.5\log(0.5) = 1$$

Decision trees: Example



$x_1 \leq 3$

$x_0 \leq -9$



$$S = -\sum_{k \in K} p(k)\ \log_2 p(k)$$

$S = -0.5\log(0.5) - 0.5\log(0.5) = 1$

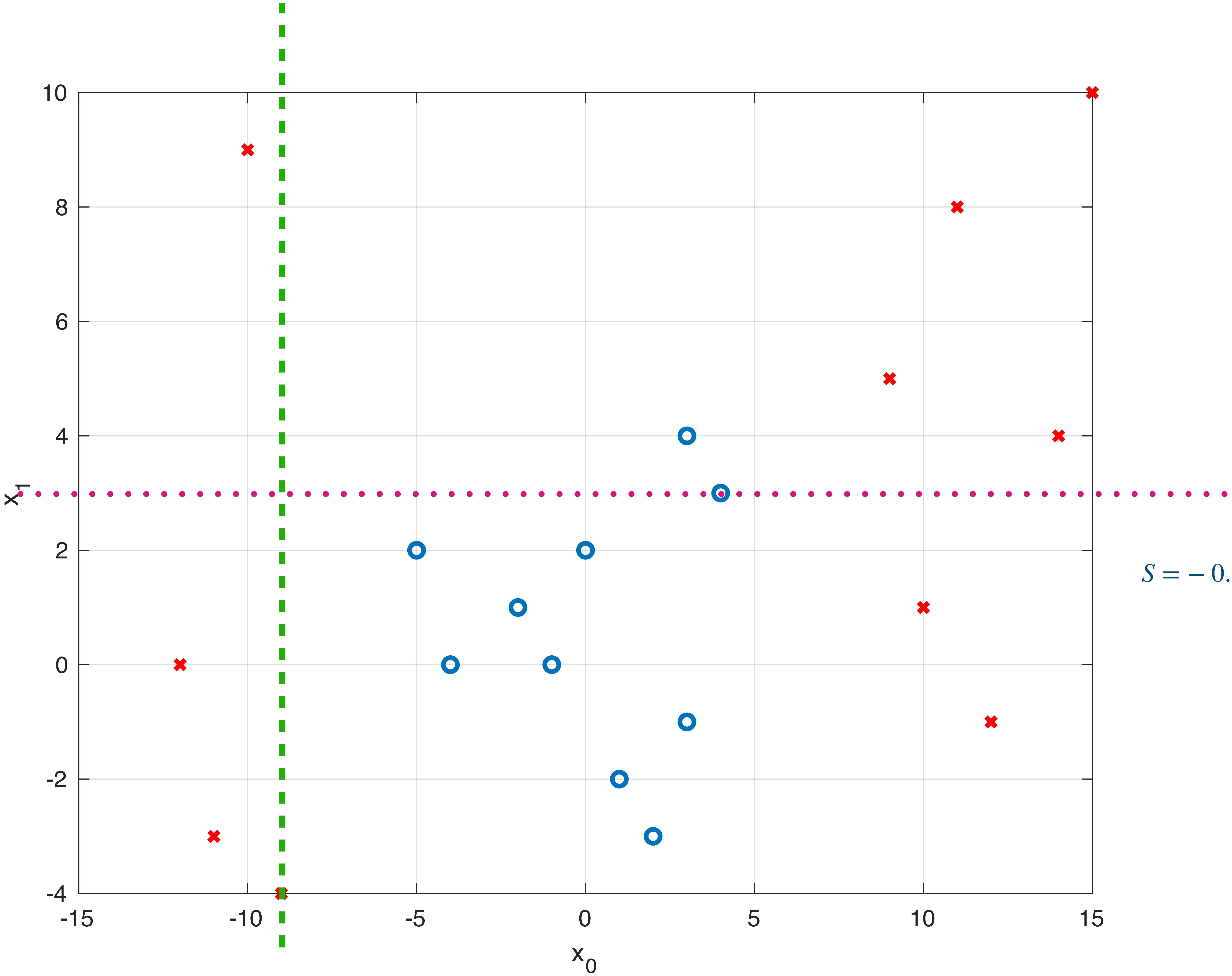
$S = -0.65\log(0.65) - 0.35\log(0.35) = 0.94$

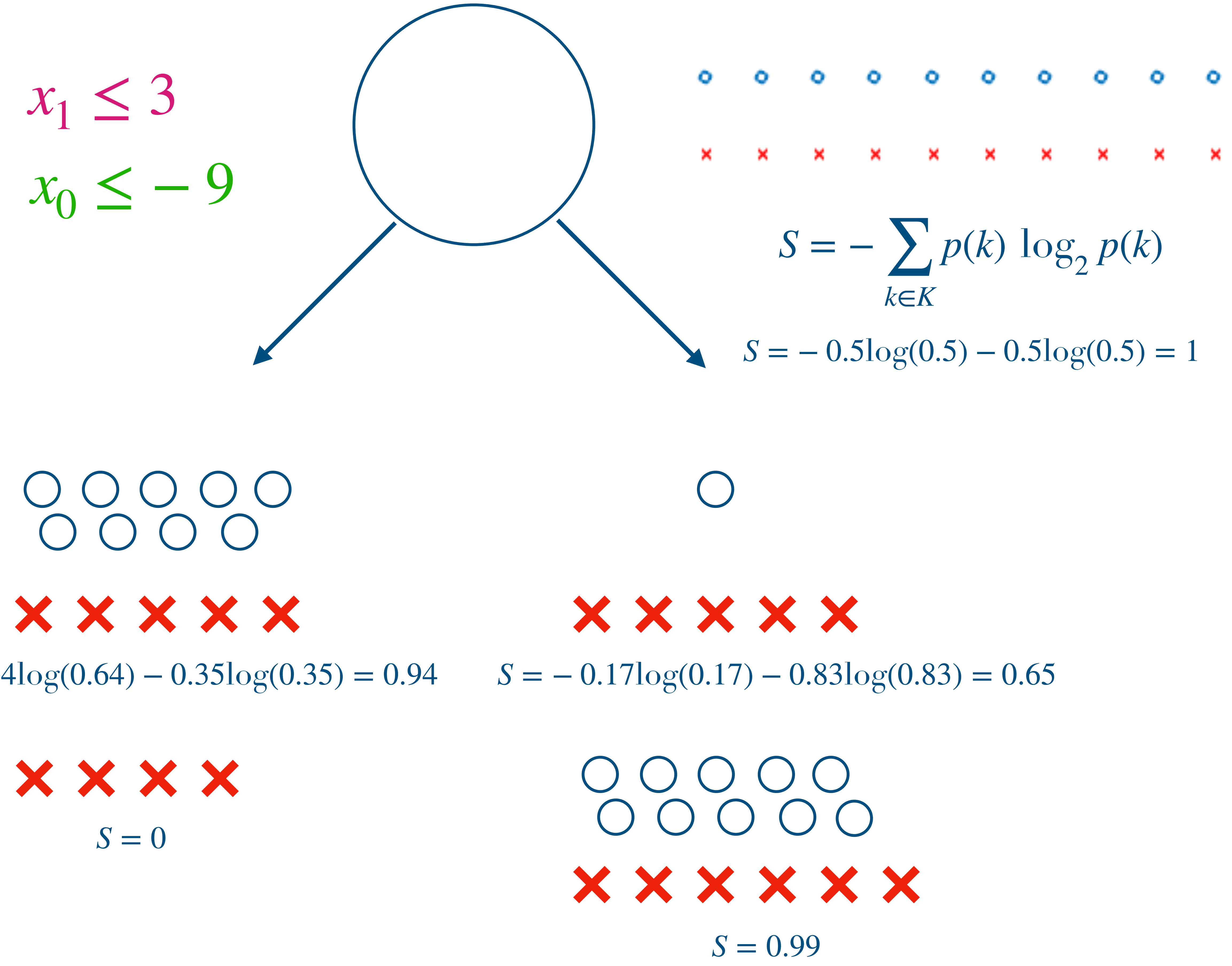
$S = -0.17\log(0.17) - 0.83\log(0.83) = 0.65$

$S = 0$

$S = 0.99$

Decision trees: Example





$x_1 \leq 3$

$x_0 \leq -9$

$$S = -\sum_{k \in K} p(k)\ \log_2 p(k)$$

$S = -0.5\log(0.5) - 0.5\log(0.5) = 1$

$S = -0.64\log(0.64) - 0.35\log(0.35) = 0.94$

$S = -0.17\log(0.17) - 0.83\log(0.83) = 0.65$

$S = 0$

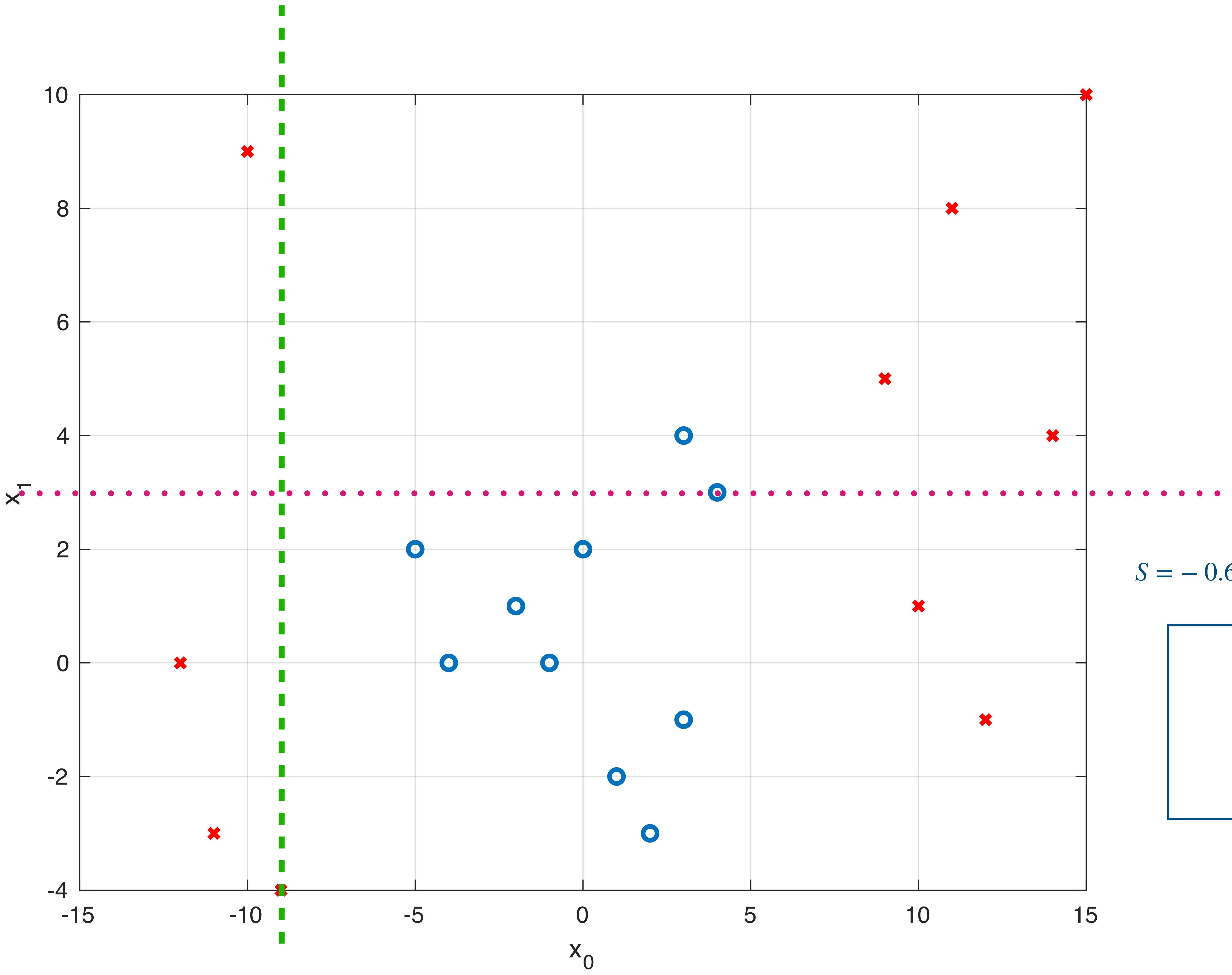
$S = 0.99$

$$\Theta = S(p) - \sum w_i S(c_i)$$

$$\Theta_1 = 1 - \frac{14}{20} * 0.94 - \frac{6}{20} * 0.65 = 0.14$$

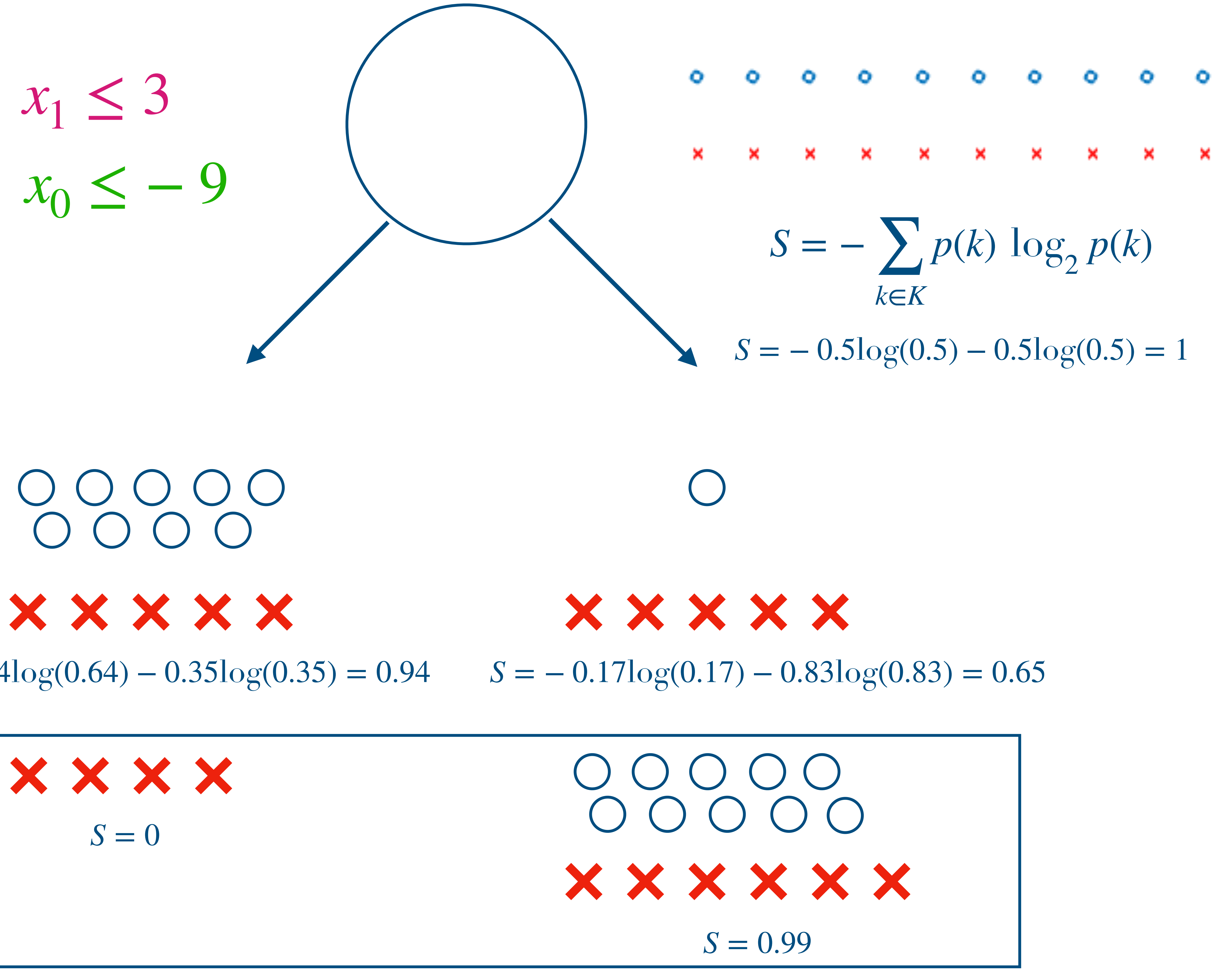
$$\Theta_2 = 1 - \frac{4}{20} * 0 - \frac{16}{20} * 0.99 = 0.2$$

Decision trees: Example



$x_1 \leq 3$

$x_0 \leq -9$



$$S = -\sum_{k \in K} p(k) \log_2 p(k)$$

$S = -0.5\log(0.5) - 0.5\log(0.5) = 1$

$S = -0.64\log(0.64) - 0.35\log(0.35) = 0.94$

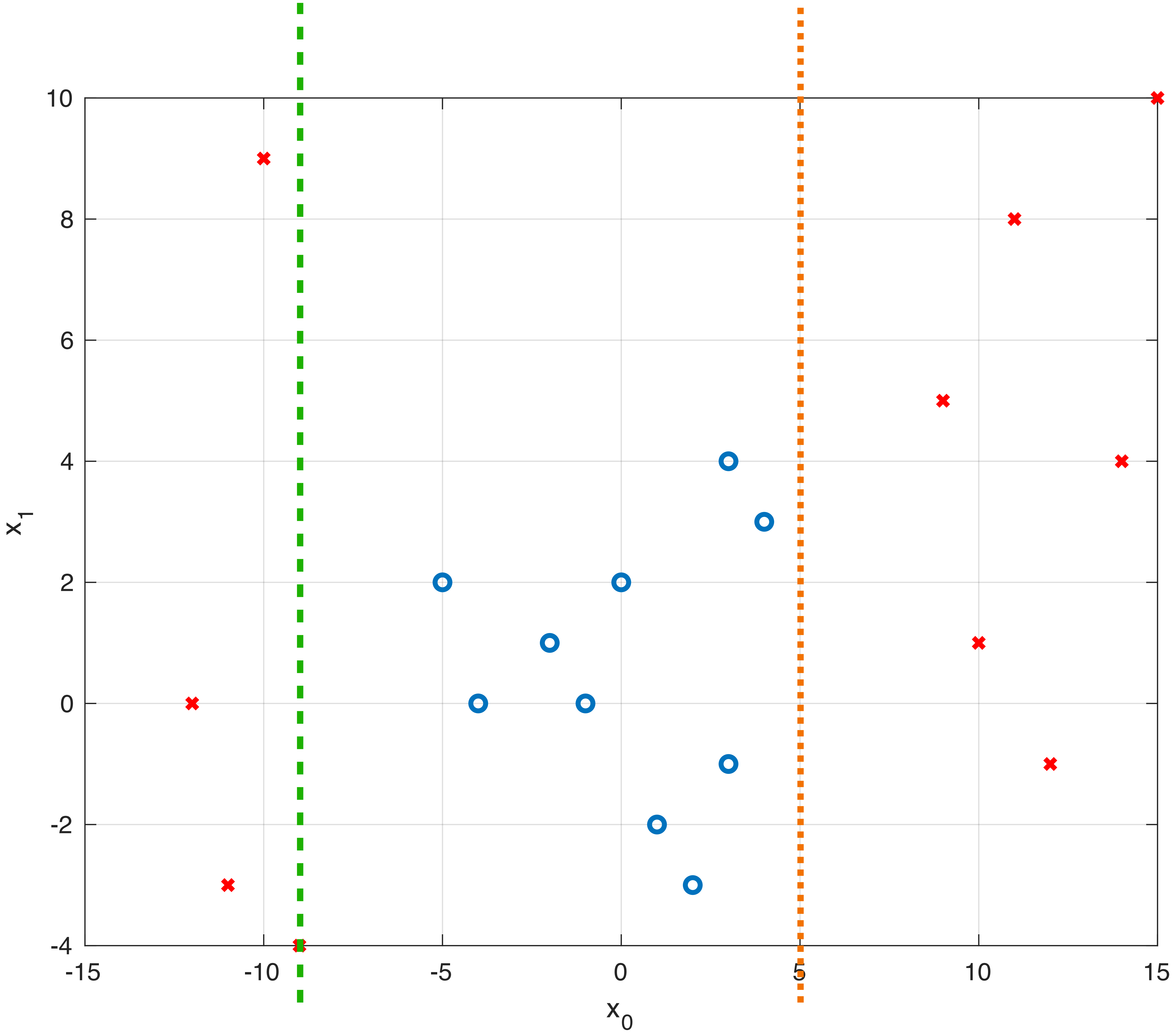
$S = -0.17\log(0.17) - 0.83\log(0.83) = 0.65$

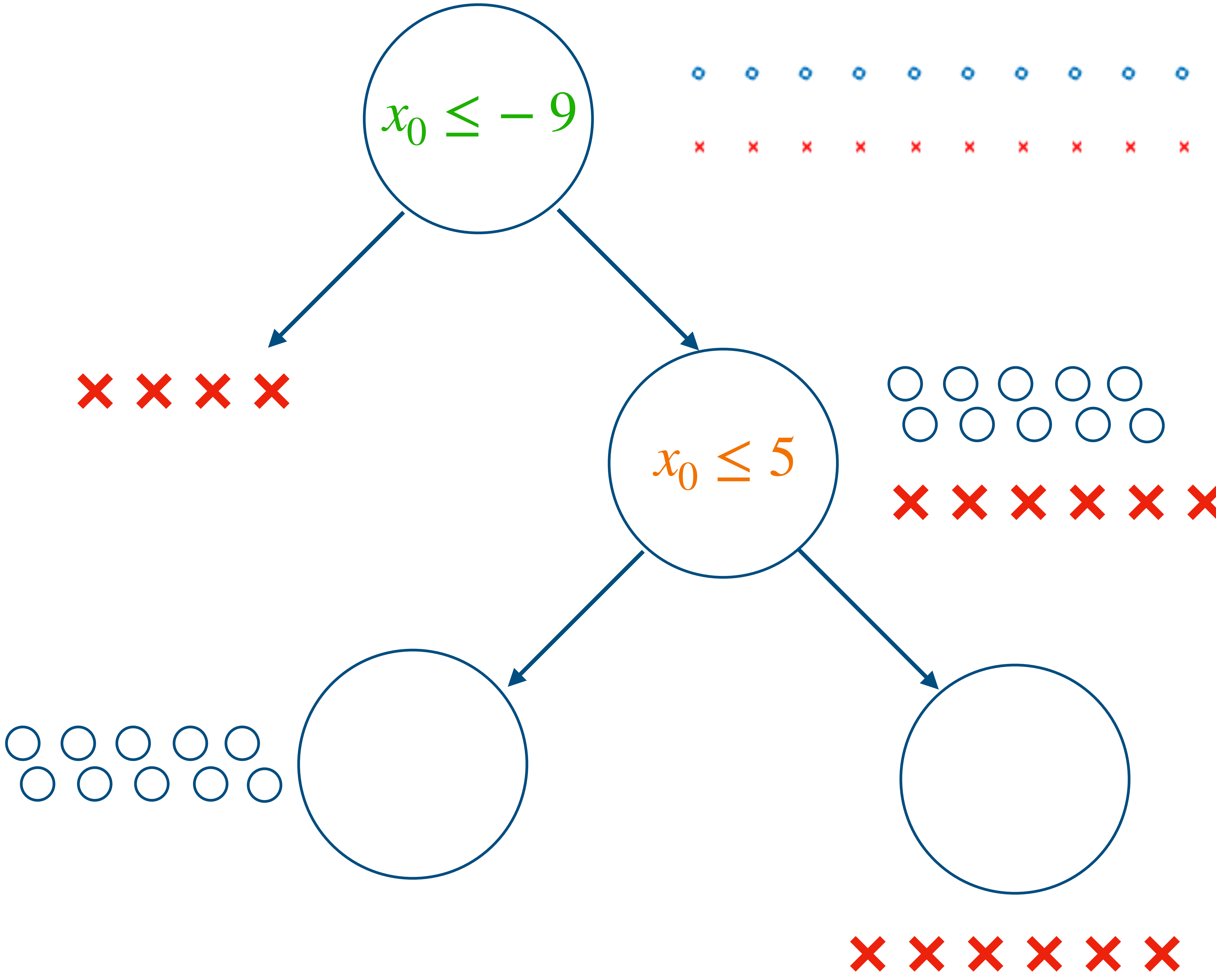
$S = 0$

$S = 0.99$

$$\Theta_2 > \Theta_1$$

Decision trees: Example





Decision trees

Advantages

Simple to understand: visual representations of decision trees make them easier to understand

Little to no data preparation

Flexible: can be leveraged for both classification and regression tasks

Disadvantages

Prone to overfitting: Complex decision trees tend to overfit

High variance estimator: Small variations within data can produce a very different decision tree

Decision trees in Matlab

Classification and regression tree (CART): Breiman et al., 1984

- Decision trees, or classification trees and regression trees, predict responses to data in Matlab
- Allows you to predict a response following the decisions in the tree from the root (beginning) node down to a leaf node
- The leaf node contains the response
- Classification trees give nominal responses, such as `'true'` or `'false'`
- Regression trees give numeric responses.

Decision trees in Matlab

HOME
PLOTS
APPS
EDITOR
PUBLISH
VIEW
Search Documentation
Sign In
New
Open
Save
Compare
Print
FILE
Go To
Find
Bookmark
NAVIGATE
Refactor
CODE
Profiler
Analyze
ANALYZE
Run Section
Section Break
Run and Advance
Run to End
SECTION
Run
Step
Stop
RUN
/ Users lazperez Library Mobile Documents com~apple~CloudDocs MATLAB GPHS
Current Folder
Name
Details
Editor – /Users/lazperez/Library/Mobile Documents/com~apple~CloudDocs/MATLAB/GPHS/gphs_dectrees.m
gphs_rSquared.m
gphs_dectrees.m
Command Window
Workspace
Command History

Current Folder:

- gphs_rSquared.asv
- StarWarsCharacters.csv
- train_SW_Classifier.m
- gphs_dectrees.m
- gphs_rSquared.m
- gphs_tableCreation.m
- dectree.pdf

```
%% View Decision Tree
% This example shows how to view a classification or regression tree
% There are two ways to view a tree: view(tree) returns a text description and view(tree,'mode','graph') returns a graphic description of the tree.
% Create and view a classification tree.

load fisheriris % load the sample data
ctree = fitctree(meas,species); % create classification tree
view(ctree) % text description
%% view the tree
view(ctree,'mode','graph') % graphic description
```

Command Window:

```
Decision tree for classification
1  if x3<2.45 then node 2 elseif x3>=2.45 then node 3 else setosa
2  class = setosa
3  if x4<1.75 then node 4 elseif x4>=1.75 then node 5 else versicolor
4  if x3<4.95 then node 6 elseif x3>=4.95 then node 7 else versicolor
5  class = virginica
6  if x4<1.65 then node 8 elseif x4>=1.65 then node 9 else versicolor
7  class = virginica
8  class = versicolor
9  class = virginica
```

Workspace:

Name	Value	Max	Min
ctree	1x1 Classific...		
meas	150x4 double	7.9000	0.100(
species	150x1 cell		

Command History:

```
2/6
8/14
-9/14*(log2(9/14)) - 5/14*(log2(5...
9/14
5/14
-1/6*(log2(1/6)) - 5/6*(log2(5/6))
1/6
```

Decision trees in Matlab

HOME
PLOTS
APPS
EDITOR
PUBLISH
VIEW
New
Open
Save
Compare
Print
FILE
Go To
Find
Bookmark
NAVIGATE
Refactor
CODE
Profil
Analy
ANALYZ
/ Users lazperez Library Mobile Documents com~apple~C
Current Folder
Name
gphs_rSquared.asv
StarWarsCharacters.csv
train_SW_Classifier.m
gphs_dectrees.m
gphs_rSquared.m
gphs_tableCreation.m
dectree.pdf
Details
Editor – /Users/lazperez/Library/Mobile Docu
gphs_rSquared.m
gphs_dectrees.m
Command Window
Classification tree viewer
File
Tools
Desktop
Tree
Window
Help
Click to display:
Identity
Magnification:
100%
Pruning level:
0 of 4
x3 < 2.45
x3 >= 2.45
setosa
x4 < 1.75
x4 >= 1.75
x3 < 4.95
x3 >= 4.95
virginica
x4 < 1.65
x4 >= 1.65
virginica
versicolor
virginica
Search Documentation
Sign In
Workspace
Command History

```
%% View Decision Tree
% This example shows how to view a c
% There are two ways to view a tree:
% Create and view a classification t

load fisheriris % load the sample da
ctree = fitctree(meas,species); % cr
view(ctree) % text description
%% view the tree
view(ctree,'mode','graph') % graphic
```

```
Decision tree for classification
1  if x3<2.45 then node 2 elseif x3>=2.4
2  class = setosa
3  if x4<1.75 then node 4 elseif x4>=1.7
4  if x3<4.95 then node 6 elseif x3>=4.9
5  class = virginica
6  if x4<1.65 then node 8 elseif x4>=1.65 then node 9 else versicolor
7  class = virginica
8  class = versicolor
9  class = virginica
```

Name	Value	Max	Min
ctree	1x1 Classific...		
meas	150x4 double	7.9000	0.100
species	150x1 cell		

```
2/6
8/14
-9/14*(log2(9/14)) - 5/14*(log2(5...
9/14
5/14
-1/6*(log2(1/6)) - 5/6*(log2(5/6))
```