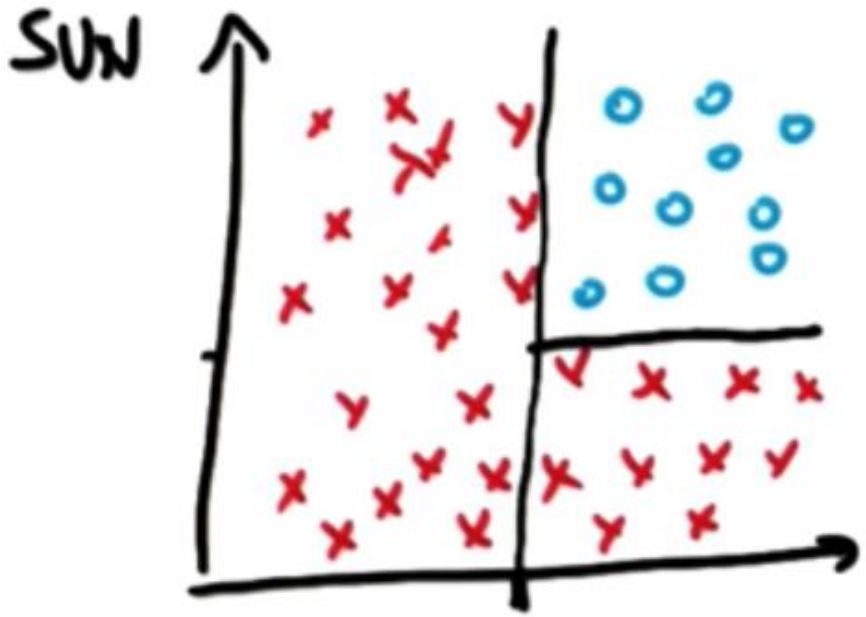
DECISION TREES

WIND SURF

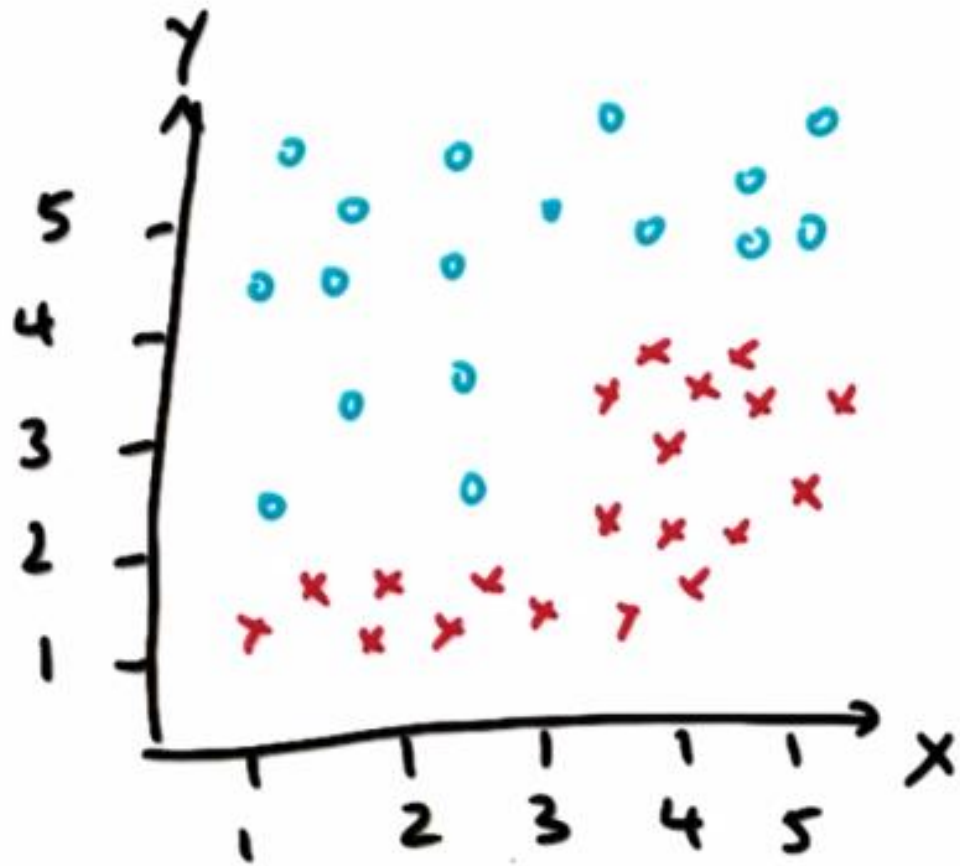
IS THIS DATA LINEARLY SEPARABLE ?

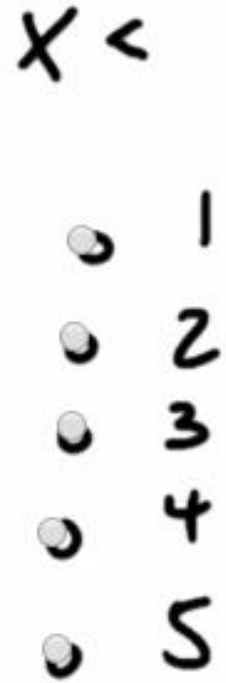
○ YES ○ NO

DECISION TREES

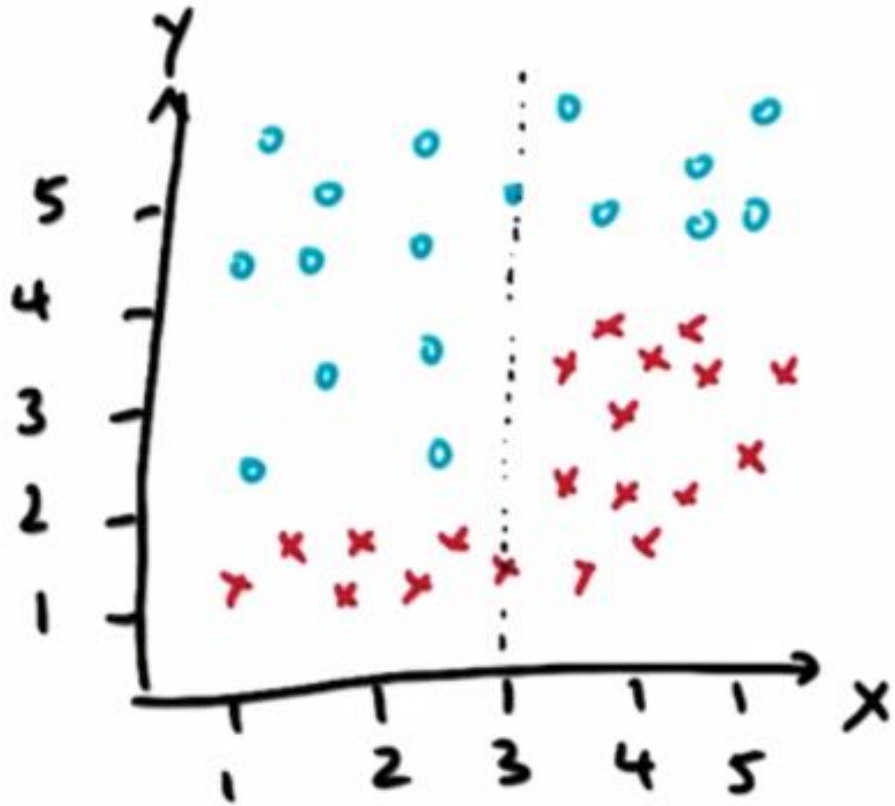


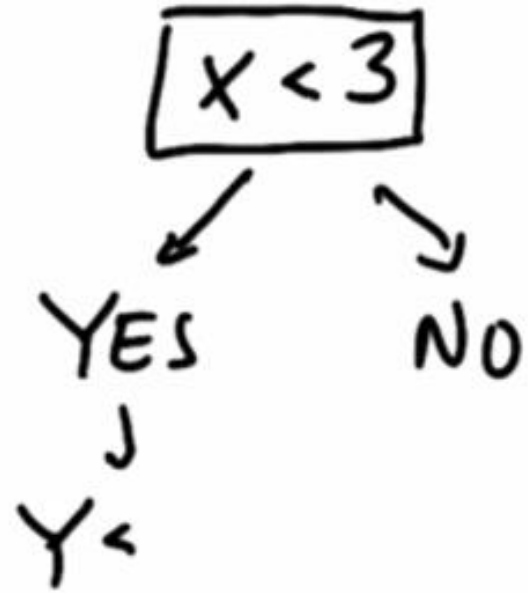
DECISION TREE

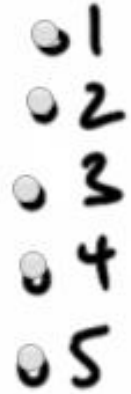


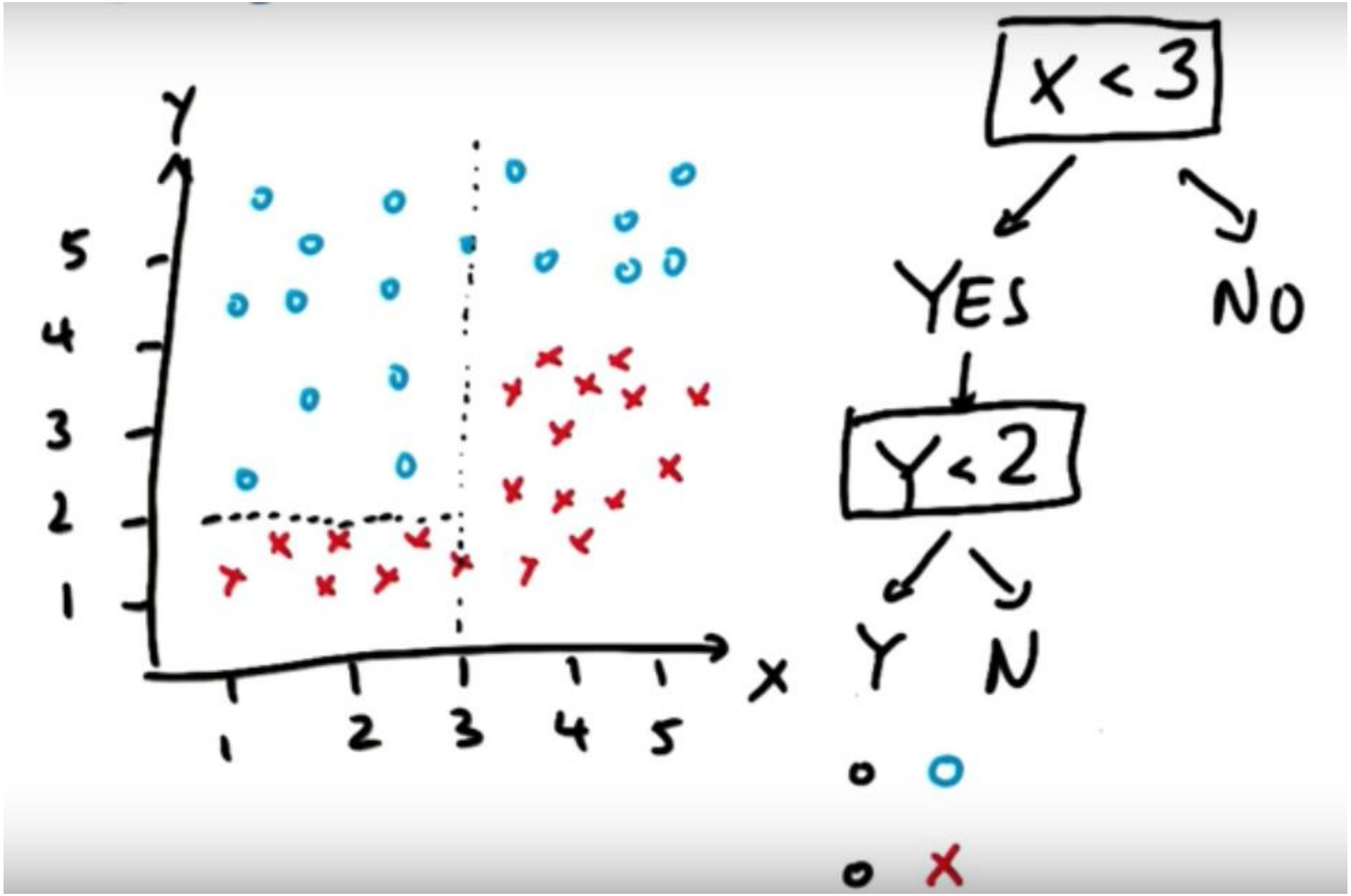


DECISION TREE








Y
5
4
3
2
1
X
1
2
3
4
5
X < 3
YES
NO
Y < 2
Y
N
o
o
o
X

Entropy

← controls how a DT decides where to split the data

definition: measure of impurity in a bunch of examples

all examples are same class

→ entropy = 0

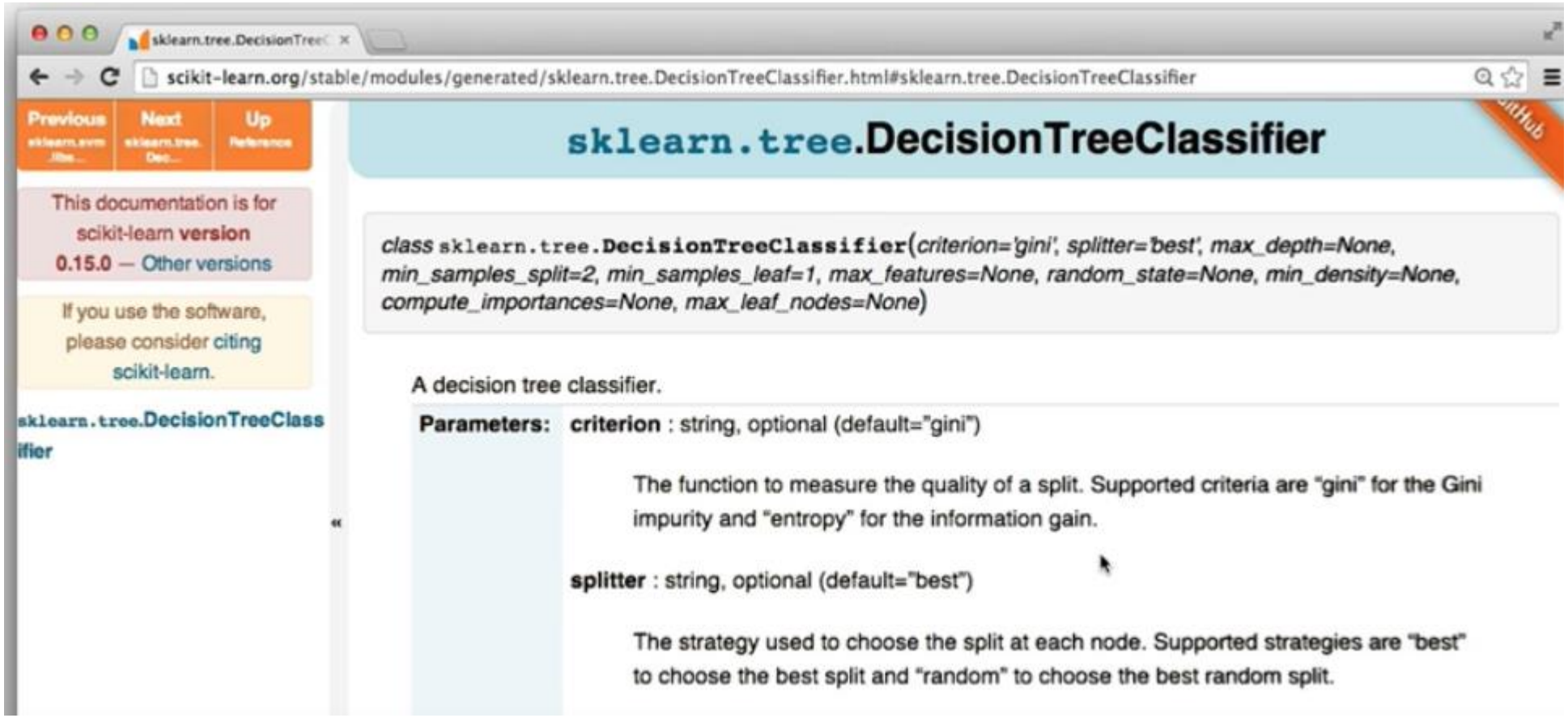
examples are evenly split between classes

→ entropy = 1.0

Information Gain

information gain = entropy (parent) − [weighted average] entropy (children)

decision tree algorithm : maximize information gain.


sklearn.tree.DecisionTreeC
scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier
Previous
Next
Up
This documentation is for scikit-learn version 0.15.0 — Other versions
If you use the software, please consider citing scikit-learn.
sklearn.tree.DecisionTreeClassifier
sklearn.tree.DecisionTreeClassifier
class sklearn.tree.DecisionTreeClassifier(criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features=None, random_state=None, min_density=None, compute_importances=None, max_leaf_nodes=None)
A decision tree classifier.
Parameters:
criterion : string, optional (default="gini")
The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.
splitter : string, optional (default="best")
The strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.

FORTALEZAS

* Facil de manipular y entender

DEBILIDADES

* Propensión a overfitting (profundidad del árbol)