

Decision and Regression Trees

Applied machine learning (EDAN96)

Lecture 11 — Decision Trees

2023–12–04

Elin A. Topp

Lindholm et al, ch 2.3

Information Gain lecture of F. Aiolli: <https://www.math.unipd.it/~aiolli/corsi/0708/IR/Lez12.pdf>
various sources

About Elin

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022
- AI for HRI, robot skills and robotic learning, mixed-initiative interaction in SAR

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022
- AI for HRI, robot skills and robotic learning, mixed-initiative interaction in SAR
- Teaching in courses: IAS, Applied ML, AI (course responsible), Adv Applied ML, (Project course), (Graduate courses)

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022
- AI for HRI, robot skills and robotic learning, mixed-initiative interaction in SAR
- Teaching in courses: IAS, Applied ML, AI (course responsible), Adv Applied ML, (Project course), (Graduate courses)
- Supervising / examining MSc theses

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022
- AI for HRI, robot skills and robotic learning, mixed-initiative interaction in SAR
- Teaching in courses: IAS, Applied ML, AI (course responsible), Adv Applied ML, (Project course), (Graduate courses)
- Supervising / examining MSc theses
- WASP (Wallenberg AI, Autonomous Systems and Software Program) Graduate School Management; LU profile area NAC: 1, 2, many; LTH profile area AI and Digitalisation; COMPUTE Steering Group; DIGG advisory board; Associate editor ACM THRI; Local co-chair HRI2023 in Stockholm

About Elin

- MSc CS Karlsruhe University (now KIT) in 2003 (well, stamped 2004 ;-)
MSc thesis project conducted at KTH, Stockholm
- PhD in 2009, KTH, Stockholm.
Thesis: “Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping” (defended Oct 2008).
- At Lund University since spring 2009, lecturer since 2012, docent since 2019, member of LTH’s pedagogical academy (ETP) since 2022
- AI for HRI, robot skills and robotic learning, mixed-initiative interaction in SAR
- Teaching in courses: IAS, Applied ML, AI (course responsible), Adv Applied ML, (Project course), (Graduate courses)
- Supervising / examining MSc theses
- WASP (Wallenberg AI, Autonomous Systems and Software Program) Graduate School Management; LU profile area NAC: 1, 2, many; LTH profile area AI and Digitalisation; COMPUTE Steering Group; DIGG advisory board; Associate editor ACM THRI; Local co-chair HRI2023 in Stockholm
- Director of graduate studies, (deputy) head of department of Computer Science

Today's agenda

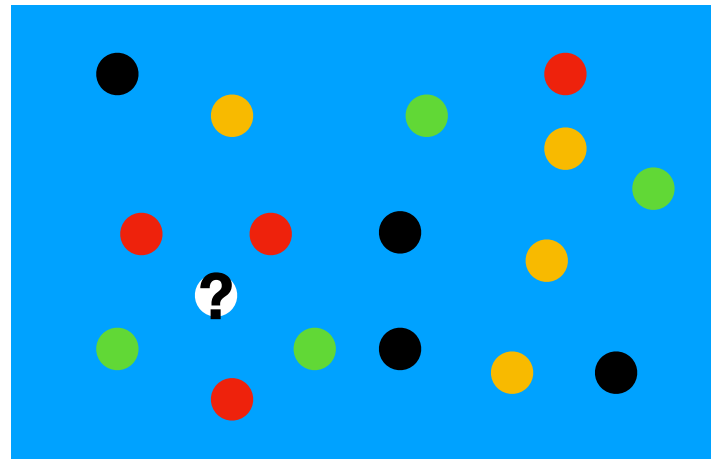
Instance based supervised learning, is a ML paradigm where the model makes predictions based on specific examples from the training data. Instead of learning a general model that captures the underlying patterns of the entire dataset, instance-based learning focuses on storing and memorising the individual training instances

- Instance based supervised learning (see also Lindholm et al, chapter 2):
 - (briefly) k-NN (“clustering”)
 - (mainly) Decision Trees (rule based “clustering”)
 - The “ConceptData” toy data set for illustration
 - Information Gain calculations (“Toy Data”, not used in the assignment)
 - Transfer of the DT idea to regression

recap: k-Nearest Neighbour

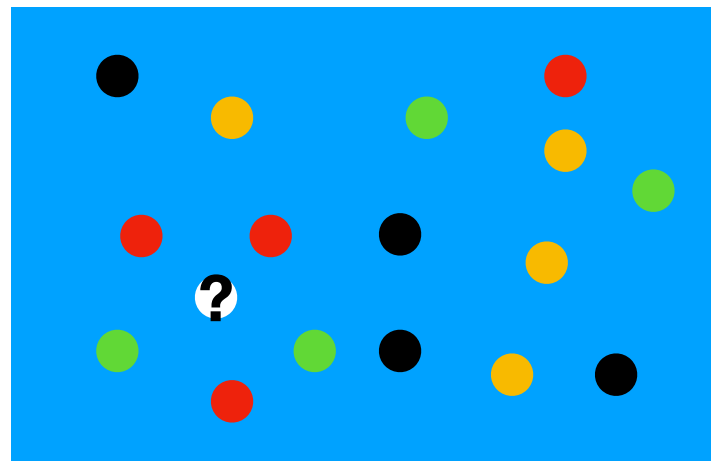
The most common type of instance-based learning is k-Nearest Neighbors (k-NN), where k is a user-defined parameter representing the number of nearest neighbors to consider. When a new instance needs to be classified or predicted, the algorithm looks at the k training instances that are closest to the input instance in the feature space and assigns the majority class or averages the values of those neighbors.

- k-Nearest Neighbour classifier votes over the k closest points (use average for regression)



recap: k-Nearest Neighbour

- k-Nearest Neighbour classifier votes over the k closest points (use average for regression)



$$? = \text{red dot}$$

(Revisiting?) The concept learning problem: Classification

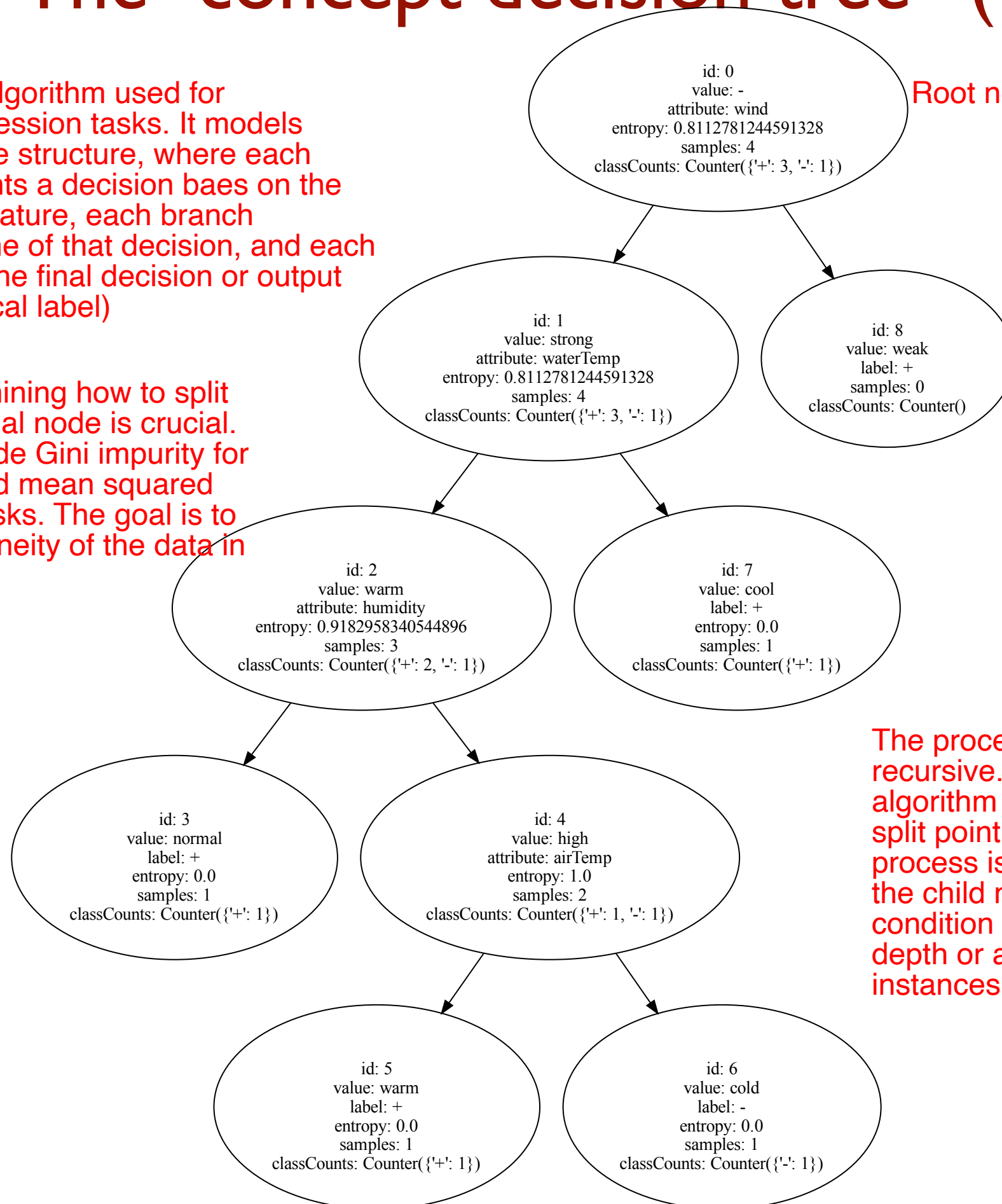
Example	Wind	Water	Humidity	AirTemp	Sky	Forecast	EnjoySport?
1	Strong	Warm	Normal	Warm	Sunny	Same	Yes
2	Strong	Warm	High	Warm	Sunny	Same	Yes
3	Strong	Warm	High	Cold	Rainy	Change	No
4	Strong	Cool	High	Warm	Sunny	Change	Yes

- We can make the decision based on a Decision Tree.
- Go through the examples attribute by attribute and split the data into subsets according to their attribute value
- Stop, when there are no more attributes to test or a sample set is “pure” (only contains examples of one class)
- When using the tree (classifying an unseen sample), follow the decisions until a leaf is reached, assign the class as the outcome of a majority vote of samples in that leaf node.

The “concept decision tree” (?)

A decision tree is an algorithm used for classification and regression tasks. It models decisions as a tree-like structure, where each internal node represents a decision baes on the value of a particular feature, each branch represents the outcome of that decision, and each leaf node represents the final decision or output (class label or numerical label)

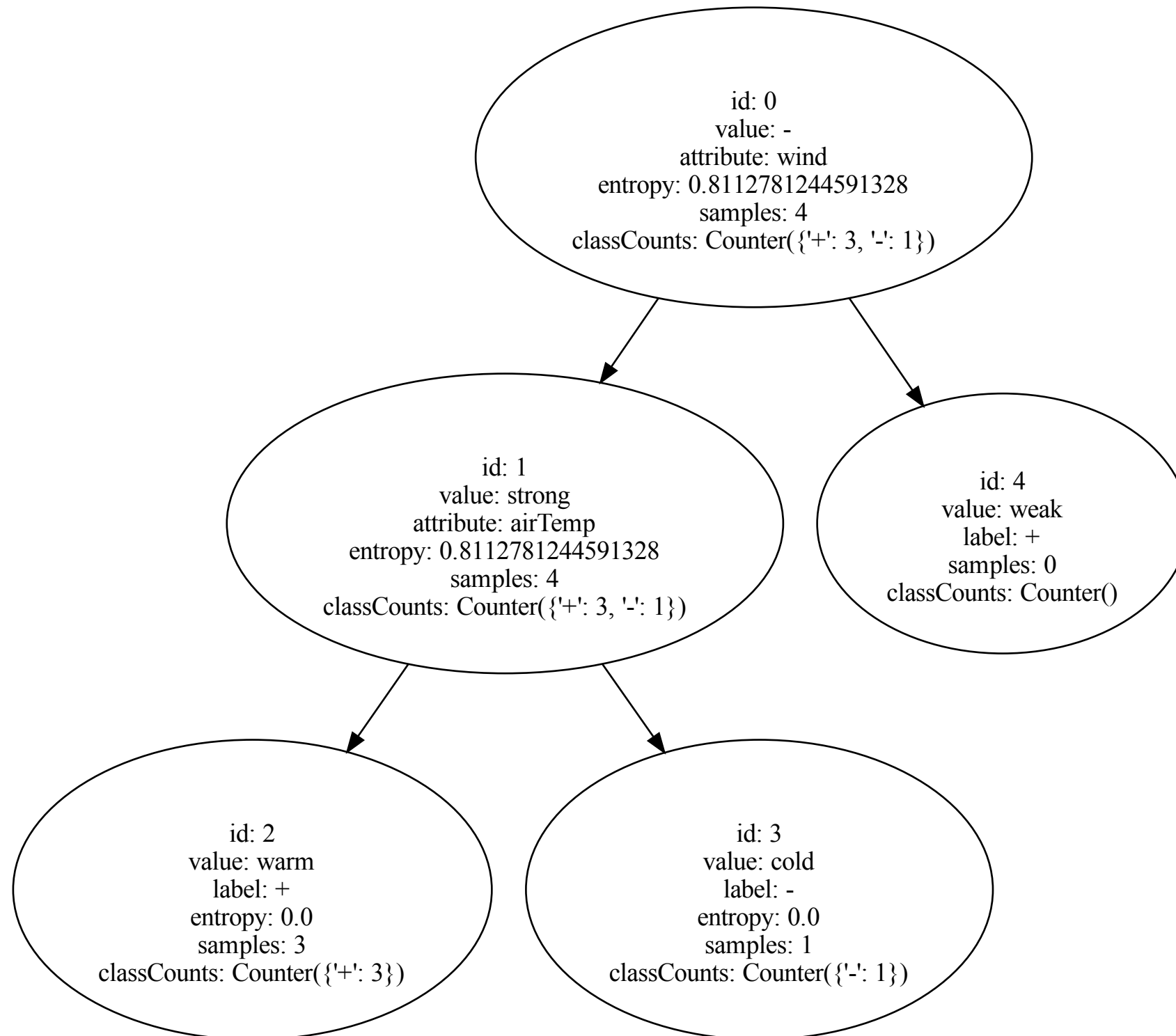
The process of determining how to split the data at each internal node is crucial. Common criteria include Gini impurity for classification tasks and mean squared error for regression tasks. The goal is to maximize the homogeneity of the data in each branch.



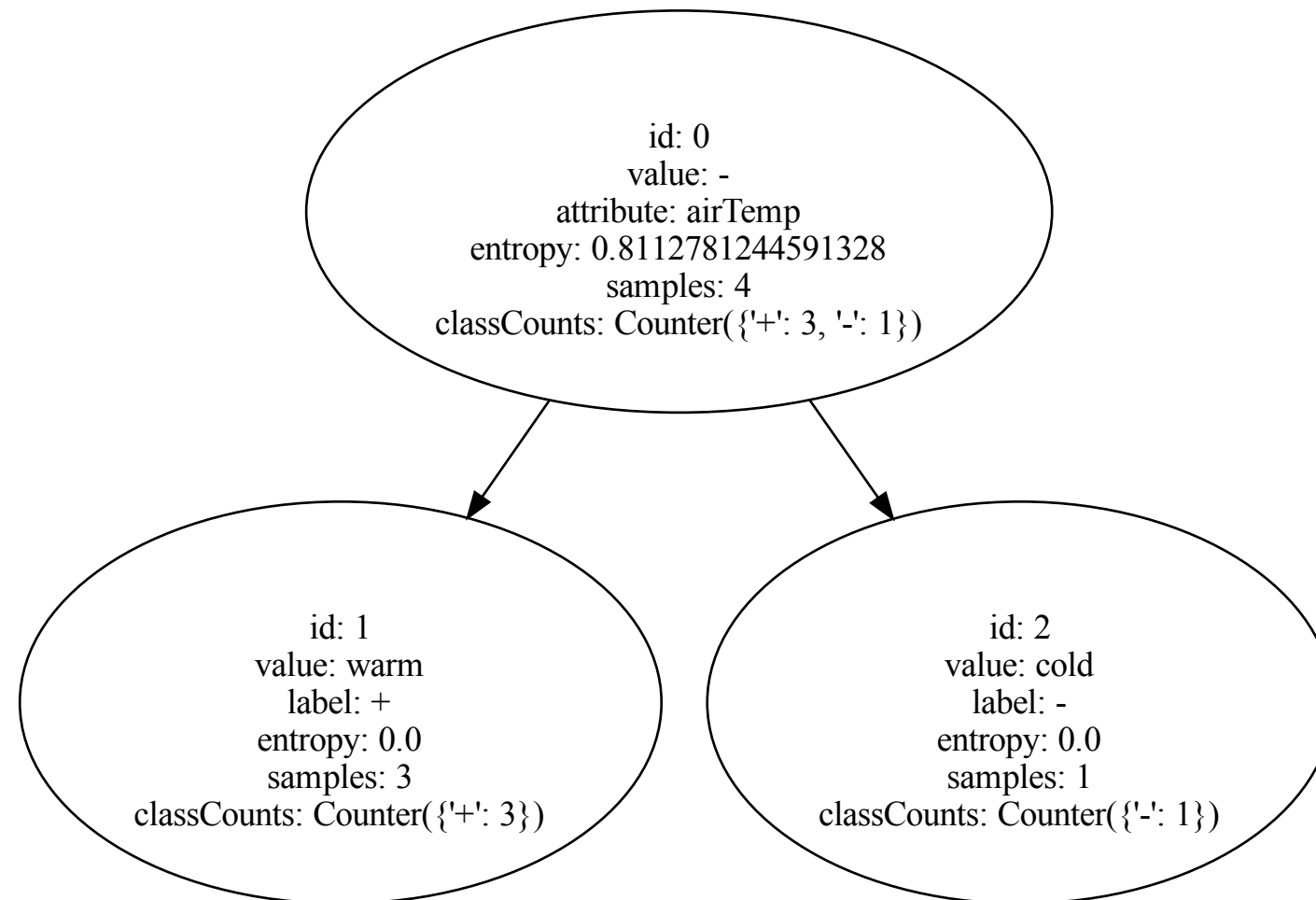
Root node = entire dataset

The process of decision-making is recursive. At each internal node, the algorithm selects the best feature and split point to partition the data. This process is repeated for each subset at the child nodes until a stopping condition is met, such as a maximum depth or a minimum number of instances in a node.

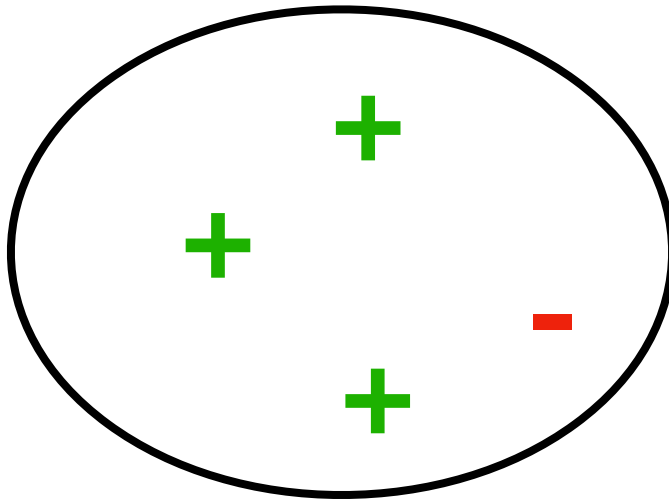
Can we do better?



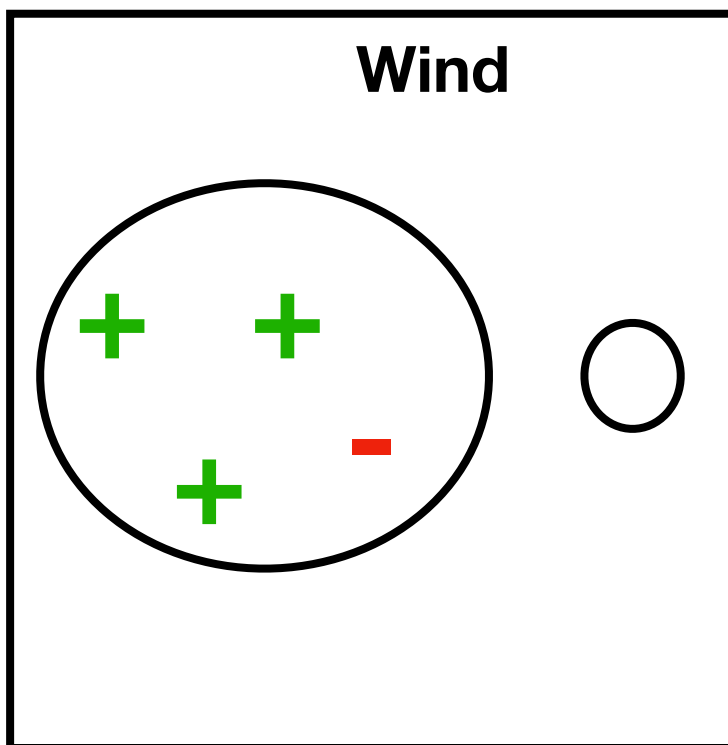
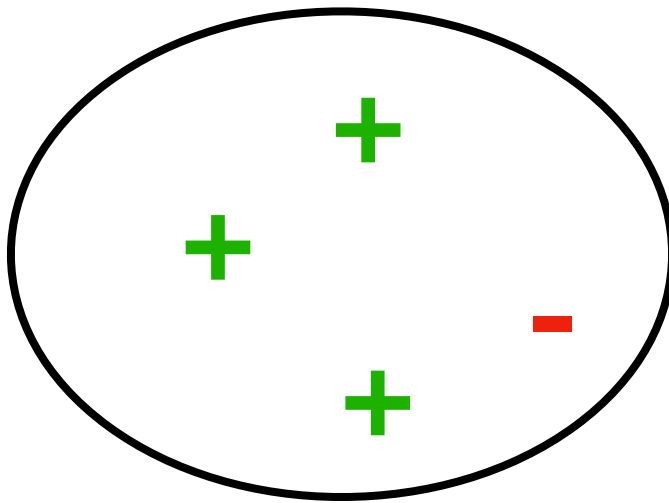
Can we do better?



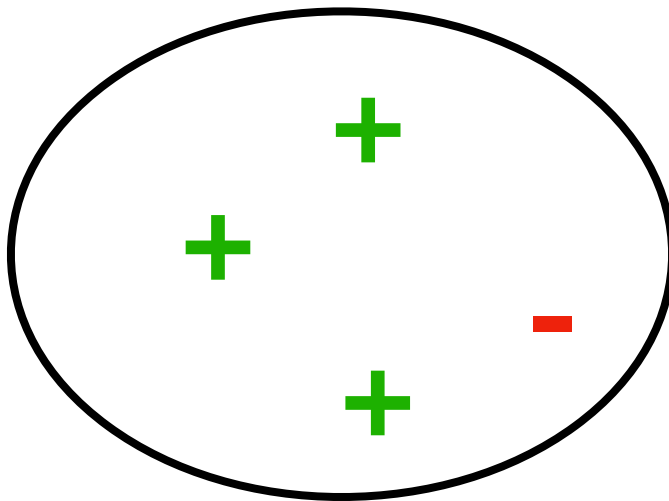
Improving decisions: Finding the best split attribute intuitively



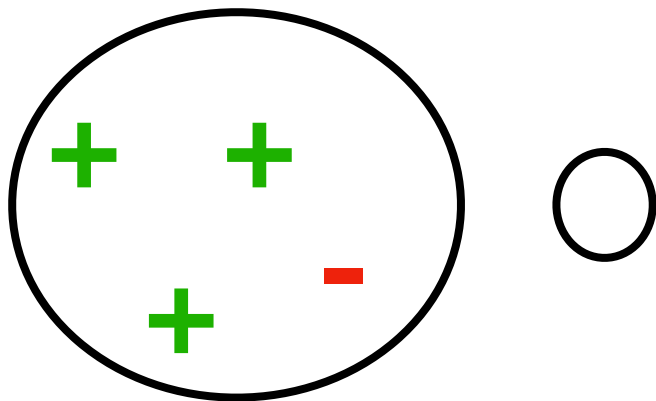
Improving decisions: Finding the best split attribute intuitively



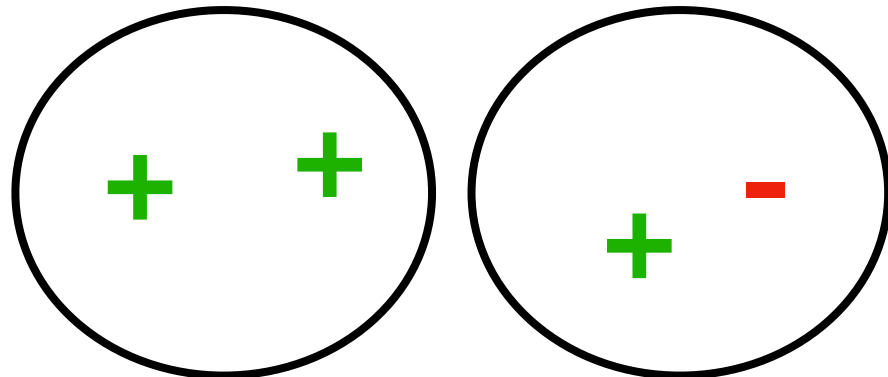
Improving decisions: Finding the best split attribute intuitively



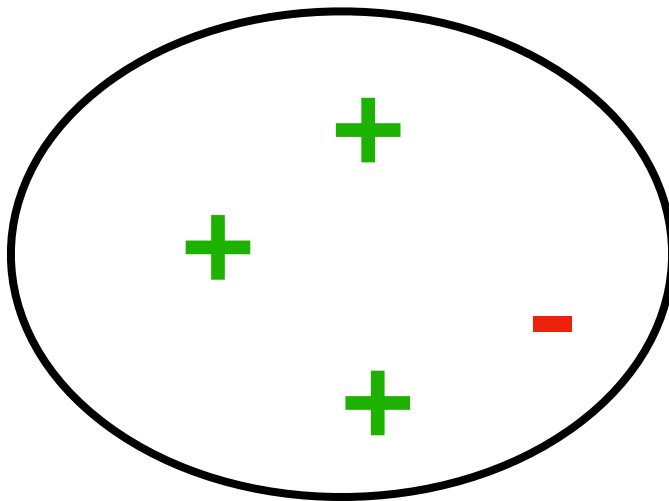
Wind



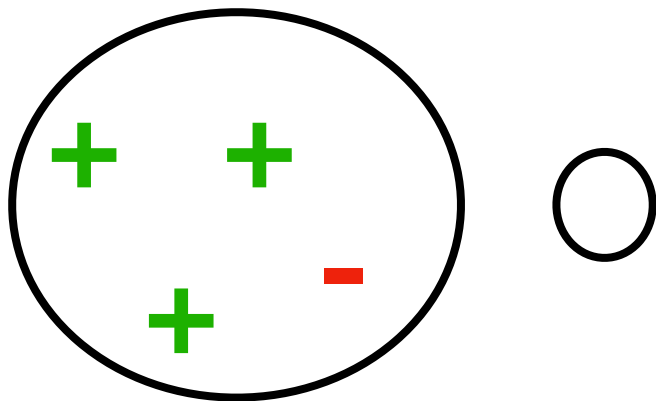
Forecast



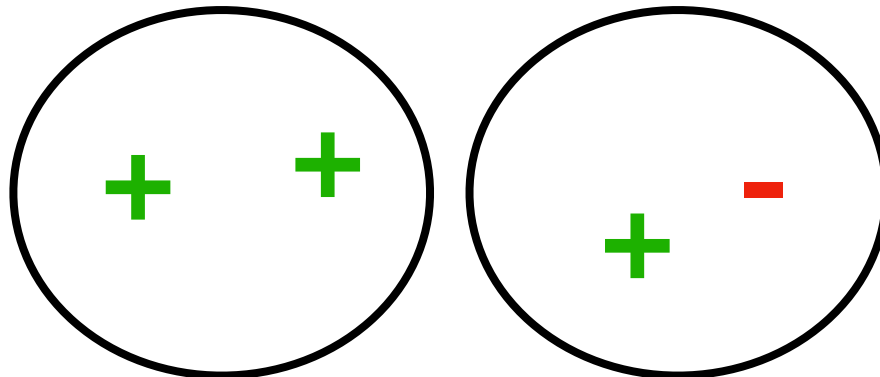
Improving decisions: Finding the best split attribute intuitively



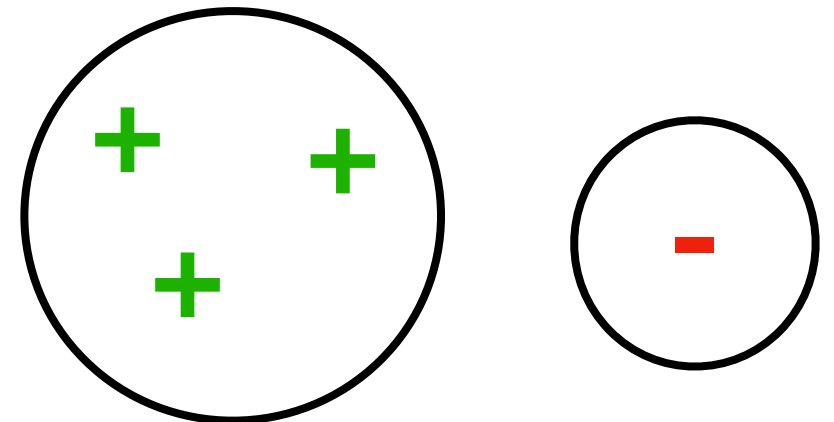
Wind



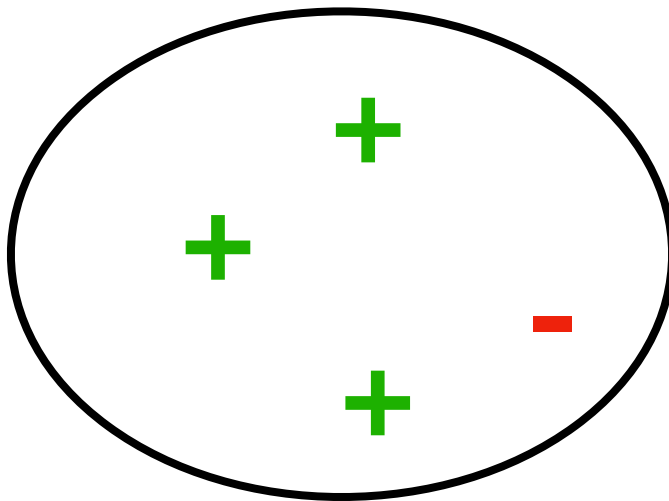
Forecast



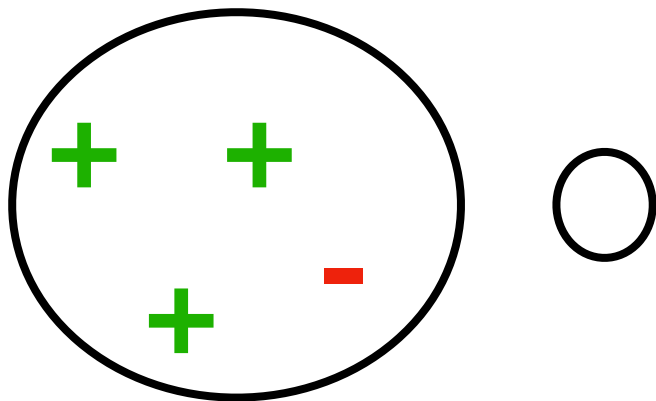
AirTemp



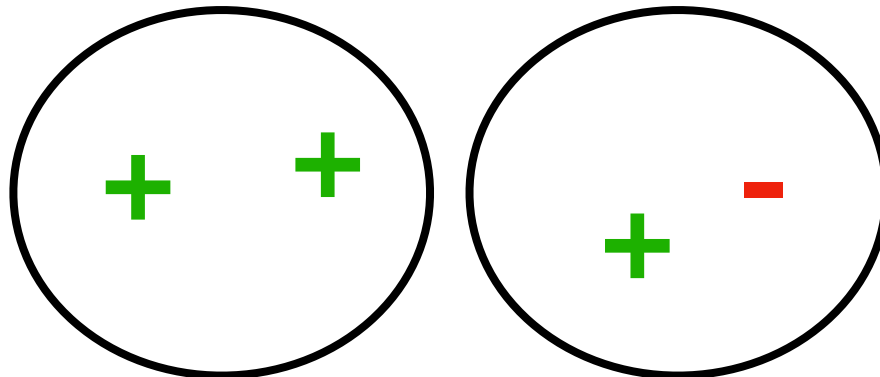
Improving decisions: Finding the best split attribute intuitively



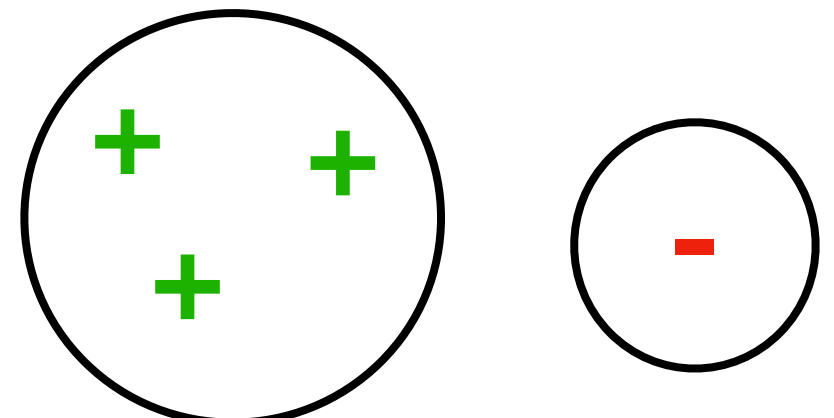
Wind



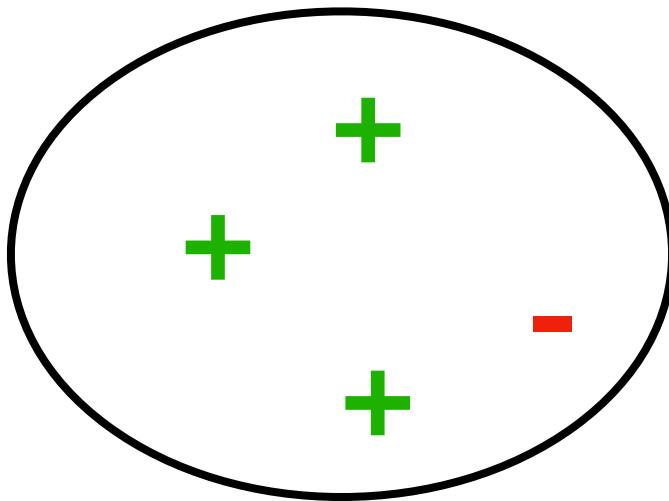
Forecast



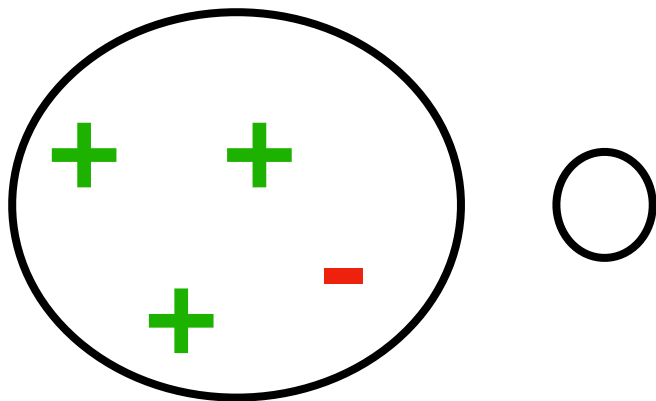
AirTemp



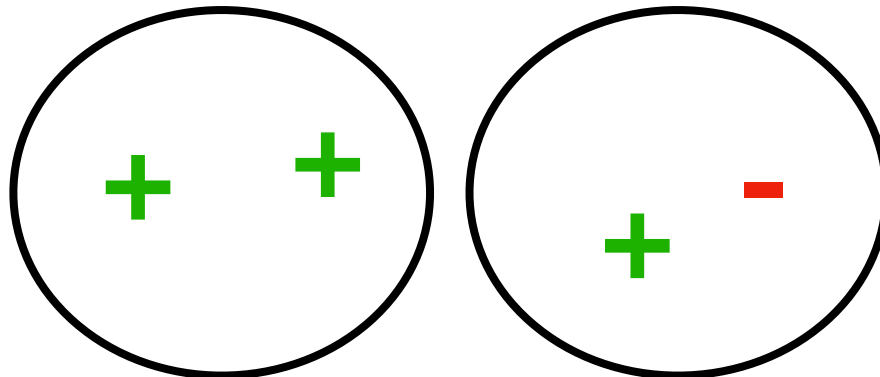
Improving decisions: Finding the best split attribute intuitively



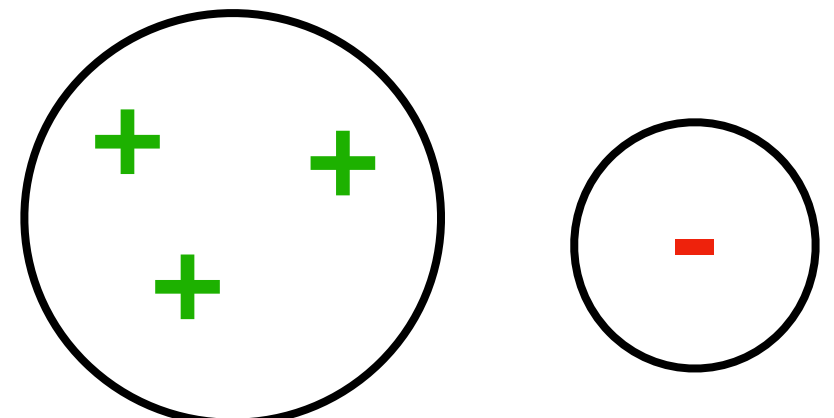
Wind



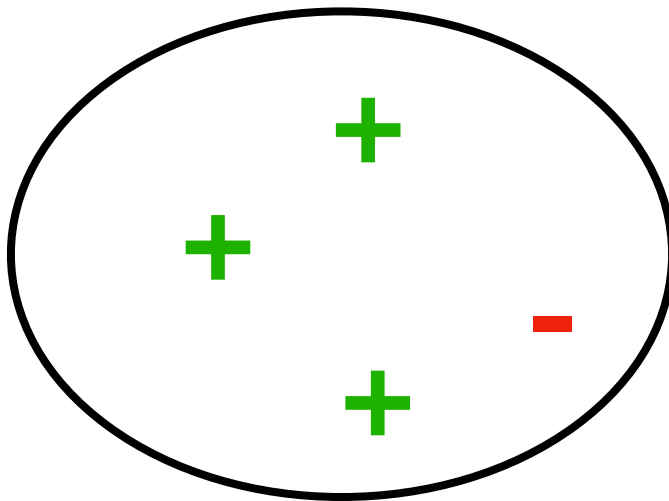
Forecast



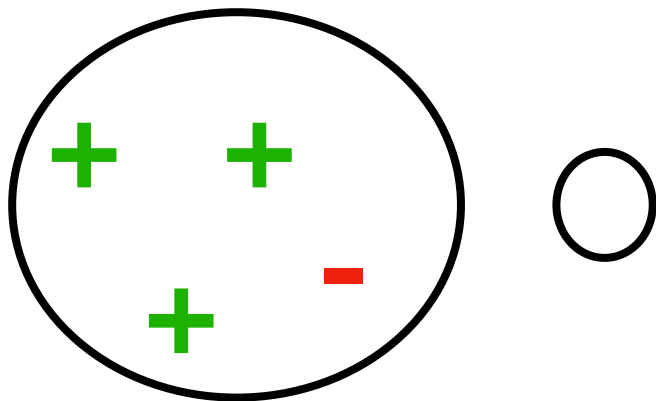
AirTemp



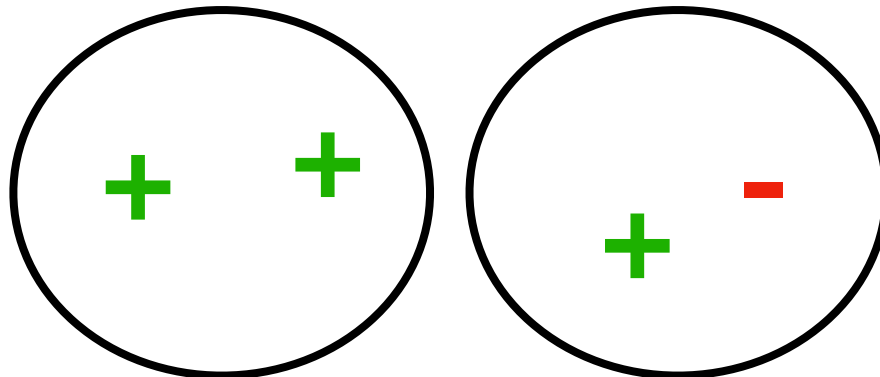
Improving decisions: Finding the best split attribute intuitively



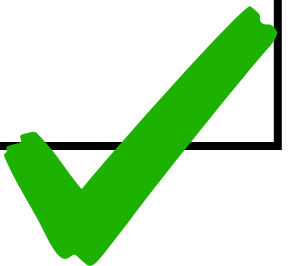
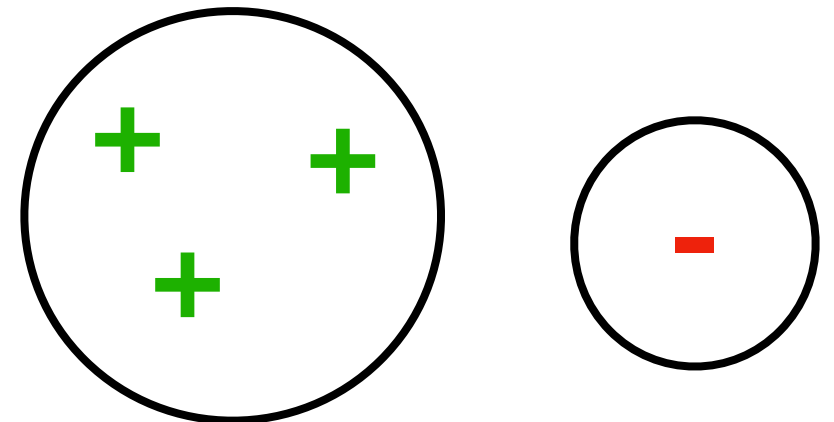
Wind



Forecast



AirTemp



Finding the best split attribute for DT - Gini Impurity Index

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution.

- **Minimising** Gini Impurity (standard for **CART** implementation in SciKitLearn)
 - Index is based on the probability for getting a sample of a particular class when drawing from the set - the higher this is for one single class, the better (index will be lower)
 - Is computed for set S with K classes and p_i the probability for class _{i} as

$$I_{Gini}(S) = \sum_{i=1}^K p_i \sum_{j \neq i} p_j = \sum_{i=1}^K p_i (1 - p_i) = \sum_{i=1}^K (p_i - p_i^2) = \sum_{i=1}^K p_i - \sum_{i=1}^K p_i^2 = 1 - \sum_{i=1}^K p_i^2$$

Finding the best split attribute for DT - Information Gain

Information Gain measures the effectiveness of an attribute in classifying the data.

- **Maximising** Information Gain (standard for ID3)

- Finding (over the possible splits) the highest possible reduction of entropy between the current node (data set) and its children (the data subsets), if this split were chosen
- Entropy (Information) $I(S)$ of a data set S with elements belonging to K classes $class_i$ and $p(class_i)$ the probability of observing $class_i$ in S :

$$I(S) = - \sum_i^K p(class_i) \cdot \log_2(p(class_i)) = \sum_i^K p(class_i) \cdot \log_2 \left(\frac{1}{p(class_i)} \right)$$

- The Information Gain $G(S,A)$ of a split of S at Attribute A is then the reduction in entropy we get as the difference between the current entropy $I(S)$ and the entropies of the subsets $I(S_v)$ over the different values of A , weighted by the proportion of samples ending up in the respective subset

$$G(S, A) = I(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} I(S_v)$$

The attribute with the highest Information Gain is the one that reduces uncertainty about the class labels the most. By choosing attributes that lead to the greatest reduction in entropy, decision trees can efficiently partition the data into homogeneous subsets.

Information Gain example

Colour	Size	Shape	edible?
yellow	small	round	+
yellow	small	round	-
green	small	irregular	+
green	large	irregular	-
yellow	large	round	+
yellow	small	round	+
yellow	small	round	+
yellow	small	round	+
green	small	round	-
yellow	large	round	-
yellow	large	round	+
yellow	large	round	-
yellow	large	round	-
yellow	large	round	-
yellow	small	irregular	+
yellow	large	irregular	+

- 16 items in S
- 9 items belong to class '+', 7 to '-'
- Entropy in S (using proportions as probabilities)

$$\begin{aligned}
 I(S) &= - \sum_i^K p(class_i) \cdot \log_2(p(class_i)) \\
 &= - \left(\frac{9}{16} \cdot \log_2 \left(\frac{9}{16} \right) + \frac{7}{16} \cdot \log_2 \left(\frac{7}{16} \right) \right) \\
 &\approx 0.9887
 \end{aligned}$$

- Three attributes, which splits best?

Information Gain example

testing the attributes

Colour	Size	Shape	edible?
yellow	small	round	+
yellow	small	round	-
green	small	irregular	+
green	large	irregular	-
yellow	large	round	+
yellow	small	round	+
yellow	small	round	+
yellow	small	round	+
green	small	round	-
yellow	large	round	-
yellow	large	round	+
yellow	large	round	-
yellow	large	round	-
yellow	large	round	-
yellow	small	irregular	+
yellow	large	irregular	+

- 13 samples in S_{yellow} , 3 in S_{green} .
- In S_{yellow} , 8 belong to class '+', 5 belong to '-'
- In S_{green} , 1 belongs to class '+', 2 belong to '-'
- Entropies for S_{yellow} and S_{green}

$$I(S_{yellow}) = - \left(\frac{8}{13} \cdot \log_2 \left(\frac{8}{13} \right) + \frac{5}{13} \cdot \log_2 \left(\frac{5}{13} \right) \right)$$

$$\approx 0.9612$$

$$I(S_{green}) = - \left(\frac{2}{3} \cdot \log_2 \left(\frac{2}{3} \right) + \frac{1}{3} \cdot \log_2 \left(\frac{1}{3} \right) \right)$$

$$\approx 0.9183$$

- Information Gain:

$$0.9887 - \left(\frac{13}{16} \cdot 0.9612 + \frac{3}{16} \cdot 0.9183 \right) \approx 0.0355$$

Information Gain example

testing the attributes

Colour	Size	Shape	edible?
yellow	small	round	+
yellow	small	round	-
green	small	irregular	+
green	large	irregular	-
yellow	large	round	+
yellow	small	round	+
yellow	small	round	+
yellow	small	round	+
green	small	round	-
yellow	large	round	-
yellow	large	round	+
yellow	large	round	-
yellow	large	round	-
yellow	large	round	-
yellow	small	irregular	+
yellow	large	irregular	+

- 8 samples in S_{small} , 8 in S_{large} .
- In S_{small} , 6 belong to class '+', 2 belong to '-'
- In S_{large} , 3 belong to class '+', 5 belong to '-'
- Entropies for S_{small} and S_{large}

$$\begin{aligned}
 I(S_{small}) &= - \left(\frac{6}{8} \cdot \log_2 \left(\frac{6}{8} \right) + \frac{2}{8} \cdot \log_2 \left(\frac{2}{8} \right) \right) \\
 &\approx 0.8113
 \end{aligned}$$

$$\begin{aligned}
 I(S_{large}) &= - \left(\frac{3}{8} \cdot \log_2 \left(\frac{3}{8} \right) + \frac{5}{8} \cdot \log_2 \left(\frac{5}{8} \right) \right) \\
 &\approx 0.9544
 \end{aligned}$$

- Information Gain:

$$0.9887 - \left(\frac{1}{2} \cdot 0.8113 + \frac{1}{2} \cdot 0.9544 \right) \approx 0.1058$$

Information Gain example

testing the attributes

Colour	Size	Shape	edible?
yellow	small	round	+
yellow	small	round	-
green	small	irregular	+
green	large	irregular	-
yellow	large	round	+
yellow	small	round	+
yellow	small	round	+
yellow	small	round	+
green	small	round	-
yellow	large	round	-
yellow	large	round	+
yellow	large	round	-
yellow	large	round	-
yellow	large	round	-
yellow	small	irregular	+
yellow	large	irregular	+

- 12 samples in S_{round} , 4 in $S_{irregular}$.
- In S_{round} , 6 belong to class '+', 6 belong to '-'
- In $S_{irregular}$, 3 belong to class '+', 1 belongs to '-'
- Entropies for S_{round} and $S_{irregular}$

$$I(S_{round}) = - \left(\frac{6}{12} \cdot \log_2 \left(\frac{6}{12} \right) + \frac{6}{12} \cdot \log_2 \left(\frac{6}{12} \right) \right)$$

$$= -1 \cdot \log_2 \left(\frac{1}{2} \right) = 1.0$$

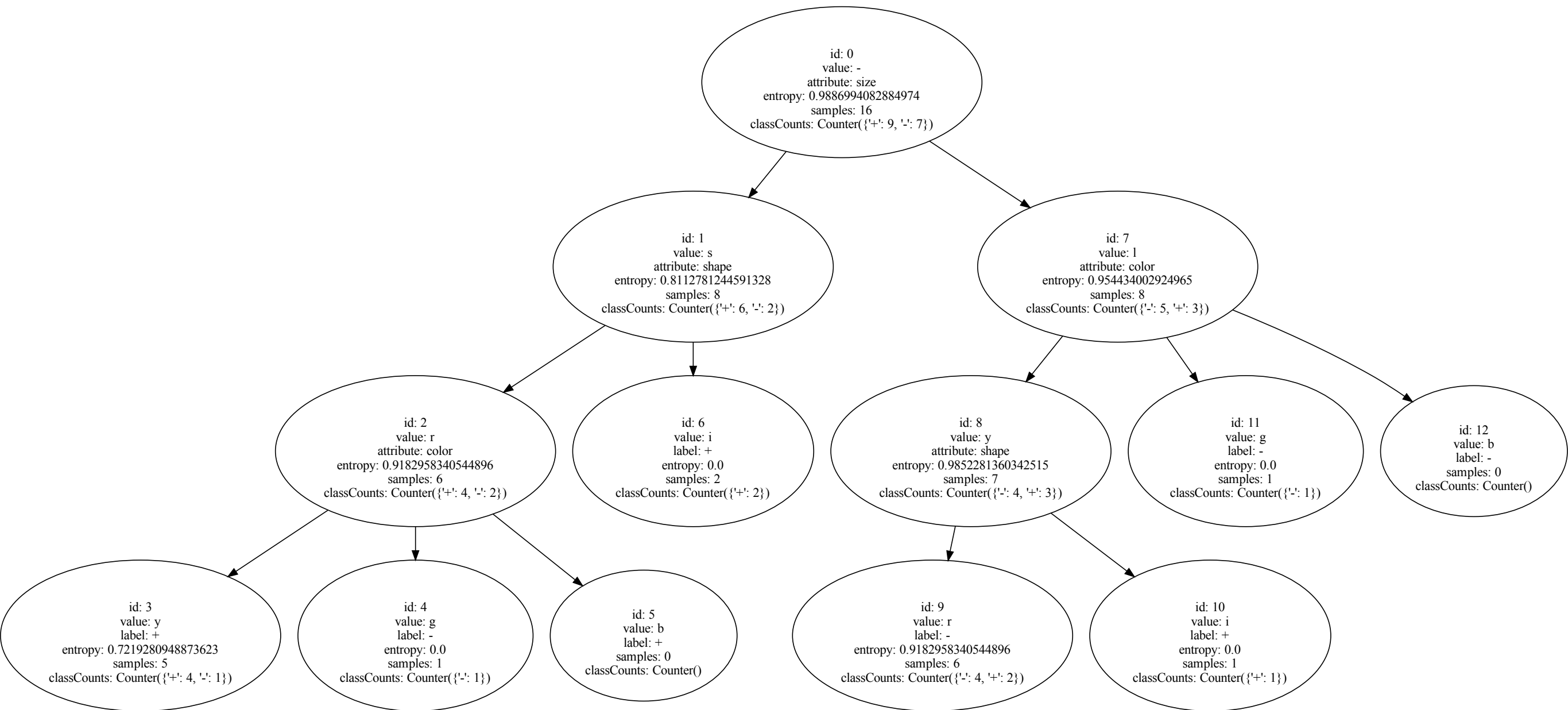
$$I(S_{irregular}) = - \left(\frac{3}{4} \cdot \log_2 \left(\frac{3}{4} \right) + \frac{1}{4} \cdot \log_2 \left(\frac{1}{4} \right) \right)$$

$$\approx 0.8113$$

- Information Gain:


$$0.9887 - \left(\frac{3}{4} \cdot 1.0 + \frac{1}{4} \cdot 0.8113 \right) \approx 0.0359$$

ID3-Decision Tree based on maximum Information Gain



Revisiting the concept learning problem:

Regression (value prediction)

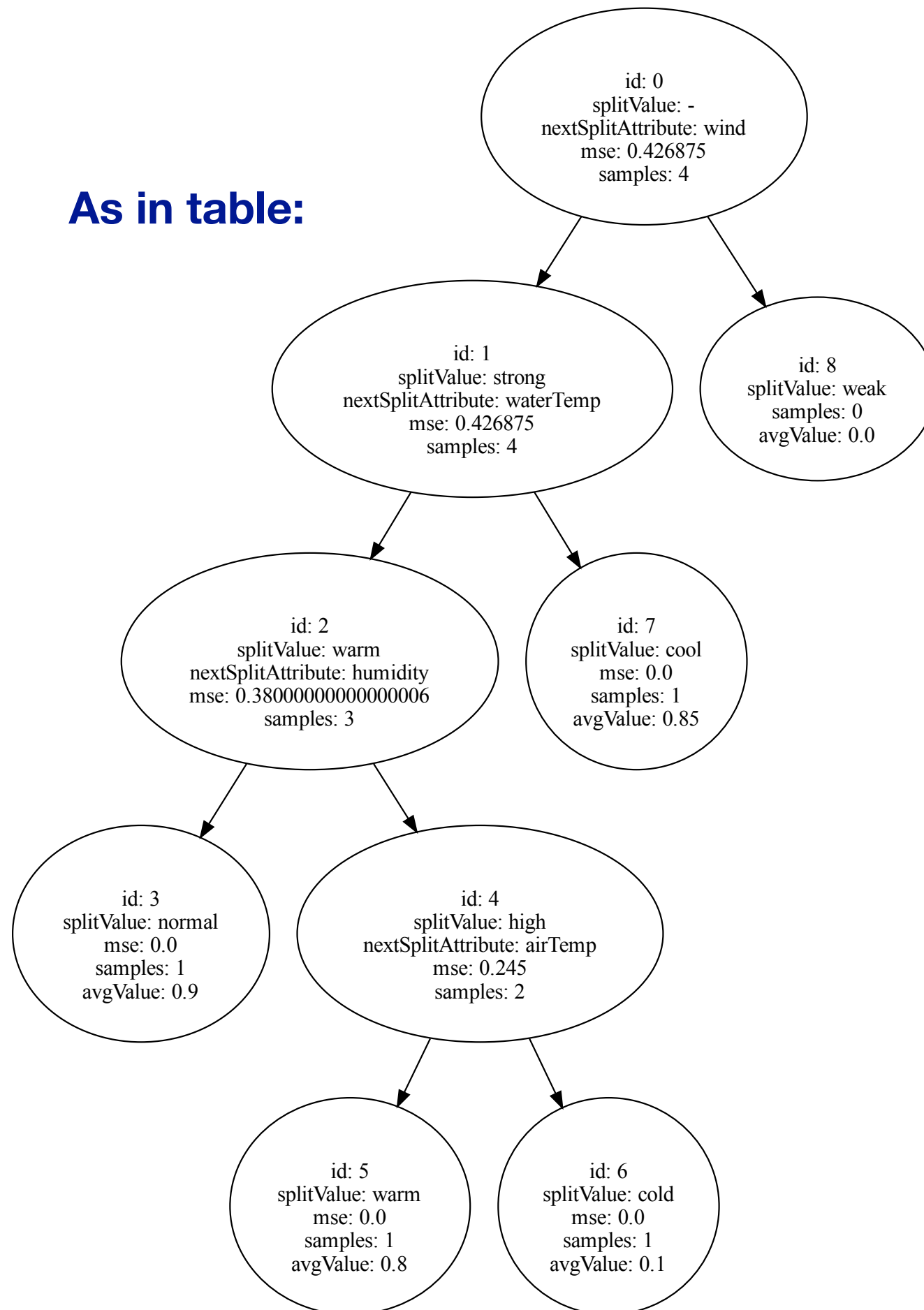


Example	Wind	Water	Humidity	AirTemp	Sky	Forecast	<u>EnjoySport?</u>
1	Strong	Warm	Normal	Warm	Sunny	Same	0.9
2	Strong	Warm	High	Warm	Sunny	Same	0.8
3	Strong	Warm	High	Cold	Rainy	Change	0.1
4	Strong	Cool	High	Warm	Sunny	Change	0.85

- We can predict a day's value (probability that one enjoys sports) with a regression tree
- Go through the examples attribute by attribute and split the data into subsets according to their attribute value
- Stop, when there are no more attributes to test or a stop criterion is reached
- When using the tree (predicting the value for an unseen sample), follow the decisions until a leaf node is reached. The predicted value is then the average of the values in the leaf node.

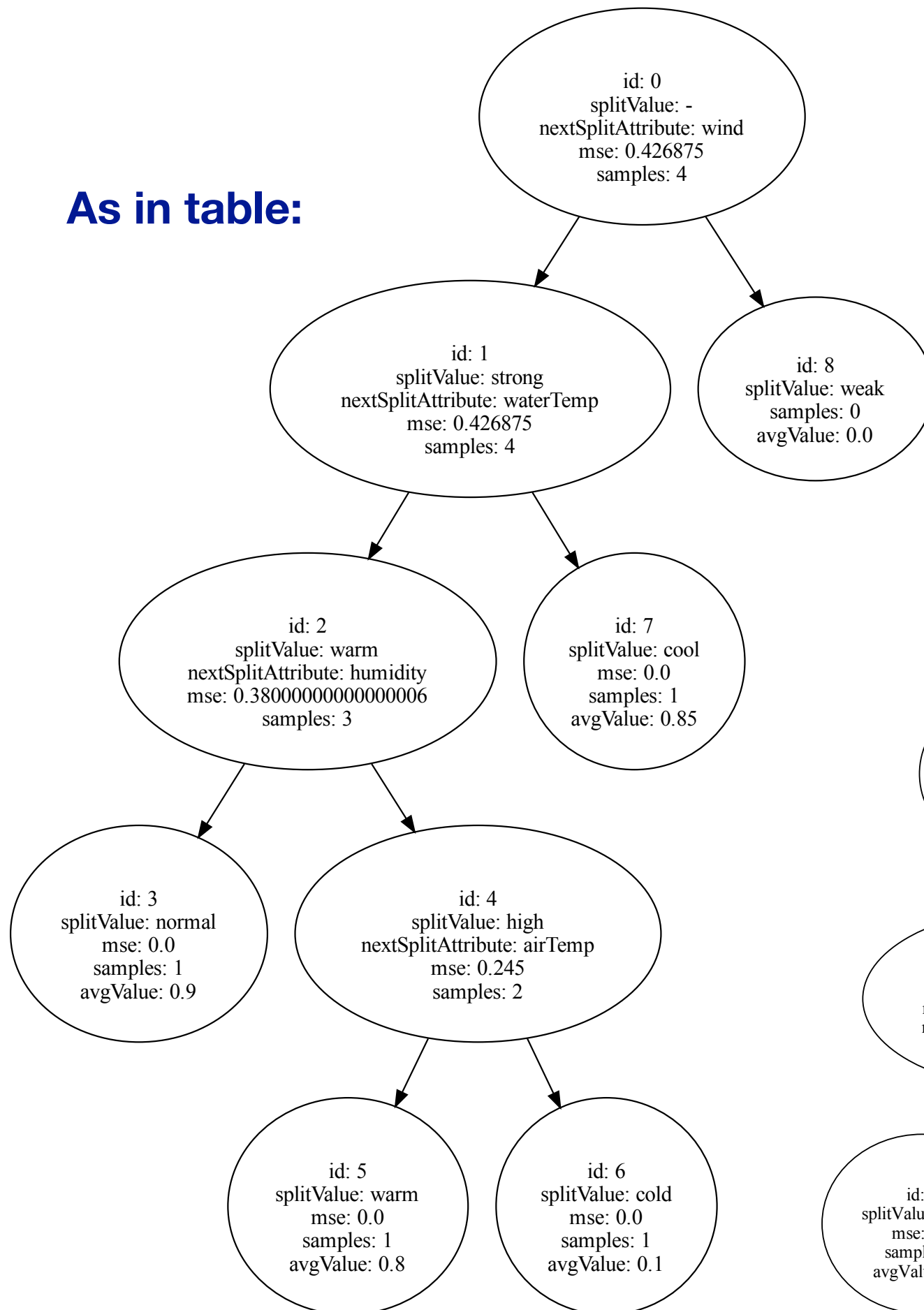
The “concept regression tree” (?)

As in table:

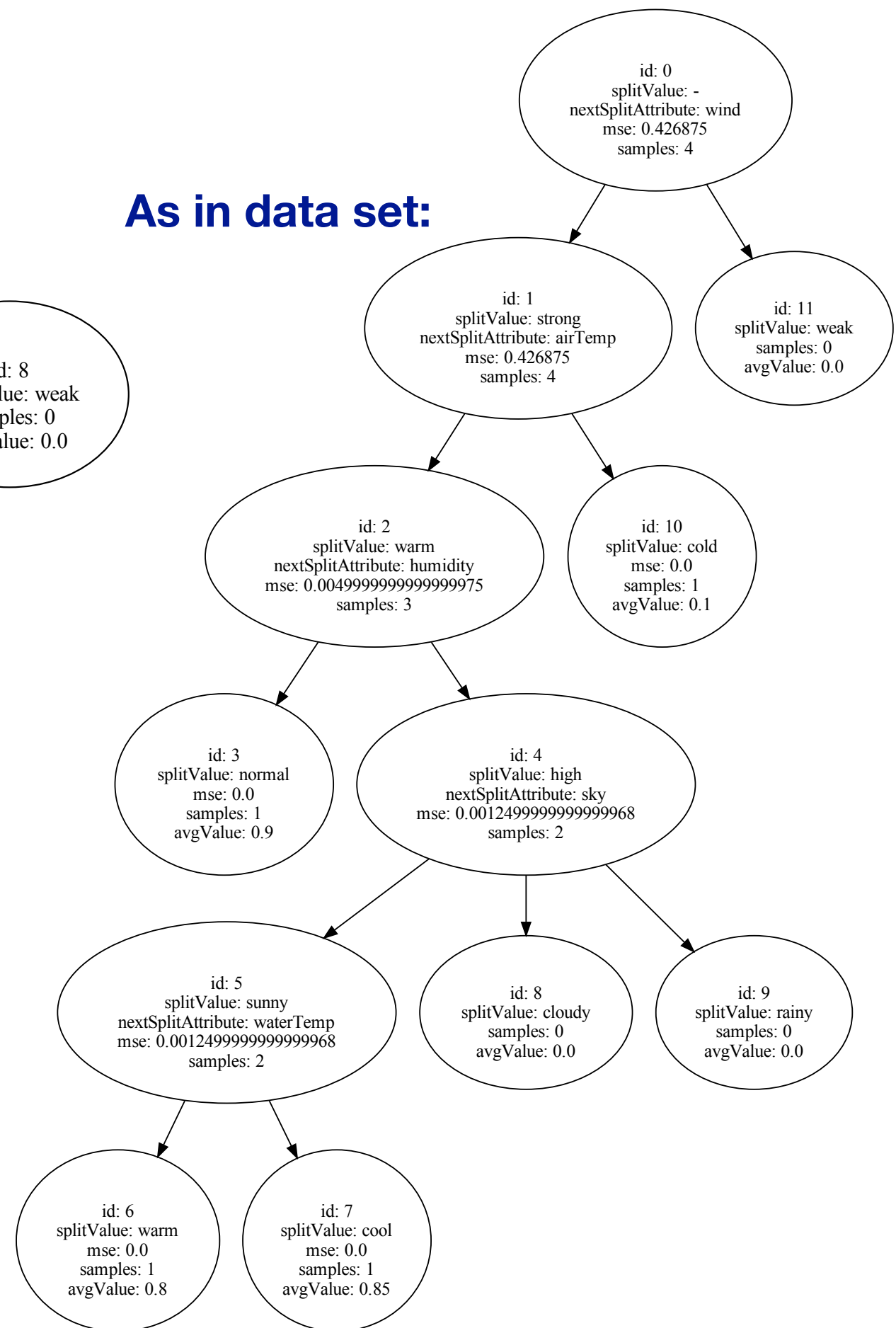


The “concept regression tree” (?)

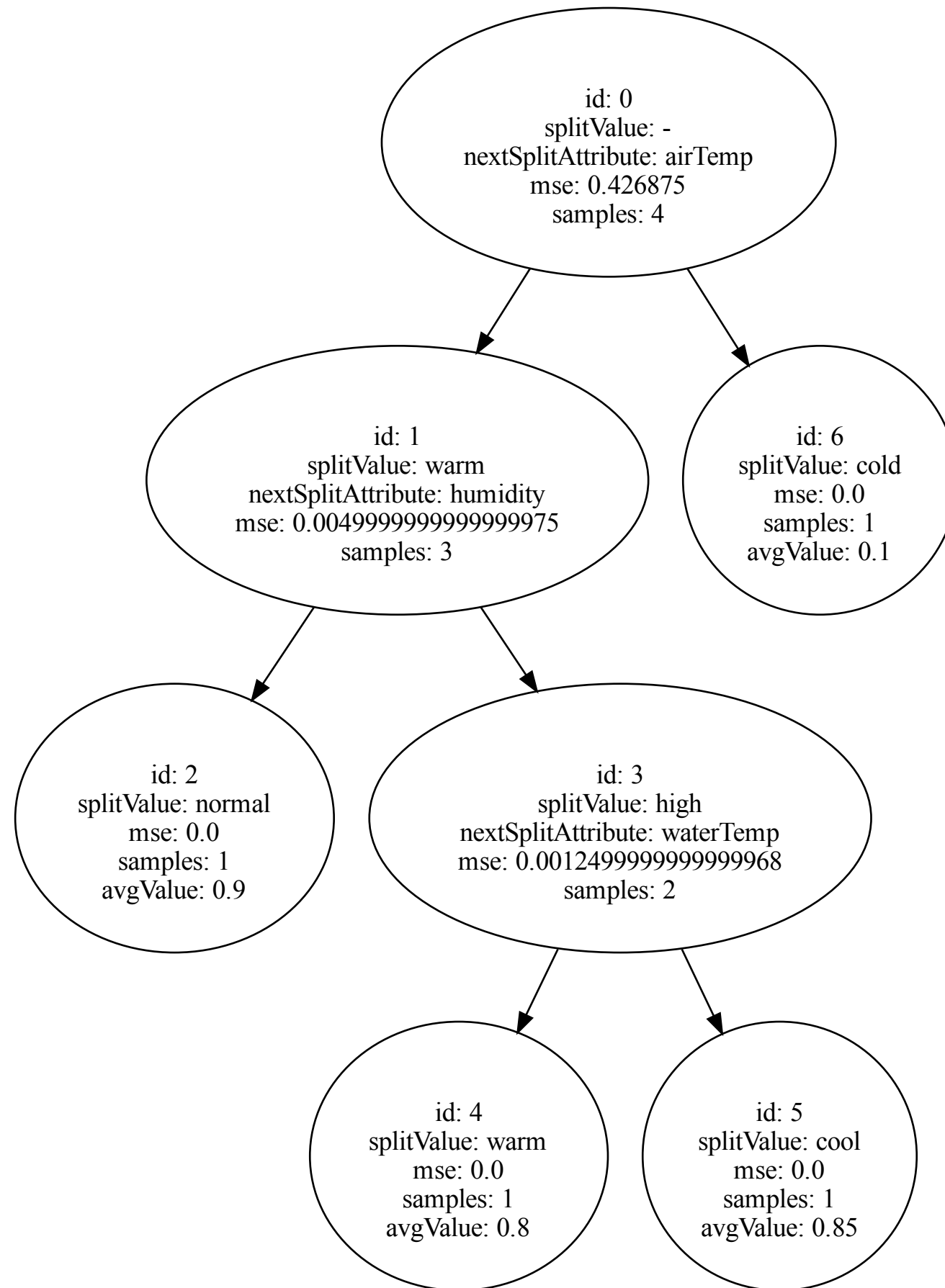
As in table:



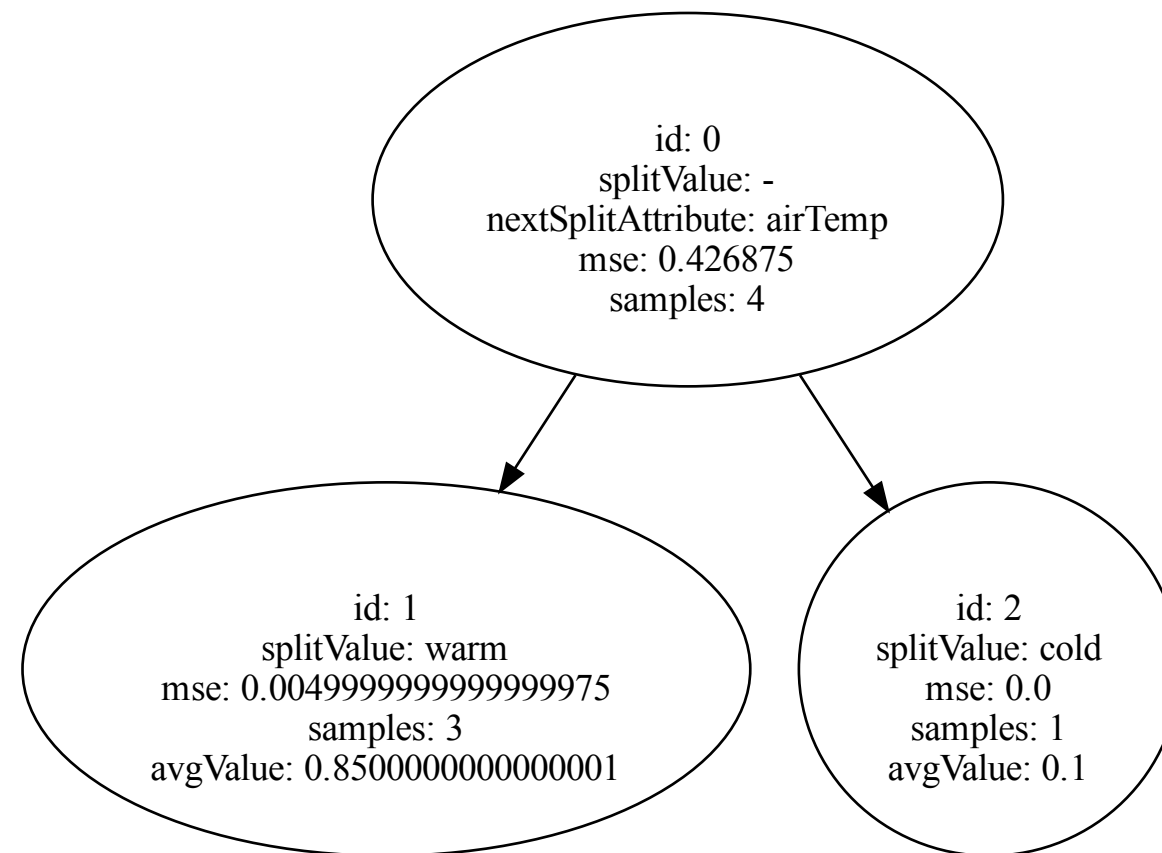
As in data set:



Can we do better?



Can we do better?



Finding the best split attribute for RT

- Minimising Mean Squared Error (MSE) (still working with ID3, i.e. multiple value evaluation)

Find (over the possible splits) the **lowest** possible *overall MSE*, if this split were chosen

- **MSE** of a data set S with n elements \mathbf{x}_i , respective target values y_i and the average value (prediction value) \hat{y} :

$$MSE(S) = \sum_{i:\mathbf{x}_i \in S}^n (y_i - \hat{y}(S))^2$$

- The *overall MSE* $oMSE(S, A)$ of a split of S at attribute A is then the **sum over the MSEs of the subsets (S_v)** over the V different values of A :

$$oMSE(S, A) = \sum_{S_v = \{x \in S : value(A) = v\}}^V MSE(S_v)$$

- The **best split attribute** A' for S is then the **A that minimises $oMSE(S, A)$** :

$$splitA(S) = \underset{A}{\operatorname{argmin}}(oMSE(S, A))$$

- In case of a strictly binary tree (CART), find also the best split value v' for A' (see Lindholm et al, page 29)

Issues with Decision and Regression Trees

- Consider a new example with which you want to modify your tree...
- Consider a very unbalanced data set (like the concept learning example)
- Consider really unseen examples - how well does the tree generalise?

Today's summary

- Discussed the basics of Decision and Regression Trees
- Mentioned the connection to other instance-base approaches like k-NN
- Reading:
 - Lecture slides lecture 4, 2018
 - Mitchell, chapter 3, Decision Trees
 - Lindholm et al, chapter 2

Outlook on programming assignment 5

- Regression: California housing prices
- Two notebooks / Python scripts, making two parts of the assignment
 - Part 1: “Tutorial” / walkthrough notebook to explore RTs (SciKitLearn implementation, CART based) - and ensemble methods
 - Part 2: Implement split routine (finding best split attribute) for ID3-based, own implementation, evaluate in comparison to the SciKitLearn version
- Reading:
 - SciKitLearn documentation
 - ID3 description, e.g. Wikipedia