Decision Trees

Applied machine learning (EDAN95)
Lecture 03 — Decision Trees
2019–11–11
Elin A. Topp

Goodfellow chapter 3, Géron chapter on DTs
Information Gain lecture of F. Aiolli: https://www.math.unipd.it/~aiolli/corsi/0708/IR/Lez12.pdf
various sources

Today’s agenda

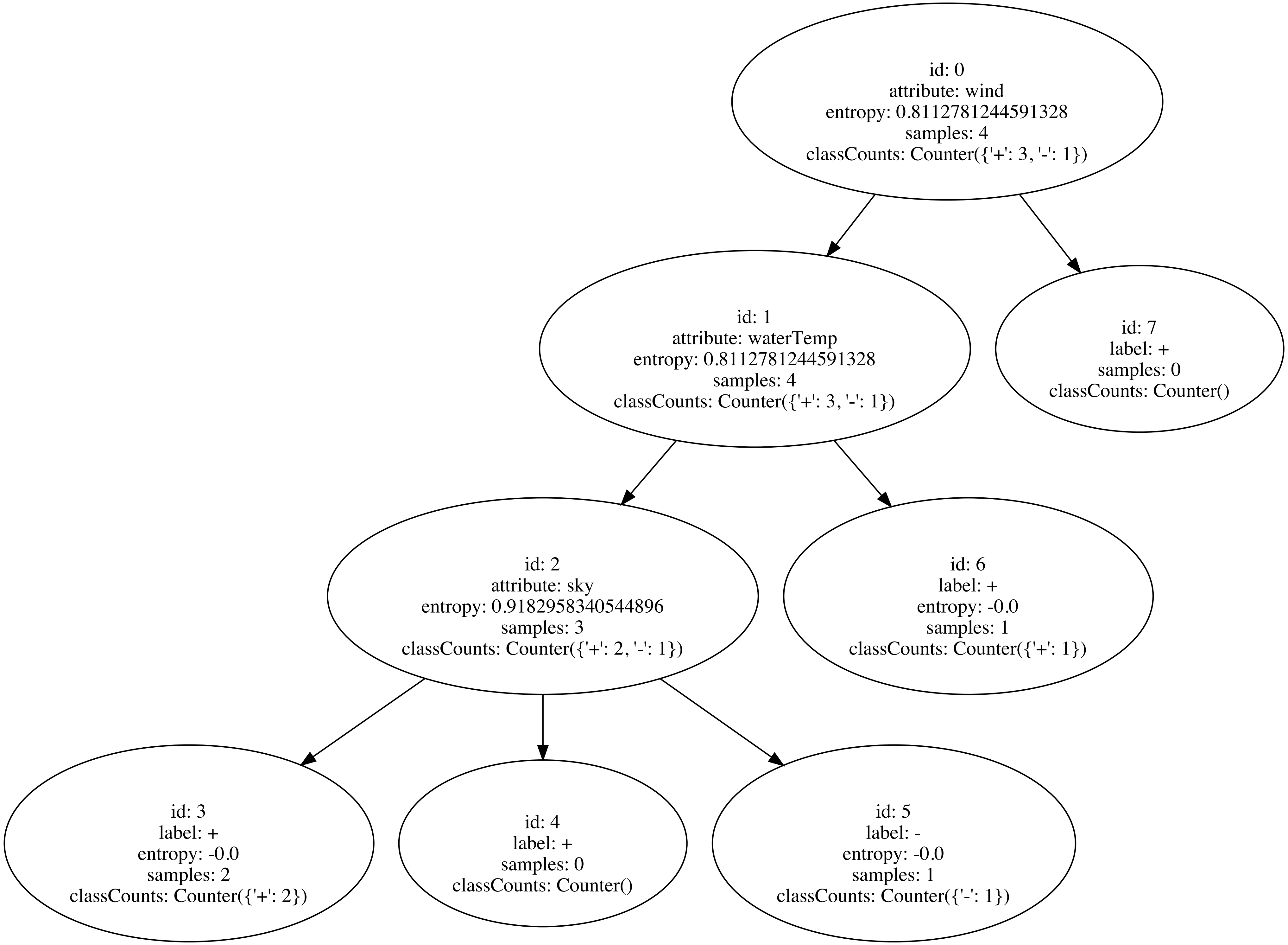
- Decision Trees
- (Recap) Information theory (Goodfellow chapter 3) and some Probability Theory
- Metrics (outlook)

Revisiting the concept learning problem

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

The “concept decision tree” (?)



id: 0
attribute: wind
entropy: 0.8112781244591328
samples: 4
classCounts: Counter({'+': 3, '-': 1})

id: 1
attribute: waterTemp
entropy: 0.8112781244591328
samples: 4
classCounts: Counter({'+': 3, '-': 1})

id: 7
label: +
samples: 0
classCounts: Counter()

id: 2
attribute: sky
entropy: 0.9182958340544896
samples: 3
classCounts: Counter({'+': 2, '-': 1})

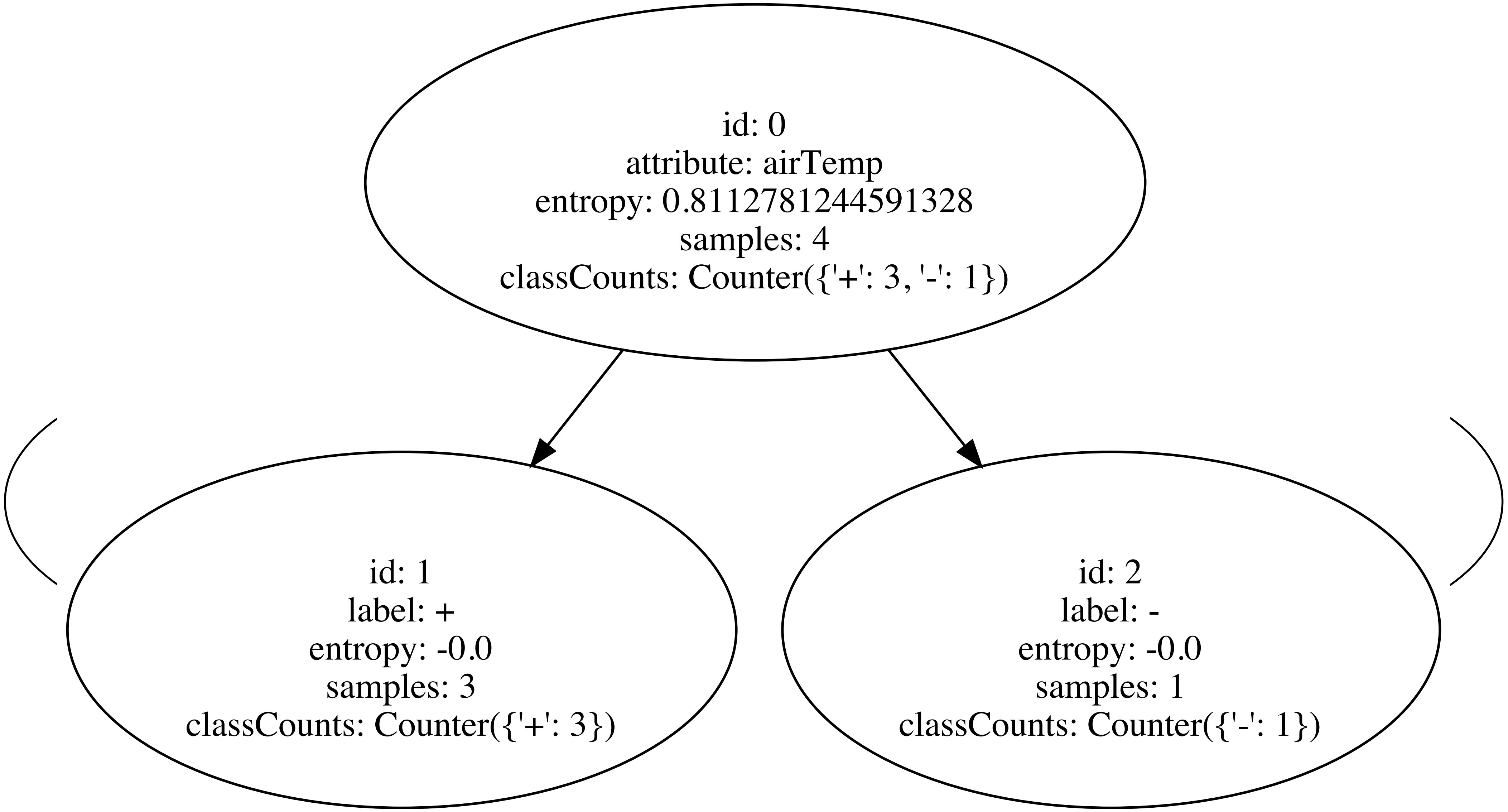
id: 6
label: +
entropy: -0.0
samples: 1
classCounts: Counter({'+': 1})

id: 3
label: +
entropy: -0.0
samples: 2
classCounts: Counter({'+': 2})

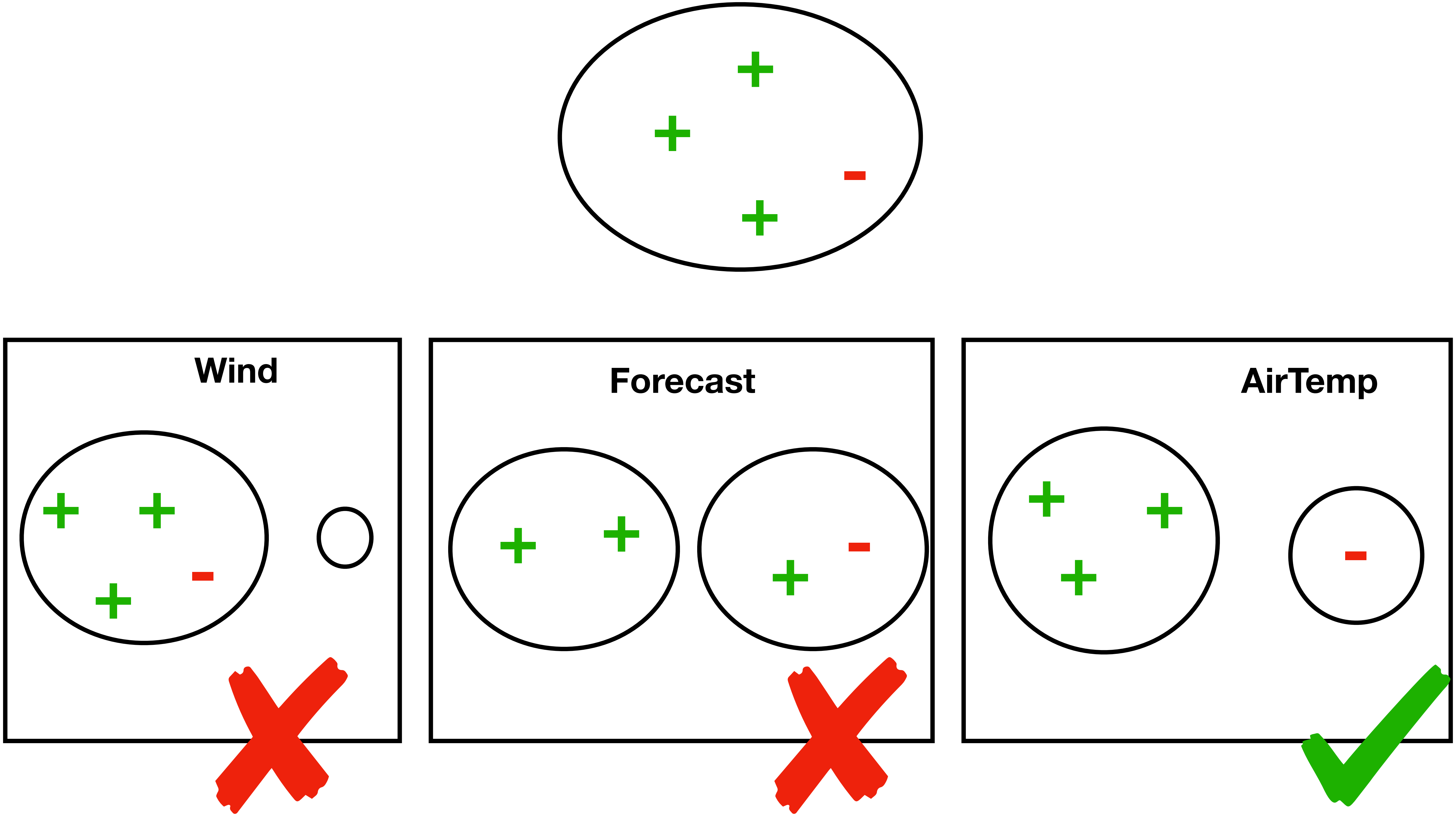
id: 4
label: +
samples: 0
classCounts: Counter()

id: 5
label: -
entropy: -0.0
samples: 1
classCounts: Counter({'-': 1})

Can we do better?


id: 0
attribute: airTemp
entropy: 0.8112781244591328
samples: 4
classCounts: Counter({'+': 3, '-': 1})
id: 1
label: +
entropy: -0.0
samples: 3
classCounts: Counter({'+': 3})
id: 2
label: -
entropy: -0.0
samples: 1
classCounts: Counter({'-': 1})

Information Gain intuitively



Probability and Information Theory

- Goodfellow, chapter 3
  - Probability basics and Information Theory (entropy, information gain, impurity) (this lecture)
  - More probabilistic representation (uncertainty) and graphical models → Lecture 10/11
- Other material:
  - Russel, Norvig, AI - A modern Approach, chapters 13/14/15
  - EDAF70, Lecture slides on Probabilistic Representation VT2018/VT2019
  - Murphy, Machine Learning - A Probabilistic Approach
  - Géron, Hands-on ML, Material on Github (see Lab 1) (spec Decision Trees)

Finding the best split attribute

- 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:*

$$I(S) \quad = \quad -\sum_{i}^{K} p(class_i) \cdot log_2(p(class_i))$$

  - 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 *(S$_v$)* over the different values of *A*, weighted by the proportion of samples ending up in the respective subset

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

- Minimising Gini Impurity (standard for CART implementation in SciKitLearn)
  - for set *S* with *K* classes and *p$_i$* the probability for *class$_i$*

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

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)

$$I(S) \quad = \quad -\sum_{i}^{K} p(class_i) \cdot log_2(p(class_i))$$

$$= -\left(\frac{9}{16} \cdot log_2\left(\frac{9}{16}\right) + \frac{7}{16} \cdot log_2\left(\frac{7}{16}\right)\right)$$

$$\approx 0.9887$$

- 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}

$$I(S_{small}) = -\left(\frac{6}{8} \cdot log_2\left(\frac{6}{8}\right) + \frac{2}{8} \cdot log_2\left(\frac{2}{8}\right)\right)$$

$$\approx 0.8113$$

$$I(S_{large}) = -\left(\frac{3}{8} \cdot log_2\left(\frac{3}{8}\right) + \frac{5}{8} \cdot log_2\left(\frac{5}{8}\right)\right)$$

$$\approx 0.9544$$

- 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 $S_{irregular}$

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

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

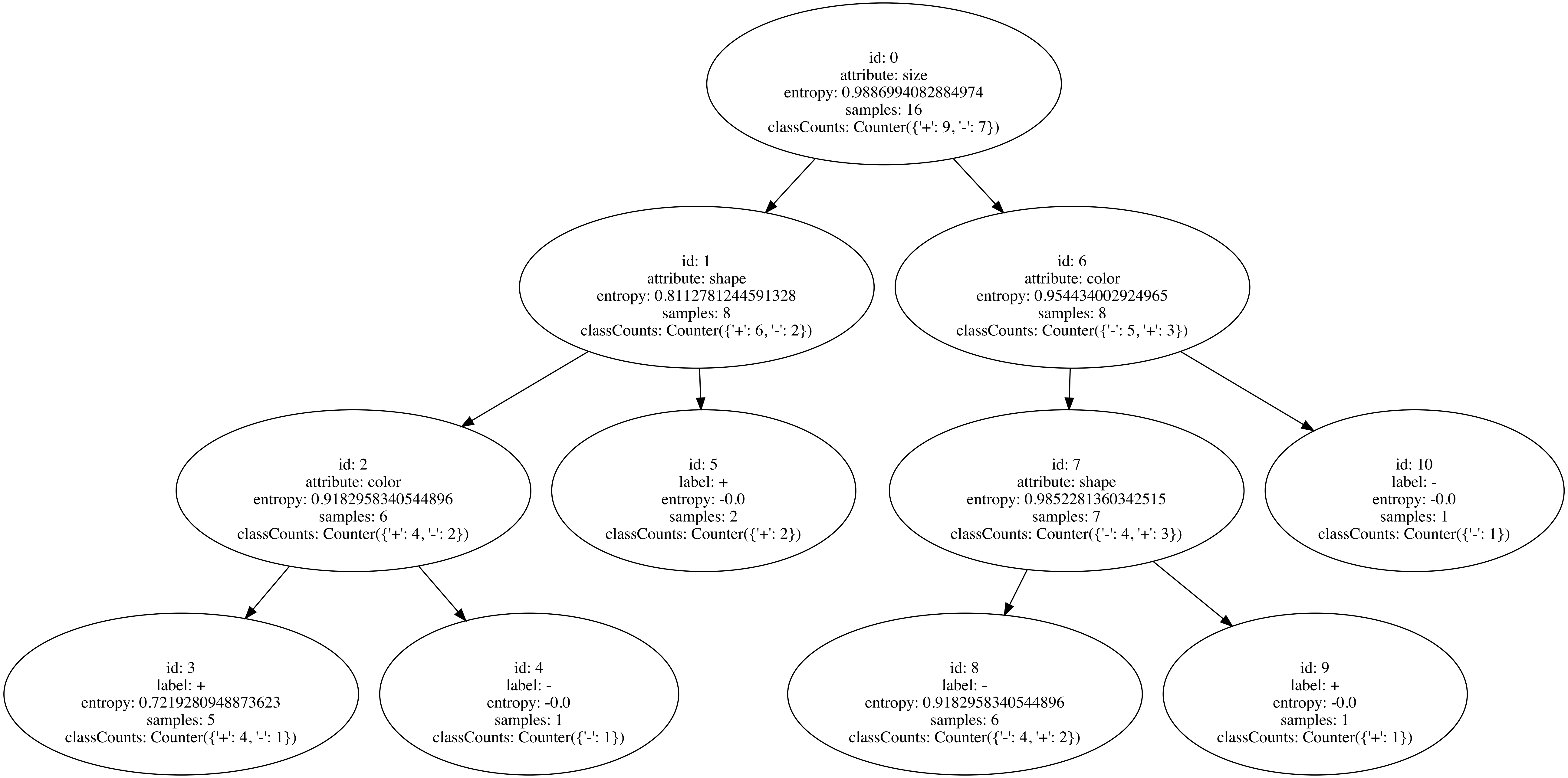
$$I(S_{irregular}) = -\left(\frac{3}{4} \cdot log_2\left(\frac{3}{4}\right) + \frac{1}{4} \cdot log_2\left(\frac{1}{4}\right)\right)$$

$$\approx 0.8113$$

- 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



id: 0
attribute: size
entropy: 0.9886994082884974
samples: 16
classCounts: Counter({'+': 9, '-': 7})

id: 1
attribute: shape
entropy: 0.8112781244591328
samples: 8
classCounts: Counter({'+': 6, '-': 2})

id: 6
attribute: color
entropy: 0.954434002924965
samples: 8
classCounts: Counter({'-': 5, '+': 3})

id: 2
attribute: color
entropy: 0.9182958340544896
samples: 6
classCounts: Counter({'+': 4, '-': 2})

id: 5
label: +
entropy: -0.0
samples: 2
classCounts: Counter({'+': 2})

id: 7
attribute: shape
entropy: 0.9852281360342515
samples: 7
classCounts: Counter({'-': 4, '+': 3})

id: 10
label: -
entropy: -0.0
samples: 1
classCounts: Counter({'-': 1})

id: 3
label: +
entropy: 0.7219280948873623
samples: 5
classCounts: Counter({'+': 4, '-': 1})

id: 4
label: -
entropy: -0.0
samples: 1
classCounts: Counter({'-': 1})

id: 8
label: -
entropy: 0.9182958340544896
samples: 6
classCounts: Counter({'-': 4, '+': 2})

id: 9
label: +
entropy: -0.0
samples: 1
classCounts: Counter({'+': 1})

How good is the tree?

- Normally, we split our entire data set in some (larger) part for training, and a (smaller) part for testing / validation. Why?
- After training, all samples in the test set are run through the tree (just following the branch that corresponds to the sample’s value for the respective attribute in each decision until a leaf is reached). We can then compare the (ground truth) classification with the prediction delivered by the tree.
- By doing so, we can construct a matrix *C*, where rows correspond to the true (expected) values, and columns to what the classifier predicted (OBS: make sure that both sets follow the same order when doing this!). We get then

$$C_{ij} \quad = \quad |expected\_in\_i\_but\_predicted\_as\_j|$$

Ideally, *C* is a diagonal matrix. Is that even possible with the given data set?

Colour	Size	Shape	edible?
yellow	small	round	+
yellow	small	round	-
…	…	…	…

Other concepts in measuring quality

- True positives *TP*, false positives *FP*, true negatives *TN*, and false negatives *FN*
- How many of all samples that were classified as “X” were actually correct?

$$precision = \frac{TP}{TP + FP}$$

- How many of all samples that should have been classified as “X” were actually found as “X”?

$$recall = \frac{TP}{TP + FN}$$

- How many samples were classified correctly as “X” or “not X”?

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Combining precision and recall to weight them directly against each other:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Other concepts in measuring quality

Test data for the example tree:

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

Target: $[\,+\,,-\,,+\,,-\,,+\,]$ Prediction: $[\,+\,,+\,,+\,,-\,,-\,]$ Confusion matrix: $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$

Classification report $$\left[\begin{array}{ccccc} & precision & recall & f1-score & support \\ + & 0.67 & 0.67 & 0.67 & 3 \\ - & 0.50 & 0.50 & 0.50 & 2 \\ accuracy & & & 0.60 & 5 \\ macro\,avg & 0.58 & 0.58 & 0.58 & 5 \\ weighted\,avg & 0.60 & 0.60 & 0.60 & 5 \end{array}\right]$$

Issues with Decision 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

- Introduced Decision Trees
- Recap Information Theory / Entropy
- Information Gain / Impurity

- Introduced some metrics to understand the quality of a learning (classification) approach

- Reading:
    - Géron, Hands-on ML, Material on Github (spec Decision Trees)
    - Mitchell, chapter 3, Decision Trees