Reading 3

☆ K-Nearest Algorithm Video

- a super simple way to classify data

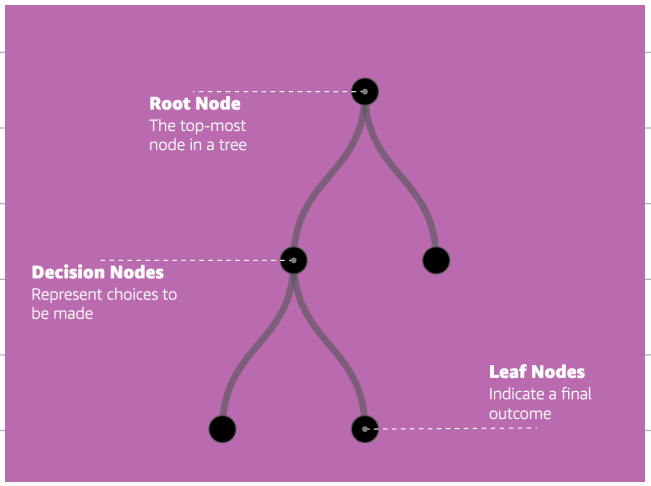
- K is the number of neighbors we look at to determine unknown data point.
  - ↳ if K is too small it can be subject to outliers
  - ↳ if K is too large it can be very challenging to classify unknown data into a small category.

- Step 1: Start with a dataset with known categories.
- Step 2: Add a new cell, with unknown category, to the plot.
  - ↳ We want to classify this cell
- Step 3: We classify the new cell by looking at the nearest annotated cells. (i.e. neighbors)

- The data used for the initial clustering is called "training data."

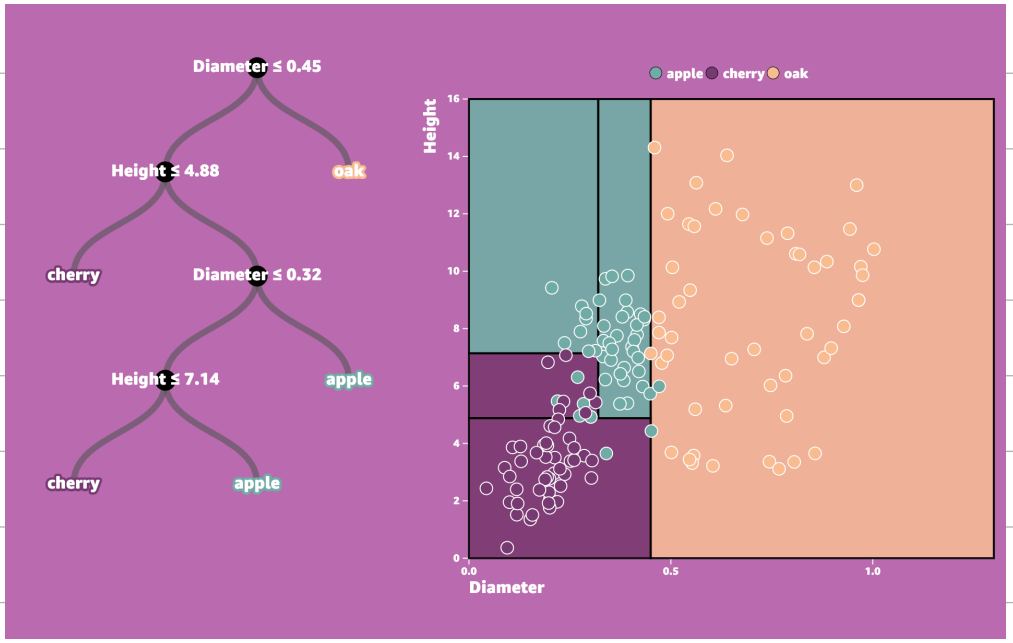
☆ Blogpost (MLU-Explain): Decision Trees

- Decision Trees are widely used algorithms for supervised machine learning.
  - ↳ they work for both regression and classification problems.



- To train a Decision Tree from data means to figure out the order in which the decisions should be assembled from the root to the leaves.

- Ex:



- Entropy measures the amount of information of some variable or event.
  - ↳ used to identify regions consisting of a large number of similar (pure) or dissimilar (impure) elements.

Given a certain set of events that occur with probabilities $(p_1, p_2, \ldots, p_n)$, the total entropy H can be written as the negative sum of weighted probabilities:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

The quantity has a number of interesting properties:

Entropy Properties

1. $H = 0$ only if all but one of the p_i are zero, this one having the value of 1. Thus the entropy vanishes only when there is no uncertainty in the outcome, meaning that the sample is completely unsurprising.
2. H is maximum when all the p_i are equal. This is the most uncertain, or 'impure', situation.
3. Any change towards the equalization of the probabilities $(p_1, p_2, \ldots, p_n)$ increases H.

- The entropy can be used to quantify the impurity of a collection of labeled data points.
  - ↳ A node containing multiple classes is impure. (high entropy)
  - ↳ A node including only one class is pure. (zero entropy)

- An alternative to the entropy for the construction of Decision Trees is the Gini impurity.

- **Information Gain**: measures the amount of information we can gain.
  - ↳ The idea is to subtract from the entropy of our data before the split the entropy of each possible partition thereafter.

- The core algorithm to calculate information gain is **ID3**. It calculates the difference in entropy at each depth!

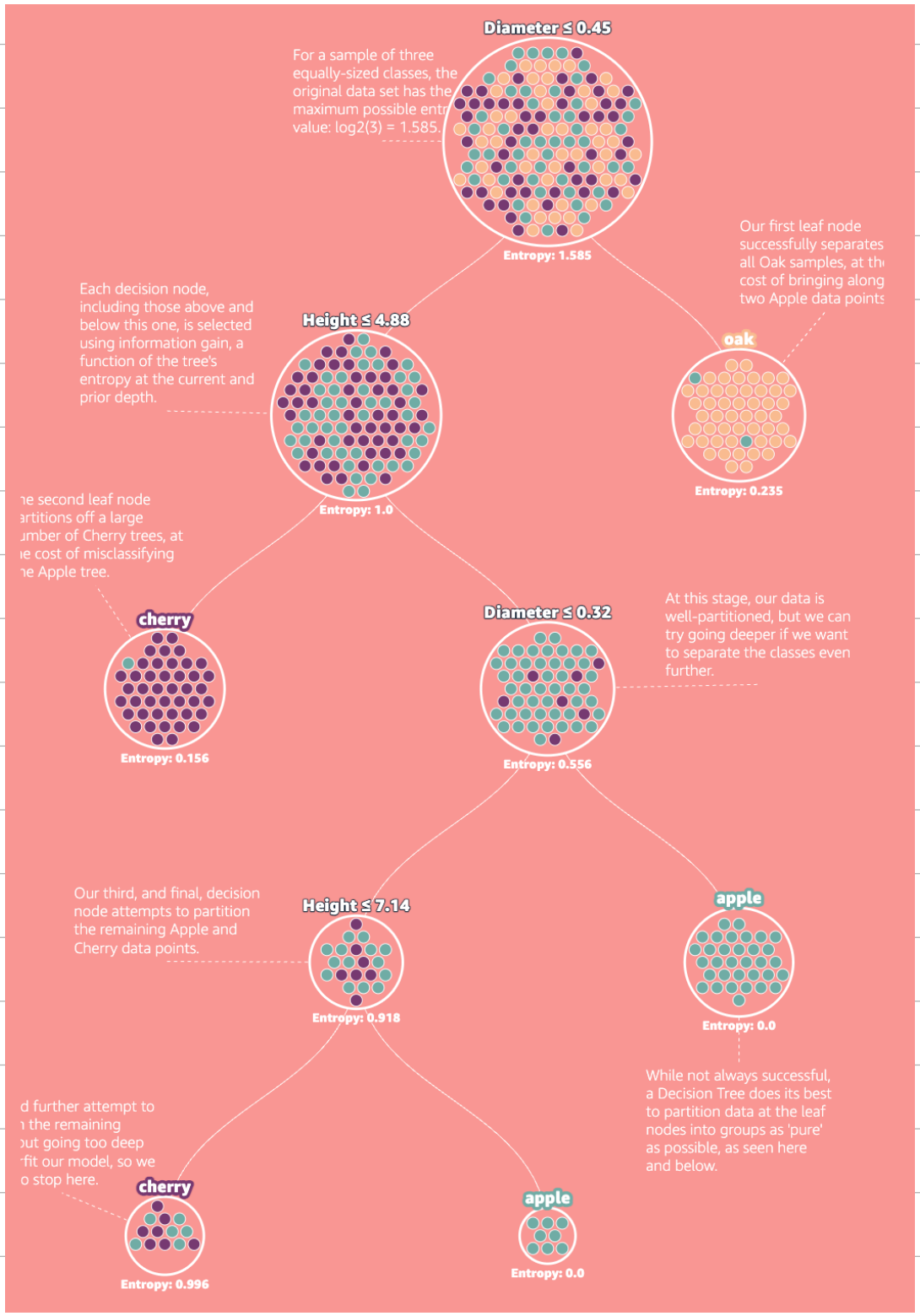
$$\Delta IG = H_{\text{parent}} - \frac{1}{N} \sum_{\text{children}} N_{\text{child}} \cdot H_{\text{child}}$$

To be specific, the algorithm's steps are as follows:

ID3 Algorithm Steps

1. Calculate the entropy associated to every feature of the data set.
2. Partition the data set into subsets using different features and cutoff values. For each, compute the information gain ΔIG as the difference in entropy before and after the split using the formula above. For the total entropy of all children nodes after the split, use the weighted average taking into account N_{child}, i.e. how many of the N samples end up on each child branch.
3. Identify the partition that leads to the maximum information gain. Create a decision node on that feature and split value.
4. When no further splits can be done on a subset, create a leaf node and label it with the most common class of the data points within it if doing classification or with the average value if doing regression.
5. Recurse on all subsets. Recursion stops if after a split all elements in a child node are of the same type. Additional stopping conditions may be imposed, such as requiring a minimum number of samples per leaf to continue splitting, or finishing when the trained tree has reached a given maximum depth.

• Same decision tree as above through a different lens:



- Setback :
  - ↳ Decision trees can be extremely sensitive to small pertubations in the data
    - ↳ A minor change in training data can result in a drastic change in the structure of the Decision Tree.
  - ↳ Decision Tree's are unstable